CENTRO UNIVERSITÁRIO SENAI
PROGRAMA DE PÓS-GRADUAÇÃO STRICTO SENSU
MODELAGEM COMPUTACIONAL E TECNOLOGIA INDUSTRIAL

CLEÔNIDAS TAVARES DE SOUZA JÚNIOR

Método para analisar autoria de textos baseado em regras de associação e redes de palavras

Outubro, 2022

Cleônidas Tavares de Souza Júnior

Método para analisar autoria de textos baseado em regras de associação e redes de palavras

Tese de Doutorado apresentada ao Programa de Pós-Graduação Stricto Sensu do Centro Universitário SENAI CIMATEC como requisito para a obtenção do título de **Doutor em Modelagem Computacional e Tecnologia Industrial**.

Orientador: Prof. Dr. Valter de Senna

Coorientador: Prof. Dr. Hernane Borges de Barros Pereira

Salvador
2022

Ficha catalográfica elaborada pela Biblioteca do Centro Universitário SENAI CIMATEC

S719m Souza Junior, Cleônidas Tavares de

Método para analisar autoria de textos baseado em regras de associação e redes de palavras / Cleônidas Tavares de Souza Junior. – Salvador, 2022.

135 f. : il. color.

Orientador: Prof. Dr. Valter de Senna.

Tese (Doutorado em Modelagem Computacional e Tecnologia Industrial) – Programa de Pós-Graduação, Centro Universitário SENAI CIMATEC, Salvador, 2022.
Inclui referências.

1. Análise de autoria. 2. Regras de associação. 3. Regras de palavras. 4. Combinações de palavras. I. Centro Universitário SENAI CIMATEC. II. Senna, Valter de. III. Título.

CDD 004

NDI - 01

Nota sobre o estilo do PPGMCTI

Esta tese de doutorado foi elaborada considerando as normas de estilo (i.e. estéticas e estruturais) propostas e aprovadas pelo colegiado do CENTRO UNIVERSITÁRIO SENAI e estão disponíveis em formato eletrônico (*download* na Página Web http://ead.fieb.org.br/portal_faculdades/dissertacoes-e-teses-mcti.html ou por solicitação via e-mail à secretaria do programa) e em formato impresso somente para consulta.

Ressalta-se que o formato proposto considera diversos itens das normas da Associação Brasileira de Normas Técnicas (ABNT), entretanto opta-se, em alguns aspectos, por seguir um estilo próprio elaborado e amadurecido pelos professores do programa de pós-graduação supracitado.

Sistema FIEB | SENAI CIMATEC
PELO FUTURO DA INOVAÇÃO

Centro Universitário SENAI CIMATEC

Doutorado em Modelagem Computacional e Tecnologia Industrial

A Banca Examinadora, constituída pelos professores abaixo listados, leu e aprovou a Tese de doutorado, intitulada "Método para Analisar Autoria de Textos Baseado em Regras de Associação e Redes de Palavras", apresentada no dia 25 de outubro de 2022, como parte dos requisitos necessários para a obtenção do Título de Doutor em Modelagem Computacional e Tecnologia Industrial.

Electronically signed by:
VALTER de Senna
CPF: ***.290.367-**
Date: 10/26/2022 5:29:58 PM -03:00

Orientador: Prof. Dr. Valter de Senna
SENAI CIMATEC

Assinado eletronicamente por:
Hernane Borges de Barros Pereira
CPF: ***.464.505-**
Data: 26/10/2022 10:55:45 -03:00

Coorientador: Prof. Dr. Hernane Borges de Barros Pereira
SENAI CIMATEC

Assinado eletronicamente por:
MARCELO Albano MORET Simões Gonçalves
CPF: ***.131.127-**
Data: 22/11/2022 09:00:40 -03:00

Membro Interno: Prof. Dr. Marcelo Albano Moret Simões Gonçalves
SENAI CIMATEC

Assinado eletronicamente por:
Inacio de Sousa Fadigas
CPF: ***.786.685-**
Data: 26/10/2022 20:27:01 -03:00

Membro Externo: Prof. Dr. Inacio de Sousa Fadigas
UEFS

Assinado eletronicamente por:
Marcelo do Vale Cunha
CPF: ***.534.205-**
Data: 01/11/2022 15:18:14 -03:00

Membro Externo: Prof. Dr. Marcelo do Vale Cunha
IFBA

Assinado eletronicamente por:
MARCOS GRILO ROSA
CPF: ***.945.715-**
Data: 26/10/2022 14:57:23 -03:00

Membro Externo: Prof. Dr. Marcos Grilo Rosa
UEFS

Agradecimentos

Agradeço:

À Deus;
À FABESB (Termo de Outorga de Bolsa nº: BOL0241/2018);
Aos meus orientadores Valter de Senna e Hernane de Borges;
Ao grupo de pesquisa Fuxicos e Boatos;
Aos membros da banca;
Ao Senai-CIMATEC;
Aos meus colegas e amigos;
À minha Família Brasileira;
À minha Nova Família Angolana;
À minha esposa Lidia e a minha filha Ana Kiami da Silva de Souza

Salvador, Brasil
25 de Outubro, 2022

Cleônidas Tavares de Souza Júnior

Resumo

Esta tese trata de métodos para análise e verificação de autoria. De modo mais específico, aborda-se a análise de autoria (AA) de textos escritos, em especial, obras literárias em língua portuguesa do Brasil e de Portugal. A AA de textos investiga o quanto uma obra de autoria desconhecida é semelhante ao conjunto de obras de um autor conhecido (LAGUTINA *et al.*, 2019; ROCHA *et al.*, 2017; VENCKAUSKAS *et al.*, 2015; BOUANANI; KASSOU, 2014; TAMBOLI; PRASAD, 2013; STAMATATOS, 2009; HOLMES, 1985). Autenticar o real autor de uma obra é importante, pois previne que falsas atribuições de autoria sejam feitas e evita, por exemplo, que a notoriedade de um escritor seja usada para difundir ideias que, originalmente, não são suas. Para analisar as semelhanças entre as obras, a AA extrai, organiza e compara estruturas que aparecem nos textos como quantidades de letras, comprimento das frases, repetição de palavras etc. (ROCHA, 2019; JAMIL; MUSTAFA, 2018; REXHA *et al.*, 2007; MARKOV; BAPTISTA; PICHARDO-LAGUNAS, 2017; VOROBEVA, 2016). Com o presente trabalho, identificou-se a carência, na AA, de estudos que verifiquem a autoria de um texto a partir das combinações de palavras (i.e. conjuntos de palavras que recorrentemente aparecem entre as frases de um escritor e que não aparecem nos conjuntos de palavras de outros escritores). Nesse sentido, esta tese tem como objetivo apresentar um novo método para verificação de autoria. Assume-se que as combinações das palavras não acontecem de modo aleatório; elas ocorrem em conformidade com o conhecimento sintático e semântico que os autores têm e evidenciam de sua língua (CHOMSKY, 2018; CHOMSKY, 1994; MIOTO; SILVA; LOPES, 2007; FRANCHI; NEGRAO; MULLER, 1998). A vantagem de se analisar combinações de palavras está em descobrir padrões relativos a cada autor e aos contextos de produção e publicação de cada obra. O método proposto nesta tese extrai combinações de palavras por meio de regras de associação, consolida as combinações em redes de palavras e, a partir de dezesseis métricas de rede, analisa e infere, em obras literárias, os períodos das edições, as variedades de língua portuguesa utilizadas, as escolas literárias e os autores. Nesse sentido, esta tese contribui para a AA com um método de trabalho que, além de verificar autorias, evidencia os contextos que, supostamente, um texto inédito de um autor deveria apresentar.

Palavras-chave: Análise de autoria, regras de associação, redes de palavras, combinações de palavras.

Abstract

This thesis deals with methods for authorship analysis and verification. More specifically, authorship analysis (AA) of digital texts, in particular, literary works in Portuguese from Brazil and Portugal. The AA investigates how similar is a work by an unknown author to a set of works by a known author (LAGUTINA *et al.*, 2019; ROCHA *et al.*, 2017; VENCKAUSKAS *et al.*, 2015; BOUANANI; KASSOU, 2014; TAMBOLI; PRASAD, 2013; STAMATATOS, 2009; HOLMES, 1985). Authenticating the real author of a work is important, it prevents false authorship attributions and prevents, for example, that the notoriety of a writer is used to spread ideas that, originally, are not his. To analyze the similarities between the works, the AA extracts, organizes and compares structures that appear in the texts, such as number of letters, length of sentences, repetition of words, etc. (ROCHA, 2019; JAMIL; MUSTAFA, 2018; REXHA *et al.*, 2007; MARKOV; BAPTISTA; PICHARDO-LAGUNAS, 2017; VOROBEVA, 2016). In AA, we identified a lack of studies that verify the authorship of a text from combinations of words (i.e. sets of words that recurrently appear among the sentences of a writer and that do not appear in the sets of words of other writers). In this sense, this thesis aims to present a new method for authorship verification. Word combinations are assumed not to happen randomly; they occur in accordance with the syntactic and semantic knowledge that the authors have and evidence of their language (CHOMSKY, 2018; CHOMSKY, 1994; MIOTO; SILVA; LOPES, 2007; FRANCHI; NEGRAO; MULLER, 1998). The advantage of analyzing word combinations lies in discovering patterns related to each author and the contexts of production and publication of each work. The method proposed in this thesis extracts word combinations through association rules, consolidates the combinations into word networks and, from sixteen network metrics, analyzes and infers, in literary works, the periods of editions, the varieties of language Portuguese used, the literary schools and the authors. In this sense, this thesis contributes to AA with a working method that, in addition to verifying authorship, highlights the contexts that, supposedly, an unpublished text by an author should present.

Keywords: Authorship analysis, association rules, word networks, word combinations.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Siglas

AA	Análise de autoria
K-NN	*K-Nearest Neighbors*
PCA	*Principal Component Analysis*
POS	*Parts of Speech*
RNA	Redes Neurais Artificiais
SVM	*Support Vector Machine*
PPGMCTI	Modelagem Computacional e Tecnologia Industrial

Capítulo Um

Introdução

Análise de autoria (AA) é o termo conferido aos estudos interessados em atribuir crédito a um documento (ou a partes de um documento) inferindo, a partir dele (e para ele), um ou mais autores. Na análise de autoria de textos, acredita-se que todo autor deixa marcas nos documentos que produz e que essas marcas podem ser encontradas em diferentes partes dos textos produzidos. Ao identificar o real autor de um documento, a AA contribui para: o combate às falsas atribuições de autorias (MENDONCA, 2002; G1, 2020); o combate às tentativas de sequestro de contas online (NIRKHI; DHARASKAR; THAKARE, 2016); o bloqueio de publicações falsas ou mal intencionadas (SOLORIO; HASAN; MIZAN, 2014); o combate à desinformação (G1, 2020; SOLORIO; HASAN; MIZAN, 2014) etc.

Na análise de autoria de textos digitais, uma etapa importante é a organização dos documentos que servirão de subsídio para a investigação. Textos produzidos em situações informais (e.g. mensagens para amigos, conversas em ambientes de bate-papo etc.) tendem a ser mais flexíveis em relação às normas ortográficas[1] e gramaticais (assumindo o português padrão[2]). Por outro lado, em textos produzidos em situações formais (e.g. *e-mail* de trabalho, obra literária etc.) observam-se, muitas vezes, os rigores ortográficos e gramaticais sendo aplicados. Em um conjunto de documentos, misturar textos produzidos em diferentes situações comunicativas torna a análise de autoria confusa e imprecisa (GRIEVE, 2007; HONORIO *et al.*, 2007; HOLMES, 1985). Entre os documentos analisados, a diversidade de situações comunicativas em que estes são produzidos é um problema para a AA, pois não se sabe ao certo o que está motivando a distinção entre as predições de autorias; não se sabe, por exemplo, se a diferença inferida entre dois autores está relacionada aos formalismos nos textos ou a características linguísticas próprias de cada autor.

Nos textos, os contextos de produção nem sempre estão explícitos. Nas obras literárias, por exemplo, o contexto pode aparecer por meio de marcas que evidenciam o período de publicação da edição, a variedade linguística, a escola literária e o autor. Explorar as relações entre os contextos implícitos e os autores nas obras literárias é uma área pouco explorada no âmbito da AA. Faltam, na análise de autoria, métodos que evidenciem, nos

[1] A ortografia é uma convenção cujo objetivo é regular a representação dos sons da fala na modalidade escrita da língua. As regras ortográficas não interferem no funcionamento do sistema linguístico, mas têm papel político importante, por exemplo, no que diz respeito à planificação linguística (considerando o estatuto de uma língua ou variedade) e ao compartilhamento de documentos escritos (FIORIN, 2009).

[2] Assume-se que uma propriedade inerente às línguas naturais é a variação. É possível pensar na variação linguística em sentido amplo, quando são consideradas as diferentes línguas do mundo; e em sentido estrito, quando se considera que, dentro de uma mesma língua, há variação (BELINE, 2010). Nesse sentido, é possível falar, por exemplo, em português padrão (como uma abstração da língua); português culto e português popular (usos reais da língua) (BAGNO, 2007).

textos, as marcas que diferenciam, além da autoria, os contextos de produção textual ligados a cada autor e as obras estudadas.

1.1 Definição do problema

Na análise de autoria, o pesquisador busca, nos documentos investigados, elementos textuais que sirvam de subsídio para inferir o real autor de um texto. A busca por esses elementos pode ser realizada de diferentes formas: pela análise de caracteres (ROCHA, 2019; JAMIL; MUSTAFA, 2018; MARKOV; BAPTISTA; PICHARDO-LAGUNAS, 2017); tamanhos das frases (VOROBEVA, 2016; SOUSA-SILVA *et al.*, 2010); frequências de palavras (VARELA; ALBONICO; ASSIS, 2019; GALINA; FLORES; KOMATI, 2019; SAVOY, 2018) etc.

Por outro lado, uma proposta de análise de combinações de palavras recorrentes (i.e. conjuntos de palavras[3] que aparecem muitas vezes em uma mesma frase) nos textos dos autores é uma linha de pesquisa ainda pouco explorada. O problema desse tipo de análise é que, mesmo em um texto com poucos parágrafos, podem existir milhões de combinações de palavras e essa quantidade de dados torna o processo de análise oneroso. Uma solução seria sintetizar essas combinações em uma estrutura que reduza o volume dos dados e mantenha as relações entre as palavras de forma que seja possível extrair informações relacionadas ao conjunto de combinações analisadas. A vantagem de se analisar as combinações de palavras nos textos está em encontrar elementos que possam contribuir para a verificação de autoria e contextos relativos a cada obra.

1.2 Questões e hipóteses

Nesta seção, são apresentadas as questões e as hipóteses que guiam o desenvolvimento das discussões apresentadas nesta tese.

Os textos são produzidos a partir de regras que combinam, de forma coerente e coesa, palavras lexicais e gramaticais/funcionais[4]. Os textos literários, por exemplo, além de carregarem marcas que são características a cada autor, registram marcas que são próprias do contexto de produção (período de publicação da edição, variedade linguística e escola literária).

[3] Não importa se as palavras estão próximas (i.e. lado-a-lado) ou distantes (i.e. separadas por outras palavras) em uma frase.

[4] Grosso modo, palavras lexicais são as que nomeiam os objetos no mundo (real ou não) e palavras gramaticais/funcionais são as que auxiliam o funcionamento da língua (NEVES, 2006; CAVALCANTI, 2004).

Questão 1: A Teoria e Ciência das Redes[5] é usada para analisar relacionamentos em diferentes tipos de sistemas. A partir de um conjunto de combinações de palavras, o uso da Teoria e Ciência das Redes contribui para a verificação de autoria e do contexto de produção de uma obra (período de publicação da edição, variedade linguística e escola literária)?

Hipótese 1.1: Se os textos são estruturados a partir da combinação de palavras, então usar a estrutura da rede para sintetizar (i.e. agrupar, reunir) as combinações de palavras dos autores é uma solução para a verificação de autoria e do contexto de produção de um texto.

Hipótese 1.2: Se as palavras apresentam características específicas (i.e. são lexicais ou gramaticais/funcionais) no funcionamento da língua e na estruturação de um texto, então organizar as redes agrupando-as de formas diferentes (baseando-se em suas características) identifica as distinções das marcas entre os diferentes contextos de produção.

Questão 2: Como, a partir de um conjunto de combinações de palavras, o uso da Teoria e Ciência das Redes contribui para a verificação de autoria?

Hipótese 2.1: Nas redes, existem métricas que avaliam a importância dos vértices de diferentes formas; entre as métricas dos vértices das redes, os valores próximos indicam que estas pertencem a um mesmo autor.

Hipótese 2.2: Se os conjuntos de combinações de palavras que distinguem um autor de outro são muito diferentes, as diferenças entre dois autores se manifestam com maiores números entre as métricas de redes.

Questão 3: Como, a partir da combinação de palavras, o uso da Teoria e Ciência das Redes infere as informações sobre o contexto de produção de uma obra?

Hipótese 3.1: Se as obras literárias, ao longo do tempo, passam por alterações (e.g. atualização ortográfica), então as redes construídas a partir da combinação de palavras (lexicais e gramaticais) apresentam propriedades capazes de distinguir edições produzidas em períodos diferentes de tempo.

Hipótese 3.2: Se as combinações de palavras caracterizam variedades de português (brasileiro e europeu[6]), então as métricas de redes evidenciam essa distinção nos conjuntos de

[5] Os estudos das redes nascem de um ramo da matemática chamada grafos; esse ramo da matemática trabalha com conjuntos de objetos e as relações existentes entre eles.

[6] Distinguir uma língua de outra passa por aspectos formais (sistema de regras do funcionamento da língua), aspectos históricos, políticos, sociais e identitários (ainda que dois sistemas linguísticos apresentem profunda semelhança, podem ser considerados línguas distintas, pois essa distinção também está associada ao reconhecimento e percepção de seus falantes). Há linguísticas, como Marcos Bagno, que defendem que português brasileiro

palavras.

Hipótese 3.3: Se as escolhas das palavras e suas formas de organização carregam marcas do contexto de produção, então as redes construídas a partir de combinações de palavras (lexicais e gramaticais) possuem propriedades capazes de distinguir textos entre duas escolas literárias.

1.3 Objetivos

Este trabalho tem como objetivo propor um método de análise que, com o auxílio da Teoria e Ciência das Redes, use combinações de palavras para verificar autorias e contextos de produção (e.g. período de edição, variedade linguística e escolas literárias) em obras literárias escritas em português.

Os objetivos específicos são:

1. Construir um *corpus* a partir da reunião de obras literárias de diferentes autores, de distintos períodos de edição, escritas em português brasileiro e europeu e filiadas a diferentes escolas literárias;
2. Criar um algoritmo que extraia subconjuntos de obras com base em parâmetros delineados a partir de características da produção de um texto;
3. Analisar quais tipos de combinações de palavras (lexicais e gramaticais) e métricas de redes são relevantes para caracterizar o período de edição, a variedade linguística, as escolas literárias e a autoria de uma obra (ou conjunto de obras).

1.4 Limites e limitações

Esta pesquisa limita-se a investigar a autoria, as escolas literárias e os anos de edições de obras literárias escritas em português brasileiro e português de europeu publicadas entre 1.843 e 2.016 (conjunto de textos no Apêndice A). As análises de autoria podem ser feitas com textos escritos em qualquer língua, no entanto, algumas premissas devem ser respeitadas. É premissa, por exemplo, que o conjunto de textos analisados deve pertencer

e português europeu são línguas distintas. Nesta tese, assume-se, porém, que português brasileiro e português europeu, embora reconhecidas todas as suas distinções, como atestam diferentes pesquisas (MATEUS *et al.*, 2003; MULLER; OLIVEIRA, 2004; CYRINO, 2020), são variedades de uma mesma língua, a língua portuguesa, compreendida aqui como um sistema de nível mais alto de abstração que englobaria regras comuns subjacentes às diferentes variedades (SILVA, 1988).

a escritores de mesma língua. Nas Seções 3.3 e 4.3, estão as motivações para assumir essa e outras premissas também importantes para os estudos sobre as análises de autorias.

O método apresentado nesta tese cria redes a partir da combinação de palavras em conjuntos de textos organizados de acordo com a natureza das palavras (i. e. se lexicais ou gramaticais, conforme será explicitado na Seção 2.2). A lista de palavras gramaticais/funcionais, aqui categorizadas como *stopwords*, está apresentada no Apêndice B e contem palavras como artigos, pronomes, preposições e numerais.

A proposta deste método de análise de autoria é avaliar, por meio das redes, as combinações de palavras recorrentes nos textos de um escritor. Optou-se por analisar combinações com no máximo três palavras, pois um número maior que esse demanda significativo uso das memórias do computador causando lentidão e travamento.

1.5 Importância da pesquisa

Um pesquisador que trabalha com análise de autoria busca, nos textos, elementos que possam ser usados para inferir o real autor de uma obra. Faz parte do processo de trabalho do pesquisador avaliar os textos sob investigação considerando diferentes métricas de análise e contextos, pois um falsificador pode usar determinadas marcas textuais características de um escritor para se passar por ele e, assim, difundir ideias que não são do real autor (BRENNAN; AFROZ; GREENSTADT, 2012; G1, 2020; MENDONCA, 2002).

Analisar a autoria de um texto com apoio da Teoria e Ciência das Redes torna o processo de identificação de fraudes mais fácil, pois, além de reproduzir as marcas textuais de um escritor, o falsificador deverá levar em consideração diversas métricas e combinações de elementos textuais que podem ser analisados nas redes (MACHICAO *et al.*, 2018). De modo distinto de outros trabalhos, o método proposto nesta pesquisa inclui, na investigação da autoria, uma ferramenta que, por meio de regras de associação, complementa a análise de autoria em redes considerando diferentes combinações de palavras em diferentes partes de um texto.

Os resultados e os métodos desenvolvidos neste trabalho dialogam com as três linhas de pesquisa do Programa de Pós-Graduação Stricto-Sensu em Modelagem Computacional e Tecnologia Industrial (PPGMCTI) do Centro Universitário SENAI CIMATEC. No que se refere à linha de pesquisa de Modelagem de Sistemas Cognitivos, o método proposto formaliza, nas redes, por meio de combinações de palavras, marcas que distinguem as escolhas de palavras pessoais de cada escritor e as usa para verificar a autoria de um documento; além das marcas pessoais, o método distingue características relativas ao período das edições, variedade linguística e escolas literárias das obras analisadas.

Na linha de pesquisa de Modelagem de Processos Industriais, esta tese apresenta uma ferramenta para investigação de autoria que pode ser usada pelo mercado editorial para avaliar a autenticidade de uma obra literária e evidenciar falsas publicações, caracterizando-se, assim, como uma contribuição para a análise de patentes e para o desenvolvimento de técnicas e métodos que contribuem para o combate à desinformação.

Na linha de pesquisa de Sistemas Complexos, além de usar teorias e algoritmos aplicados na investigação de sistemas complexos (e.g. Teoria e Ciência das Redes, algoritmos de predição etc.), este trabalho alinha teorias linguísticas, teorias estatísticas e teorias matemáticas para propor a criação de um método que visa formalizar conceitos relativos a autorias e contextos de produção de uma obra. Nesse sentido, esta tese apresenta caráter interdisciplinar, uma vez que relaciona diferentes correntes teóricas para propor um novo método de análise.

1.6 Motivação

Nesta tese, assume-se que, nas línguas, existem regras formais que dizem como as palavras devem ser combinadas para que um falante possa formular um sintagma[7] (e transmitir uma mensagem) e para que o seu interlocutor possa interpretá-lo e, assim, compreender a mensagem transmitida (CHOMSKY, 2018; CHOMSKY, 1994; MIOTO; SILVA; LOPES, 2007; FRANCHI; NEGRAO; MULLER, 1998). No entanto as escolhas de quais palavras serão combinadas (e como serão combinadas) para transmitir a mensagem variam entre as pessoas.

Como dito, ao produzir um texto, as combinações de palavras que as pessoas fazem obedecem às regras gramaticais de funcionamento da língua; mas relevam, também, as escolhas próprias do autor. Consciente ou inconscientemente, o escritor deixa, nos textos, marcas e informações que vão além das mensagens transmitidas; registram, por exemplo, a situação na qual o texto foi escrito e a variedade da língua. Explorar as combinações de palavras é uma maneira de encontrar e formalizar os contextos incorporados nos textos.

A análise de autoria busca elementos que possam ser usados para evidenciar a origem de um documento sem autoria definida. A proposta apresentada neste trabalho torna-se relevante, pois, diferente de outros métodos, ao explorar as combinações de palavras, visa encontrar elementos que caracterizam, além do autor, informações do contexto de produção presentes no texto.

[7] Um sintagma é uma estruturação de palavras que, embora superficialmente pareça uma mera sequência linear de sons ou letras, está organizada de forma hierárquica (MIOTO; SILVA; LOPES, 2007).

1.7 Organização da tese

Nesta seção, conforme exposto abaixo, estão descritas as etapas de construção e desenvolvimento desta pesquisa.

Capítulos:

- **Capítulo 1 - Introdução**: Contextualiza, em relação à análise de autoria, os temas que serão tratados neste trabalho, os problemas da pesquisa, os objetivos, motivações, limitações, hipóteses e a importância da pesquisa.
- **Capítulo 2 - Métodos usados na revisão da literatura e na organização do *corpus***: Descreve os métodos e as estratégias usadas para a revisão da literatura e a construção do *corpus* desta pesquisa.
- **Capítulo 3 - A análise de autoria de textos**: Delineia, brevemente, a história da análise de autoria em textos impressos; organiza e descreve métodos e termos técnicos da área; e, no final, lista as principais premissas teóricas da análise de autoria.
- **Capítulo 4 - Análises de autorias com redes**: Descreve algumas definições relativas à Teoria e Ciência das Redes e algumas de suas contribuições para as análises de autorias. No final do capítulo, são apresentadas algumas premissas teóricas das análises de autoria que trabalham com redes.
- **Capítulo 5 - Métodos e materiais**: Descreve o funcionamento do processo de descoberta de regras de associação e apresenta o método de análise de autoria proposto neste trabalho.
- **Capítulo 6 - Resultados e discussões**: Aplica-se o método aqui proposto para realizar predições de ano de edição, escolas literárias, variedades da língua portuguesa e autoria dos textos.
- **Capítulo 7 - Considerações finais**: Apresenta as principais contribuições desta pesquisa, mostra os resultados relevantes das análises e descreve os próximos passos da pesquisa. Os resultados alcançados pelo método apresentado neste trabalho ilustram que as combinações de palavras dos textos registram marcas que identificam os autores e os contextos de produção do texto.

Apêndices:

- **Apêndice A - Obras que compõem o *corpus* da pesquisa:** Contém uma tabela

com os nomes e país de origem dos autores; títulos das obras; ano da primeira edição; ano da edição da publicação; e fonte da obra.

- **Apêndice B - Lista de *stopwords***: Lista de palavras gramaticais que fazem parte do conjunto de palavras da *stopwords* desta pesquisa.
- **Apêndice C - Algoritmo: Extrair *Corpus*:** Apresenta o algoritmo que foi desenvolvido para extrair subconjuntos de textos do *corpus* principal da pesquisa.
- **Apêndice D - Algoritmo: Etapa 01 - Transformar textos em conjuntos de itens:** Apresenta o algoritmo desenvolvido para transformar textos em conjuntos de itens.
- **Apêndice E - Algoritmo: Etapa 02 - Transformar conjuntos de itens em regras de associação:** Algoritmo criado para transformar os conjuntos de itens em regras de associação.
- **Apêndice F - Algoritmo: Etapa 03 - Transformar regras de associação em redes**: Código do algoritmo criado para transformar regras de associação em redes.
- **Apêndice G - Algoritmo: Etapa 04 - Extrair métricas das redes:** Algoritmo criado para extrair e organizar métricas das redes em uma tabela.
- **Apêndice H - Algoritmo: Processos de Predições**: Algoritmo criado para realizar predições a partir de métricas das redes.

Capítulo Dois

Métodos usados na revisão da literatura e na organização do *corpus*

Este capítulo apresenta os métodos, os resultados e os critérios de seleção e organização dos documentos usados nesta pesquisa. Este trabalho reuniu publicações científicas (i.e. artigos, dissertações e teses) sobre a análise de autoria de textos digitais e obras literárias escritas em língua portuguesa do Brasil e de Portugal.

Na primeira seção do capítulo, estão dispostos os métodos usados para a construção da pesquisa descritiva feita sobre a análise de autoria em textos digitais. Nessa seção, estão os nomes das fontes das pesquisas, as quantidades de documentos encontrados, os descritores[1] e o método usado para selecionar as publicações sobre as análises de autorias.

Na segunda seção, são descritos os métodos de coleta e organização dos textos (i.e. *corpus*) usados pelas análises de autorias deste trabalho. O *corpus* desta pesquisa reuniu obras literárias de autores portugueses e brasileiros publicadas em diferentes períodos. Para organizar os conteúdos dessas obras e gerar material para as análises de autoria deste trabalho, foi construído um algoritmo capaz de extrair diferentes conjuntos de textos (i.e. conjuntos de textos com mesmas características literárias).

A coleta e organização dos documentos foram relevantes para a construção do referencial teórico do Capítulo 3 e do Capítulo 4. Nesses capítulos, são descritos os termos técnicos, fatos históricos, métodos de análises, vantagens e desvantagens relacionadas à análise de autoria em geral e a análise de autoria com o suporte da Teoria e Ciência de Redes e de regras de associações.

2.1 Método de revisão da literatura

As publicações analisadas nesta pesquisa foram obtidas nos sites *Google Acadêmico*[2], *Periódicos CAPES*[3] e *Web of Science*[4]. Os principais temas pesquisados foram itens relacionados à análise de autoria e suas relações com métodos de construção de redes e regras de associação. A Tabela 2.1 exibe os descritores e as quantidades de publicações localizadas em cada site antes da filtragem; a Tabela 2.2 exibe os descritores e as quantidades

[1] Termos usados para encontrar um determinado assunto entre publicações científicas.
[2] https://scholar.google.com.br/?hl=pt
[3] https://www.periodicos.capes.gov.br/
[4] http://www.webofknowledge.com/

de publicações após a filtragem.

Tabela 2.1: Descritores, fontes da pesquisa e quantidades de documentos encontrados

Descritores	***Google Acadêmico***	***Periódicos CAPES***	***Web of Science***
"atribuição de autoria"	1.640	17	5
"authorship assignment"	1	82	14
"authorship identification"	47	285	1392
"authorship identification"e "complex networks"	93	11	1
"authorship verification"	22	126	79
"identificação de autoria"	18.600*	12	0
"identificação de autoria"e "redes complexas"	7	0	0
"verificação de autoria"	93	0	0
atribuição de autoria	129.000*	657	9
identificação de autoria	192.000*	1.526	112
verificação de autoria	79.000*	418	0
"identificação de autoria"e "regras de associação"	3	0	0
"authorship identification"e "association rules"	142	7	0
"verificação de autoria"e "regras de associação"	1	0	0
"authorship verification"e "association rules"	34	0	0

*Este é o resultado da primeira busca sem filtro.

Fonte: Elaboração própria

Na Tabela 2.1, os resultados marcados com asteriscos (valores superiores a dez mil documentos) ocorreram porque o site *Google Acadêmico* buscou cada palavra do descritor no corpo principal dos textos consultados por ele; como consequência, retornou documentos que não tinham relações com este trabalho; retornou, por exemplo: (i) solicitações de revistas pedindo para que pesquisadores realizassem a *identificação de autoria* conforme as normas da revista; (ii) informes sobre a importância da *verificação de autoria* em peças jurídicas etc. Para refinar a pesquisa, por meio de filtros, solicitou-se ao *Google Acadêmico* que retornasse somente publicações relacionadas a artigos de análises de autoria; como resultado, os valores baixaram para uma média de cem documentos por busca, como ilustra a Tabela 2.2.

Tabela 2.2: Descritores, fontes da pesquisa e quantidades de documentos encontrados após a filtragem

Descritores	***Google Acadêmico***	***Periódicos CAPES***	***Web of Science***
"atribuição de autoria"	1.640	17	5
"authorship assignment"	1	82	14
"authorship identification"	47	285	1392
"authorship identification"e "complex networks"	93	11	1
"authorship verification"	22	126	79
"identificação de autoria"	101	12	0
"identificação de autoria"e "redes complexas"	7	0	0
"verificação de autoria"	93	0	0
atribuição de autoria	101	657	9
identificação de autoria	91	1.526	112
verificação de autoria	101	418	0
"identificação de autoria"e "regras de associação"	3	0	0
"authorship identification"e "association rules"	142	7	0
"verificação de autoria"e "regras de associação"	1	0	0
"authorship verification"e "association rules"	34	0	0

Fonte: Elaboração própria

Os materiais investigados pelas análises de autorias são extraídos de manuscritos, documentos digitais, pinturas, descobertas históricas etc. Por conta dessa abrangência, os documentos encontrados pelos descritores da Tabela 2.1 agruparam publicações relacionadas a diferentes materiais. Para selecionar somente publicações relativas a textos digitais[5], adotaram-se os seguintes critérios de inclusão: (i) ler os títulos e resumos das publicações nas fontes da pesquisa; (ii) baixar somente publicações sobre análise de autoria em textos digitais[6]; (iii) ler e analisar as publicações selecionadas em (ii). As pesquisas realizadas nos *Periódicos CAPES* e na *Web of Science* não necessitaram desse detalhamento de filtragem.

[5]Textos impressos em papeis ou textos em formatos eletrônicos.

[6]Durante esse processo, também foram selecionadas publicações que relacionavam as análises de autorias a diferentes tipos de redes (e.g. redes complexas, redes de palavras, redes semânticas etc)

2.2 Métodos de coleta de dados e organização do *corpus*

Quando uma pessoa escreve, ela acrescenta nos seus textos características pessoais (e.g. jargões; marcas de expressividade, como alguns sinais de pontuação; saudações etc.) e essas características textuais podem variar quando ela escreve, por exemplo, um *e-mail* para um amigo ou um *e-mail* para um escritório de advocacia. No *e-mail* para o amigo, em contraste com o *e-mail* para o escritório de advocacia, o texto tende a ser mais informal e menos atento aos rigores ortográficos e gramaticais (no sentido de seguir as normas do português padrão). Na análise de autoria, as mudanças de características textuais devem receber uma atenção especial do pesquisador. O contexto no qual foi escrito o texto investigado deve ser similar ao contexto dos textos compilados dos supostos autores (GRIEVE, 2007); a desatenção às métricas e aos estilos tornam a atribuição de autoria imprecisa (HONORIO *et al.*, 2007).

A construção do *corpus* desta pesquisa reuniu textos do tipo ficção. Os textos coletados são obras literárias escritas por autores portugueses e brasileiros e foram obtidos no *Portal Domínio Público*[7], na página *Baixe Livros*[8], na *Biblioteca Digital de Literatura de Países Lusófonos* (BLPL)[9] e no *Projeto Gutenberg*[10]. No *corpus*, não entraram obras que, originalmente, foram escritas em outras línguas e depois traduzidas para o português. Para uma análise de autoria, é importante que os textos não sofram traduções, pois os tradutores deixam marcas próprias nos textos que traduzem e tornam obscuros os padrões textuais do real autor da obra (ARUN; SURESH; MADHAVAN, 2009; ANDRADE, 2015). Desconsiderar as traduções foi um critério de exclusão de textos.

O *corpus* é composto por oitenta e nove livros pertencentes a trinta e dois autores (vinte e cinco brasileiros e sete portugueses[11]) publicados entre o período de 1.843 a 2.016 (os nomes dos autores e das obras estão no Apêndice A). As obras reunidas pertencem a nove escolas literárias. A Tabela 2.3 descreve, brevemente, as características literárias de cada escola e mostra os períodos de vigências de cada uma. Embora algumas escolas literárias tenham coexistido em um determinado momento (e.g. realismo, naturalismo e parnasianismo), semântica e sintaticamente, elas formalizaram características próprias.

Com o tempo, os textos das obras literárias são modificados pelas editoras para adequar a ortografia das palavras e fazer atualizações. Por exemplo, a primeira edição de uma obra publicada em 1.800 não é exatamente igual à edição publicada em 2.021. É comum encontrar, no prefácio das obras, notas dos editores comentando quais foram as modificações realizadas nos textos da publicação da nova edição. Há, no *corpus* desta pesquisa,

[7] http://www.dominiopublico.gov.br
[8] http://www.baixelivros.com.br
[9] https://www.literaturabrasileira.ufsc.br
[10] https://www.gutenberg.org/
[11] A diferença entre a quantidade de textos de autores brasileiros e autores portugueses deveu-se à disponiblidade de documentos na internet

obras com diferentes anos de edição. O livro *Quincas Borba*, de Machado de Assis, por exemplo, aparece duas vezes, uma com a edição de 1.891 e a outra com a edição de 1.994.

Embora os textos no *corpus* tenham em comum o mesmo gênero literário, eles apresentam características textuais distintas. Há obras de diferentes escolas literárias, publicadas com base em diferentes regras ortográficas, com autores de diferentes nacionalidades etc. Para organizar as obras e permitir que diferentes subconjuntos de textos possam ser extraídos conforme um contexto específico, alguns procedimentos foram realizados. O primeiro foi transformar as obras em arquivos do tipo texto[12].

Tabela 2.3: Escolas literárias da língua portuguesa entre o arcadismo e o pós-modernismo

Escolas Literárias	Breves descrições	Principais autores (brasileiros e portugueses)	Linha do tempo (1700 – 1800 – 1900 – 2000)
Arcadismo	No arcadismo, a arte conecta o homem à natureza; descreve-se a fuga da cidade; há valorização do *carpe diem*, promoção e retomada de valores clássicos.	Cláudio Manuel da Costa, Basílio da Gama, Santa Rita Durão etc.	
Romantismo	Promove a valorização dos sentimentos; o amor e o sofrimento são temas recorrentes; há exaltação do ufanismo e fuga da realidade.	Gonçalves Dias, Gonçalves de Magalhães, José de Alencar, etc.	
Realismo	Aponta os defeitos e qualidades dos humanos; a narrativa foca a classe alta; faz críticas aos valores burgueses.	Machado de Assis, Eça de Queiroz, Eugênio de Castro, Camilo Pessanha, Cesário Verde, Antero de Quental etc.	
Naturalismo	A narrativa foca o determinismo, o cientificismo e temas do cotidiano.	Aluísio de Azevedo, Eça de Queiroz, Adolfo Ferreira Caminha etc.	
Parnasianismo	Retrata a realidade dos fatos; a estética é muito valorizada; há impessoalidade e vocabulário culto.	Teófilo Dias, Olavo Bilac, Alberto de Oliveira etc.	
Simbolismo	Aversão ao real, musicalidade, valorização dos conceitos espirituais e metafísicos.	João da Cruz e Sousa, Alphonsus de Guimaraens, Alberto Guerra Vidal etc.	
Pré-modernismo	Pensamentos ecléticos, temas do cotidiano; construção de uma linguagem simples e coloquial.	Augusto dos Anjos, Coelho Neto, Euclides da Cunha, Graça Aranha, Lima Barreto, Monteiro Lobato etc.	
Modernismo	Muita liberdade de expressão, sem apego ao passado.	Carlos Drummond de Andrade, João Cabral de Melo Neto, Clarice Lispector etc.	
Pós-modernismo	Crítica à ditadura, descrição do cotidiano, exaltação dos problemas sociais etc.	João Guimarães Rosa, Clarice Lispector, Dalton Trevisan etc.	

Fonte: Elaboração própria

As obras coletadas estavam em formatos digitais do tipo imagem e do tipo texto. A

[12] Arquivos em formato *.txt são fáceis de manusear e aceitos por muitos algoritmos.

transformação dessas obras em arquivos textos obedeceram aos seguintes critérios:

i. Nas obras, só foram extraídos os textos que pertenciam ao escritor (sumários, prefácio, notas de edição, notas de rodapé, dedicatórias e epígrafes não entraram nos arquivos textos);

ii. As palavras que não foram reconhecidas pelo programa de reconhecimento de caracteres foram ajustadas manualmente conforme o vocabulário usado em cada texto (i.e. em textos antigos, as palavras ajustadas mantiveram a ortografia usada nos textos antigos e não a baseada na atual regra ortográfica);

iii. Para preservar, nos textos, caracteres e acentos antigos e de outras línguas, os arquivos foram gravados no formato UTF-8[13];

iv. Cada arquivo texto contém, aproximadamente, as cinco mil primeiras palavras[14] das respectivas obras literária;

v. A primeira linha de cada arquivo foi preenchida com os seguintes dados: *nome do autor*; *título da obra*; *ano da primeira publicação*; *ano de edição da obra*; *total de palavras no arquivo texto*; *escola literária*; *país de origem do escritor* (e.g. *Joaquim Manuel de Macedo; A luneta mágica; 1869; 1990; 5028; romantismo; Brasil*);

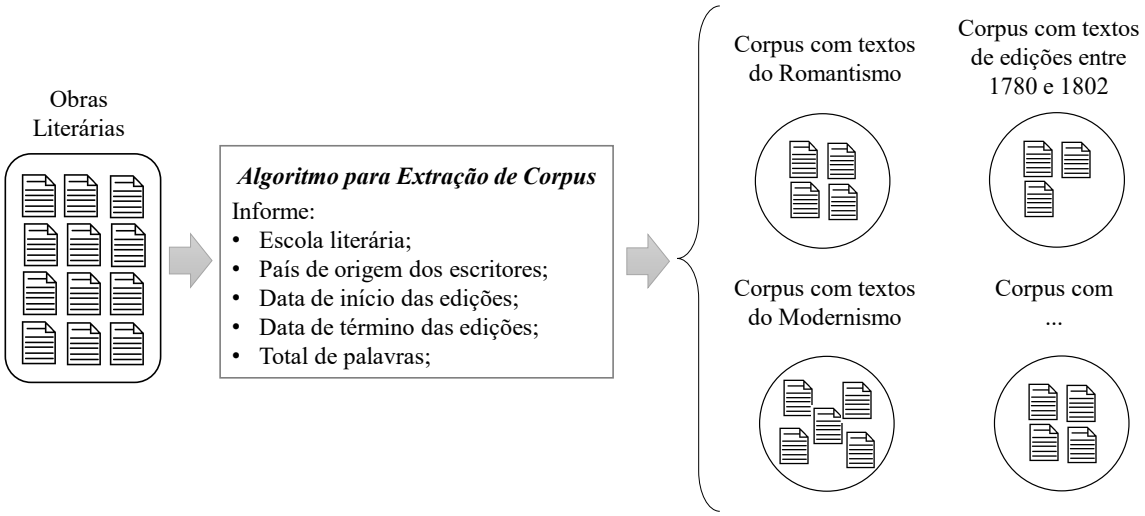
vi. o nome de cada arquivo texto seguiu o seguinte padrão: *nome do autor - ano da primeira edição da obra - título da obra - Ano da edição*.txt (e.g. *Machado de Assis - 1881 - Memórias Póstumas de Brás Cubas - 1994.txt*);

Padronizar os nomes dos arquivos e incluir metadados[15] na primeira linha dos textos foi a estratégia escolhida para estruturar os dados em um formato no qual algoritmos de mineração de textos possam reconhecer e extrair diferentes informações. Para atender as demandas desta pesquisa, o algoritmo (Apêndice C - Algoritmo: *Extrair Corpus*), representado na Figura 2.1, foi construído.

Com base nos parâmetros de entrada *escola literária, país de origem do escritor, ano de início das edições, ano de término das edições e total de palavras*, o algoritmo Extrair *corpus* (Figura 2.1) reúne subconjuntos de textos com características textuais uniformes aos parâmetros informados. A seleção por datas de edição possibilita que textos com ortografias de um mesmo período sejam extraídos; a seleção por países acrescenta à extração de *corpus* mais uma especificidade, por exemplo, selecionar só textos do romantismo português ou só textos do romantismo brasileiro.

[13] UTF-8 é um tipo de codificação de caracteres que incorpora letras e acentos dos alfabetos de diferentes línguas.
[14] Esse valor foi atribuído arbitrariamente, a proposta foi padronizar o tamanho dos textos e evitar que muitos dados sejam armazenados desnecessariamente.
[15] Informações relativas a um documento, mas que nem sempre aparecem nos conteúdos do documento (e.g. preço, local onde o documento foi criado).

Figura 2.1: Esquema básico de extração de *corpus*



Fonte: Elaboração própria

Ao executar o *algoritmo Extrair Corpus*, é solicitado que se informe: (i) o local onde estão os arquivos textos das obras literárias; (ii) o local onde o novo *corpus* será gravado; (iii) quais escolas literárias participarão da seleção; (iv) os países de origem dos escritores; (v) as datas de edição das obras (filtro por data de início e de término); (vi) a quantidade de palavras mínimas que cada texto extraído deve ter. Durante o processo de seleção dos textos, o *algoritmo Extrair Corpus* lê as primeiras linhas de cada arquivo texto e seleciona somente os arquivos que satisfazem aos parâmetros informados; à medida que o algoritmo localiza os arquivos, ele grava uma versão do arquivo texto no local de destino, porém sem a primeira linha com os metadados dos textos.

Ao estruturar e organizar as obras literárias em arquivos textos, esta pesquisa contribui para que outras pesquisas possam extrair e analisar diferentes subconjuntos de textos baseados nos metadados criados. O *algoritmo Extrair Corpus* complementa a contribuição com um método para seleção e extração de textos baseados nos parâmetros *escola literária, país de origem do escritor, ano de início das edições, ano de término das edições e total de palavras*. O *algoritmo Extrair Corpus* foi o responsável por selecionar os diferentes conjuntos de textos usados nas análises de autorias desta pesquisa.

Nesta tese, a organização dos textos coletados foi feita a partir do agrupamento destes em três formas: conjunto de textos com palavras lexicais e gramaticais (denominado de *Original*); conjunto de textos de onde são extraídas as palavras lexicais (permanecem as palavras gramaticais; denomindado de *Stopwords*); e conjunto de textos de onde é extraído o maior número possível de palavras gramaticais (permanecem as palavras lexicais; denominado de Sem *stopwords*).

Capítulo Três

A análise de autoria de textos

Em termos gerais, a análise de autoria (AA) de textos procura padrões nos textos que possam ser usados para caracterizar as obras de um autor. Em textos manuscritos (cf. (SCHOMAKER; FRANKE; BULACU, 2007; EGLIN; BRES; RIVERO, 2007)), a AA compara as características das escritas (e.g. inclinação das palavras, curvatura das letras) de diferentes autores, antes de sugerir o autor de um texto. Nos textos digitados (cf. (MENDENHALL, 1887; MARTINS *et al.*, 2019)), as características atribuídas aos autores estão relacionadas à diferentes métricas (e.g. frequência de letras, recorrência de palavras); essas características ou métricas servirão de subsídio para a AA sugerir o autor de um texto sem autoria definida.

Os estudos de AA de textos digitais têm diferentes aplicações; por exemplo, na identificação de pessoas por trás de pseudônimos em jogos *online* (KUZU; BALCI; SALAH, 2016); na identificação de autoria de *e-mails* anônimos (VEL, 2000; ZHANG; LIU; CHEN, 2015); na atribuição de autoria em postagem na internet (VOROBEVA, 2016); na verificação de autoria em contas do *Twitter*[1] atribuídas a falsas personalidades (LE; SAFAVI-NAINI, 2018); na detecção de predadores sexuais em sites de bate-papo *online* (CARDEI; REBEDEA, 2017); na identificação de autoria no desenvolvimento de *software* (ABUHAMAD *et al.*, 2019; WISSE; VEENMAN, 2015); no *marketing* para lançamento de livros publicados por pseudônimos (VERDU, 2019; SAVOY, 2018); na identificação e distinção de textos produzidos por humanos, *bots*[2] bem intencionados e *bots* maliciosos (BARBON *et al.*, 2018); na mensuração e atribuição de autoria no desenvolvimento de programas de computadores com múltiplos desenvolvedores (GRAY; SALLIS; MACDONELL, 1997; AVELINO *et al.*, 2019) etc.

Este capítulo descreve a AA de textos digitais sob a perspectiva histórica e conceitual. Apresenta, ainda, algumas premissas da AA compreendidas como relevantes para esta tese. Dessa forma, a Seção 3.1 começa com uma breve história da análise de autoria de textos; os primeiros estudos, a evolução dos métodos, as aplicações e as análises de alguns resultados. Na Seção 3.2, as linhas de pesquisas, os conceitos e os métodos são descritos, gradativamente, a partir das finalidades de cada estudo, concluindo-se com as visões gerais das principais linhas de pesquisas. Na Seção 3.3, estão algumas premissas teóricas das análises de autorias.

[1] *Twitter*: Rede social *online* por meio da qual os usuários trocam informações.
[2] *Bot*: Abreviação de *robot*; são algoritmos usados para interagir com humanos como se fossem humanos.

3.1 Uma breve história sobre a análise de autoria em textos

Um dos primeiros estudos a verificar a autoria de textos, com base em métricas estatísticas, foi feito por Mendenhall (1887). Ao observar a distribuição de frequências entre o número e o tamanho das palavras de textos de William Makepeace Thackeray e Charles John Huffam Dickens, Mendenhall (1887) identificou, nas distribuições, padrões que distinguiam os textos escritos por Thackeray dos textos escritos por Dickens.

No final da década de trinta, Yule (1939) obteve bons acertos na AA ao usar métricas (e.g. média, mediana) extraídas dos comprimentos das frases de textos de diferentes autores. Ao final do estudo, Yule (1939) concluiu que, embora tenha acertado os autores dos textos, as métricas de comprimentos de sentenças não são totalmente confiáveis para a atribuição de autoria.

As descobertas de padrões comuns em textos de um mesmo autor foram, aos poucos, fortalecendo e subsidiando a AA com novos métodos de análises. No entanto, nos primórdios da AA, a extração de métricas dos textos (e.g. total de palavras, comprimento das palavras) era trabalhosa e dispendiosa, pois era necessário contabilizar manualmente as palavras e as frases de cada texto. Com a invenção da computação moderna, os processos de mensuração de textos passaram a ser feitos por computadores e de modo programável. Essa mudança automatizou e facilitou o processo de análise.

Mosteller & Wallace (1964) foram pioneiros em usar métodos estatísticos e computacionais na análise da autoria. Eles investigaram a autoria dos documentos *"The Federalist"*[3] a partir de uma abordagem *não tradicional*[4] e apresentaram novas estratégias quantitativas para a identificação do autor mais provável dos documentos.

Em 1985, um poema inédito atribuído a Shakespeare foi encontrado em uma biblioteca na Inglaterra. A partir da análise de frequência das palavras nas obras de William Shakespeare, Thisted & Efron (1987) investigaram a autoria do poema. Apesar de o poema apresentar nove palavras que nunca apareceram em trabalhos anteriores de Shakespeare, Thisted & Efron (1987) verificam que o poema se encaixa razoavelmente ao estilo do poeta e inferem que o inédito poema foi escrito por Shakespeare.

A partir do final dos anos noventa, aumentam as produções textuais nas mídias da internet (e.g. *e-mails*, fóruns *online*, postagens) e também o número de novos métodos de extração de informações (STAMATATOS, 2009). Essa crescente produção textual impactou os estudos sobre aprendizados de máquinas, descobertas de informações e processamentos

[3] Textos que serviram de base para construção da Constituição dos Estados Unidos da América.

[4] Pode-se dizer que a abordagem de Mosteller & Wallace (1964) é *não tradicional* se comparada a estudos similares da época (STAMATATOS, 2009). Rudman (1998) classifica como *não tradicionais* os estudos que empregam métodos estatísticos e computacionais na atribuição de autoria

das linguagens naturais; o desenvolvimento desses estudos disponibilizou para a AA novas técnicas para processar grandes volumes de dados, novos algoritmos de aprendizado de máquina, ferramentas para análises estruturais de textos etc. (STAMATATOS, 2009).

Durante esse período inicial de popularização da internet, a AA apresentava alguns problemas como a falta de avaliação objetiva dos métodos de análise[5] e os métodos eram testados mais com obras literárias do que com outros tipos de textos (STAMATATOS, 2009). A falta de procedimentos operacionais padrões que assegurassem *confiança* e *validade* às análises de autorias fez com que alguns tribunais dos Estados Unidos começassem a solicitar testes que comprovassem a eficácia dos métodos de atribuição de autoria (CHASKI, 2001).

Confiabilidade e *Validade* são termos técnicos usados pela Linguística Forense para dar crédito a uma análise de autoria. Na AA, acredita-se que cada escritor tem um marcador de autoria (i.e. um conjunto de características linguísticas) que o torna único. A *validade* de um marcador de autoria implica explicar, com embasamentos teóricos, porque outros autores não têm o mesmo padrão; a *confiabilidade* está relacionada à capacidade de o método permitir sua aplicação em outras análises (GRANT; BAKER, 2001).

Um dos objetivos da AA é disponibilizar ferramentas e conhecimentos para que os pesquisadores possam investigar uma obra suspeita e elencar evidências que apontem seu autor. No entanto esses mesmos conhecimentos que ajudam os pesquisadores podem ser usados para ofuscar o real autor de uma obra. Brennan & Greenstadt (2009) e Brennan, Afroz & Greenstadt (2012) simularam ataques baseados em ofuscação[6] e imitação[7] e, com base nos critérios que usaram, obtiveram êxito em ocultar e falsificar a autoria de um texto. Juola & Vescovi (2011) refutaram, parcialmente, as hipóteses de Brennan & Greenstadt (2009) demonstrando que a análise de autoria utiliza muitas outras métricas que, combinadas, inferem o verdadeiro autor.

Os processos de ataques baseados em ofuscação e imitação se tornam mais suscetíveis a falhas à medida que mais métricas e métodos de análise são usados nas investigações. Estudos sobre redes complexas (ZAKI, 2001; CAVIQUE, 2004; CAVIQUE, 2007a; CAVIQUE, 2007b; ARUN; SURESH; MADHAVAN, 2009; VARELA, 2010; AMANCIO *et al.*, 2011; AMANCIO; OLIVEIRA-JR; COSTA, 2012a; MEHRI; DAROONEH; SHARIATI, 2012; LAHIRI; MIHALCEA, 2013; SEGARRA; EISEN; RIBEIRO, 2013; AREFIN *et al.*, 2014; DAROONEH; SHARIATI, 2014; SEGARRA; EISEN; RIBEIRO, 2014; CHAKRABORTY; CHOUDHURY, 2016; MARINHO; HIRST; AMANCIO, 2016; VALENCIA, 2017; ROZZ; MENEZES, 2018; MACHICAO *et al.*, 2018; AVELINO *et al.*, 2019;

[5] As avaliações dos resultados eram feitas, muitas vezes, de maneira intuitiva (STAMATATOS, 2009).

[6] O termo ofuscação (em inglês, *obfuscation attack*) é uma forma de ataque que acontece quando um escritor tenta disfarçar seu estilo de escrita para se passar por outra pessoa (BRENNAN; GREENSTADT, 2009).

[7] O termo imitação (em inglês, *imitation attack*) é uma forma de ataque que acontece quando uma pessoa tenta imitar o estilo de escrita de outra (BRENNAN; GREENSTADT, 2009).

EL-FIQI; PETRAKI; ABBASS, 2019; DUQUE; CARVALHO; VIMIEIRO, 2019; STANISZ; KWAPIEN; DROZDZ, 2019; HOU; HUANG, 2020), por exemplo, ao organizarem os textos em estruturas não lineares, acrescentam, nas análises de autorias, métricas não convencionais[8] e tornam o processo de ofuscação e imitação ainda mais trabalhosos.

A AA de textos digitais reúne conhecimentos para a investigação de diferentes tipos de textos e recebe, constantemente, novas publicações. Muitos desses conhecimentos estão organizados em artigos de revisões de literatura que descrevem as principais linhas de pesquisas da AA (LAGUTINA *et al.*, 2019; ROCHA *et al.*, 2017; VENCKAUSKAS *et al.*, 2015; BOUANANI; KASSOU, 2014; TAMBOLI; PRASAD, 2013; STAMATATOS, 2009; HOLMES, 1985). De modo a delinear os conceitos e definições das diferentes linhas de pesquisa da AA, a próxima seção sintetiza os principais métodos, terminologias e limitação da AA.

3.2 Conceitos e métricas da AA

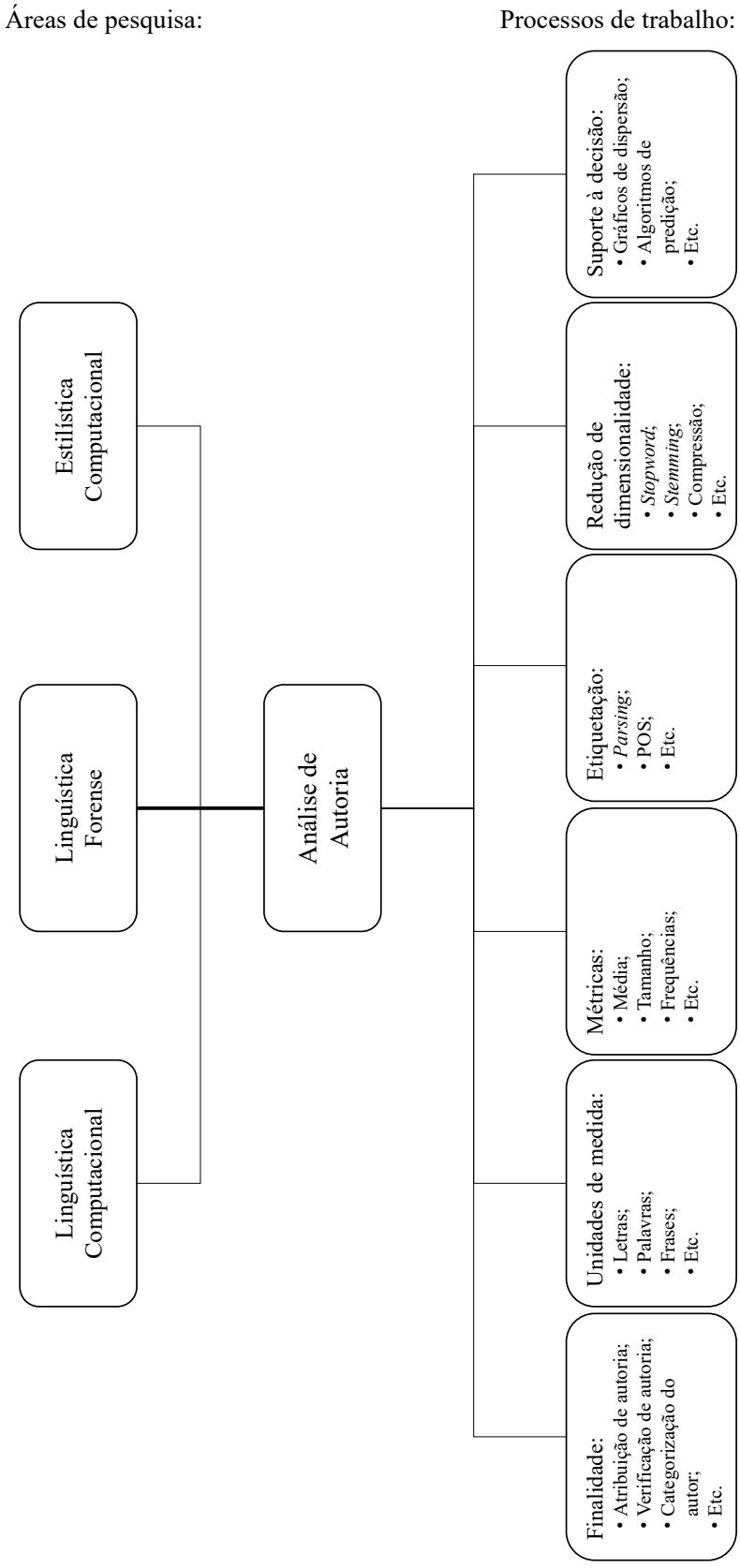
A AA não está centralizada em uma única linha de pesquisa; ela aparece na Linguística Forense, na Estilística Computacional e na Linguística Computacional. Apesar de essas linhas de pesquisas terem objetivos próprios, elas compartilham métodos, métricas e conceitos da AA (Figura 3.1).

Independentemente da linha de pesquisa, as análises de autorias alinham-se a uma finalidade de trabalho (Figura 3.1). Por sua vez, a finalidade delineia o objetivo de uma investigação e determina o modo como os dados serão coletados (Tabela 3.1). Na Tabela 3.1 estão descritos os processos de trabalho ilustrados na Figura 3.1.

Para quantificar as características dos autores, as pesquisas de AA usam unidades de medidas extraídas dos textos (e.g. letras, palavras, pontuações etc.). Essas unidades de medidas, combinadas com processos de metrificações (e.g. médias, totais etc.), geram as características dos autores. Unidades de medida e métricas confiáveis são aquelas que: (i) geram valores discrepantes para distinguir as obras de autores diferentes; (ii) geram valores semelhantes quando as obras pertencem a um mesmo autor. A Tabela 3.2 exibe alguns métodos de metrificação de textos.

[8]Neste trabalho, considera-se que as métricas convencionais estão relacionadas às medidas quantitativas (e.g. totais, médias etc.) de palavras ou letras extraídas dos textos.

Figura 3.1: Visão geral das áreas de pesquisas e dos processos ligados à AA



Fonte: Elaboração própria

Tabela 3.1: Finalidades dos estudos nas análises de autorias

Finalidade	**Definição**
Atribuição de autoria ou Identificação de autoria	Depois de estudar as obras de diferentes autores, a identificação de autoria calcula a probabilidade de um texto com autoria desconhecida pertencer a autores estudados previamente (ZHANG *et al.*, 2006). A partir de um conjunto de características estilísticas coletadas anteriormente, a atribuição de autoria infere o autor de um texto sem autoria conhecida (LAGUTINA *et al.*, 2019).
Categorização de autoria ou Caracterização do autor ou Perfil do autor	A categorização de autoria usa antecedentes culturais ou educacionais dos escritores para caracterizar o perfil de cada um deles (VEL, 2000; ZHANG *et al.*, 2006; DING *et al.*, 2016).
Detecção de mudança de estilo	Analisa a mudança de estilos em diferentes partes de um documento (LAGUTINA *et al.*, 2019).
Detecção de plágio	O foco principal desse estudo é analisar diferentes partes de um texto para detectar se são semelhantes a partes do texto de outros autores (ZHANG *et al.*, 2006; BOUANANI; KASSOU, 2014).
Detecção de similaridade ou Detecção de semelhança	Sem identificar o autor, o estudo analisa diferentes partes de um texto para determinar se foram feitas pela mesma pessoa (ZHANG *et al.*, 2006).
Determinação da intenção do autor	Kuzu & Salah (2015) atribuem esse termo a estudos que procuram propriedades ou características produzidas intencionalmente em um conjunto de textos ou documentos (e.g. para identificar a inclusão de códigos maliciosos em *scripts* de computador).
Discriminação de autoria	Sem identificar a autoria, o estudo procura determinar se um texto foi escrito por um ou por muitos autores (GRAY; SALLIS; MACDONELL, 1997).
Identificação do tradutor	Trabalha com a identificação de autoria em traduções (ELFIQI; PETRAKI; ABBASS, 2019).
Reconhecimento de autoria	Kuzu & Salah (2015) categorizam o reconhecimento de autoria (*Authorship recognition*) como uma subdivisão de uma categoria maior chamada atribuição de autoria.
Verificação de autoria	São estudos sobre a decisão de atribuir um autor à um determinado texto, geralmente, com resposta binária (e.g. "sim, o texto pertence a este escritor"ou "não, o texto não pertence a este escritor") (LAGUTINA *et al.*, 2019).

Fonte: Elaboração própria

Tabela 3.2: Métodos para metrificação de textos

Método	**Descrição**
Diversidade de vocabulário ou Riqueza de vocabulário	O pressuposto é que cada escritor trabalha com um conjunto finito de palavras e que esse conjunto é diferente dos conjuntos finitos dos outros escritores; então, ao extrair amostras de palavras de diferentes partes dos textos de um autor, obtêm-se um conjunto de palavras que define o autor (ELEWA, 2019; HOLMES, 1985).
Erros ou Estruturas textuais	Este tipo de metrificação é muito usado na Linguística Computacional e na Linguística Forense (LAGUTINA *et al.*, 2019; ROCHA *et al.*, 2017; TAMBOLI; PRASAD, 2013); trata-se de identificar padrões no modo como uma pessoa escreve e usá-los para distinguir os escritores; por exemplo, em "*João gritou: 'Mariaaaa, Mariaaa'. E Maria respondeu: 'Falaaa, Joãooo'.*", a repetição de vogais no final das palavras pode ser usada como uma métrica para caracterizar um escritor.
hápax legómenon[i]	São palavras que um escritor usa apenas uma única vez em um texto ou em toda sua obra. Essas palavras são usadas para caracterizar o escritor (REXHA *et al.*, 2007; PAVELEC *et al.*, 2009). Um texto de autoria desconhecida que repete muitas vezes os *hápax* de um escritor x, provavelmente, não foi escrito por x.
Métricas relacionadas a caracteres	São métricas (e.g. distribuição de frequência, média etc.) calculadas a partir das ocorrências de determinados caracteres nos textos. A métrica mais conhecida é a *n-gramas* que calcula o número de ocorrências de cada n números de caracteres (e.g. n = 1 (um caractere); n = 2 (dois caracteres) etc.) (ROCHA, 2019; JAMIL; MUSTAFA, 2018; MARKOV; BAPTISTA; PICHARDO-LAGUNAS, 2017). Podem ser usados caracteres alfanuméricos, pontuações, caracteres maiúsculos ou minúsculos, caracteres especiais etc. Villar-Rodriguez *et al.* (2016), por exemplo, usaram *emoticons*[ii] como unidade de medida na AA.
Tamanho das sentenças ou Comprimento das sentenças	Calcula diferentes métricas relacionadas à quantidade de palavras nas frases dos textos como a média de palavras por frase, a razão entre o número de palavras e o total de palavras do texto etc. (VOROBEVA, 2016; SOUSA-SILVA *et al.*, 2010)

Método	Descrição
Partes do discurso[iii]	É o processo de rotulação de palavras conforme uma categoria (e.g. classes de palavras, classificações sintáticas etc.) (KERNOT; BOSSOMAIER; BRADBURY, 2018; GOMEZ-ADORNO *et al.*, 2016; SOUSA-SILVA *et al.*, 2010). Nesses casos, as medições são feitas a partir das partes do discursos e não mais das palavras.
Palavras funcionais ou palavras gramaticais	São palavras com poucos sinônimos e/ou baixos valores semânticos. Elas, geralmente, executam funções gramaticas nos textos (e.g. artigo, preposição etc.). A partir de palavras funcionais são calculadas frequências, médias, pares de combinações etc. (MARINHO, 2017; VARELA, 2017)

[i]De origem grega, a expressão significa: "disse uma vez".
[ii]São imagens ou combinações de caracteres usados para representar objetos, emoções ou ações humanas.
[iii]Tradução do inglês: *Parts of Speech* (POS).

Fonte: Elaboração própria

Na Tabela 3.2, estão descritos alguns métodos de metrificações da AA. De acordo com Rudman (1998), existem mais de mil métodos para metrificação de textos. Ao metrificar as características dos autores, a AA cria condições para evidenciar onde estão as diferenças nos textos dos autores e identificá-los.

Em uma análise de autoria, as pré-definições das unidades de medida e das métricas influenciam o modo como os dados serão coletados e organizados. Para que o computador possa reconhecer e processar frases e palavras dos textos, alguns procedimentos devem ser feitos; por exemplo, é durante o processo de trabalho de Etiquetação (Figura 3.1 e Tabela 3.3) que o computador extrai as palavras dos textos e as organiza em estruturas a partir das quais algoritmos de processamento de textos conseguem interpretá-las.

Tabela 3.3: Procedimentos usados para identificação e rotulação de palavras e frases

Nome	Descrição
Analisador sintático ou *Parsing*	O analisador sintático identifica as palavras de uma frase e retorna, em uma estrutura hierárquica (i.e. uma árvore sintática), as classificações sintáticas das palavras da frase.
Tokenização	No processo de tokenização, as palavras das frases são segmentadas e armazenadas em diferentes subconjuntos. A sentença "João gosta de ler.", após passar pela tokenização, passa a ser $p = \{"João", "gosta", "de", "ler"\}$.

Nome	Descrição
Etiquetador gramatical ou *POS tagger*	O etiquetador gramatical é responsável por identificar e classificar os *tokens* (unidades mínimas) de uma frase. O etiquetador, por exemplo, ao receber a frase "Maria está feliz", retorna o conjunto $t = \{"\#s9",\ "\#v1",\ "\#n2"\}$ e o conjunto $w = \{"maria",\ "está",\ "feliz"\}$ (DUQUE; CARVALHO; VIMIEIRO, 2019). No primeiro conjunto, estão as classificações das palavras (configuradas conforme pré-cadastros no etiquetador); no segundo, estão as palavras da frase.
Símbolos de parada ou *Stopper symbols*	Para segmentar automaticamente as frases de um texto, o método de símbolos de parada usa caracteres como ",", ".", "...", "?"e "!" para identificar o fim de uma frase e o início de outra. No texto, "João é forte? Eu acho que não.", o método gera o conjunto $p = \{"joão\ é\ forte",\ "eu\ acho\ que\ não"\}$.

Fonte: Elaboração própria

A Tabela 3.3 sintetiza alguns procedimentos usados para organizar os dados textuais. Nela, os procedimentos apresentados são combinados de maneira a atender a natureza de cada pesquisa. É a natureza da pesquisa que determina quais unidades de medida, métricas e processamentos textuais serão usados.

Em pesquisas de natureza lexical (VARELA; ALBONICO; ASSIS, 2019; GALINA; FLORES; KOMATI, 2019; SAVOY, 2018; ELEWA, 2019; CARDEI; REBEDEA, 2017), as palavras e os seus atributos (e.g. classes de palavras, classificações sintáticas, propriedades morfológicas, *hápax* etc.) são o foco dos estudos. As análises são feitas com base nos repertórios de palavras usadas pelos autores.

Em pesquisas baseadas em caracteres (OLIVEIRA, 2019; JAZILAH, 2019; LE; SAFAVI-NAINI, 2018; MARKOV; BAPTISTA; PICHARDO-LAGUNAS, 2017), a extração dos dados é feita de maneira diferente das descritas na Tabela 3.3. Nas análises de caracteres, os números, as letras e os caracteres especiais são extraídos e organizados de forma a facilitar as respectivas mensurações de autorias. Nessas análises, a crença é que os autores usam proporções de caracteres diferentes entre si (e.g. o autor A usa mais vezes o caractere "w" do que o autor B).

Em estudos de natureza semântica (MARTINS *et al.*, 2019; KERNOT; BOSSOMAIER; BRADBURY, 2018), os pesquisadores organizam os conteúdos semânticos das palavras em diferentes categorias e as analisam de modo a encontrar elementos que possam distinguir textos escritos por autores diferentes. Em Martins *et al.* (2019), os autores categorizaram

as palavras com base em sentimentos (e.g. surpresa, tristeza etc.) e em seguida realizaram as análises de autorias considerando os sentimentos encontrados nos textos.

Em análises de conteúdos específicos, os pesquisadores trabalham com conjuntos de palavras que são mais prováveis em contextos específicos (i.e. um domínio); por exemplo, em propagandas de vendas de veículos, palavras como "porta" e "roda" são mais prováveis do que "lápis" ou "apontador" (ADAMOVIC *et al.*, 2019; ABUHAMAD *et al.*, 2019; REXHA *et al.*, 2007; BOUANANI; KASSOU, 2014; STAMATATOS, 2009). Como esses conjuntos de palavras têm aplicações específicas, eles não servem para outros propósitos que não o do estudo proposto. Esse tipo de abordagem é usado em estudos relacionados a crimes cibernéticos, detecção de *spams*[9], por exemplo (BARBON *et al.*, 2018; CARDEI; REBEDEA, 2017; ZHANG; LIU; CHEN, 2015; DEVI; RAVI, 2015).

Quando as quantidades de dados extraídas dos textos são muito grandes ou quando a estratégia é analisar um determinado tipo de palavra, as análises de autorias recorrem a métodos que reduzem o tamanho do *corpus*. Na Tabela 3.4, está descrito o funcionamento de alguns métodos usados para a redução de dimensionalidade dos dados.

Tabela 3.4: Processos usados na redução de dimensionalidade de textos

Nome	**Descrição**
Compressão de textos	Comprimir ou compactar um arquivo implica reduzir o seu tamanho original e salvá-lo com um novo formato. Algoritmos que compactam arquivos buscam combinações de caracteres frequentes e as trocam por combinações de menor tamanho. Esses conjuntos de caracteres modificados geram diferentes padrões no tamanho do arquivo comprimido (e.g. porcentual compactado) e, a partir desses padrões, a análise de autoria infere o autor de um texto suspeito. Os textos de um mesmo autor apresentam percentuais de compactação semelhantes entre si e diferentes dos percentuais observados na compactação de textos de outros autores. A análise da compressão de textos torna a atribuição de autoria independente de escolhas prévias de características relativas aos autores (OLIVEIRA-JUNIOR, 2011).
Remoção de números, pontuações e caracteres especiais	Consiste em remover dos textos caracteres especiais (e.g. "@", "&", "*"), números cardinais, ordinais etc. Outra abordagem é remover todas as palavras e manter no texto somente números e ou caracteres especiais. (KERNOT; BOSSOMAIER; BRADBURY, 2018; JAMIL; MUSTAFA, 2018)

[9]Propagandas enviadas por e-mail.

Nome	Descrição
Lematização	Processo usado para transformar as palavras em suas respectivas formas canônicas (e.g. "falaram" e "falou" são transformados em "falar"; "meninos" e "menininhos" são transformados em "menino") etc. (DUQUE; CARVALHO; VIMIEIRO, 2019; MARINHO; HIRST; AMANCIO, 2018; MARINHO; HIRST; AMANCIO, 2016; AMANCIO; SILVA; COSTA, 2015)
Palavras funcionais ou palavras gramaticais	São conjuntos de palavras que auxiliam o funcionamento das línguas (e.g. pronomes, preposições etc.). Nesse caso, a redução de dimensionalidade ocorre quando essas palavras são excluídas do *corpus* ou quando o *corpus* contém somente palavras funcionais. (ROZZ; MENEZES, 2018; VOROBEVA, 2016; AMANCIO, 2015a)
Stemização ou *stemming*	Processo usado para reduzir ou aproximar as palavras de suas respectivas raízes. As palavras "meninos", "meninas" e "menina"são transformadas em "menin" (i.e. sem as respectivas desinências nominais). (MARTINS *et al.*, 2019; REXHA *et al.*, 2007)
Stopword[i]	São conjuntos de palavras que aparecem muitas vezes nos textos e que, geralmente, têm pouco valor semântico. Retirar dos textos as *stopwords*, reduz o tamanho do documento e deixa no texto somente palavras com maiores valores semânticos. A remoção das *stopwords* na frase "as meninas daquela casa" deixa a frase assim "meninas casa". O conjunto de palavras que compõem a lista de *stopwords* é arbitrária e varia entre as pesquisas (MARTINS *et al.*, 2019; MACHICAO *et al.*, 2018; REXHA *et al.*, 2007).

[i]Alguns autores também a chamam de *noise words* (e.g. Arun, Suresh & Madhavan (2009).

Fonte: Elaboração própria

Apesar de diminuir a quantidade de informações, a redução de dimensionalidade promove algumas vantagens como a redução do tempo de processamento e a remoção de palavras que não seriam analisadas (LAGUTINA *et al.*, 2019; ROCHA *et al.*, 2017; ANWAR; BAJWA; RAMZAN, 2019; STAMATATOS, 2009).

Nas primeiras análises, as AAs usavam gráficos e métricas simples (e.g. médias, frequências, totais) para comparar e sugerir a quais autores os textos pertenciam. Com o passar do tempo, métodos de análises automáticas começaram a ser usados para inferir o autor de um texto suspeito. A Tabela 3.5 descreve alguns métodos usados para predições

automáticas em estudos sobre AA.

Tabela 3.5: Algoritmos de predição aplicados na análise de autoria

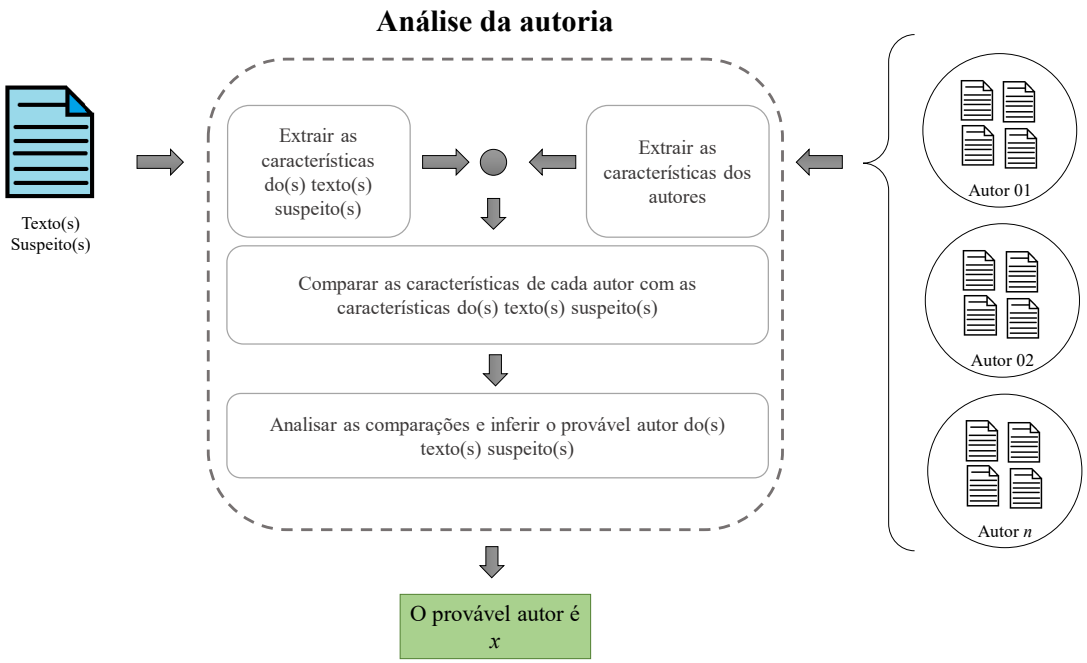
Algoritmos	**Descrição**
Análise de componentes principais ou *Principal Component Analysis* (PCA)	Resumidamente, o algoritmo usa um gráfico de dispersão para assentar os objetos e projetar os componentes (i.e. vetores). É a proximidade entre os objetos e os vetores que permite inferir as semelhanças entre os objetos. O PCA aparece nas análises de autorias dos estudos de Villar-Rodriguez *et al.* (2016); Valencia (2017) e Zhang *et al.* (2014).
K-vizinhos mais próximos (*K-Nearest Neighbors* - KNN)	Nesse algoritmo, os objetos de estudos são projetados em um espaço n-dimensional e, a partir da proximidade entre eles, o algoritmo determina a que grupo cada objeto pertence. Quanto mais próximo um objeto de classificação desconhecida estiver de vizinhos com classificação conhecida, maior a chance de o objeto desconhecido pertencer a classificação de seus vizinhos. O KNN aparece nas análises de autoria de Jazilah (2019), Adamovic *et al.* (2019) e Zhang *et al.* (2014).
Máquina de vetores de suporte ou *Support Vector Machine* (SVM)	Nesse algoritmo, os objetos também são projetados em um espaço n-dimensional, no entanto o algoritmo projeta uma linha/hiperplano entre os objetos periféricos e assim identifica os grupos de objetos de mesma classificação. O SVM aparece em Varela, Albonico & Assis (2019); ROCHA (2019) e Martins *et al.* (2019).
Naive Bayes	O algoritmo de Naive Bayes classifica os dados com base em cálculos de probabilidade. Basicamente, o algoritmo usa dados de treinamentos para formar as características dos objetos e, posteriormente, inferir a probabilidade de um determinado dado pertencer a uma das classes treinadas. Coyotl-Morales *et al.* (2006); Markov, Baptista & Pichardo-Lagunas (2017); Vazirian & Zahedi (2016) usaram Naive Bayes na AA.
Redes Neurais Artificiais (RNA)	São algoritmos inspirados em estruturas neurológicas que aprendem à medida que treinam; basicamente, a rede neural recebe sinais (dados) que são multiplicados por um número (peso) e, depois de uma soma ponderada de sinais, o algoritmo avalia se deve emitir uma resposta ou não para o conjunto de sinais de entrada. Autores que trabalharam com RNA na AA: Zhang *et al.* (2006); Brennan & Greenstadt (2009); Abuhamad *et al.* (2019).

Fonte: Elaboração própria

Os algoritmos apresentados na Tabela 3.5 reorganizam e analisam os dados coletados de modo a encontrar um padrão que possa ser usado para distinguir e determinar quais objetos têm as mesmas características. Os algoritmos de aprendizados supervisionados (e.g. Naive Bayes, SVM) preveem os nomes dos objetos[10] com base em diferentes dados [11] previamente conhecidos; os algoritmos de aprendizados não supervisionados (e.g. PCA) classificam os objetos com base nas associações e padrões observados nos dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2017).

Nas análises de autorias, os algoritmos de aprendizados (e.g. Tabela 3.5) são ferramentas que o pesquisador usa para explorar os dados e evidenciar que um texto pertence a um autor específico. Os percentuais de acertos dos métodos estão condicionados a diferentes variáveis como: tipo de texto (e.g. formal ou informal); unidades de medida; métricas etc. (ver a Seção 3.3). Embora a construção de um método de análise de autoria envolva a combinação de diferentes métricas e variáveis, as pesquisas em AA seguem um modelo de análise comum (Figura 3.2).

Figura 3.2: Modelo básico de análise de autoria



Fonte: Elaboração própria

A Figura 3.2 sintetiza um modelo de análise que aparece na Linguística Forense, na Estilística Computacional e na Linguística Computacional. No modelo (Figura 3.2), as características dos textos e dos respectivos autores são extraídas e comparadas com base

[10] Nas análises de autorias, os objetos são os autores.
[11] Nas análises de autoria, os dados são as características dos autores.

em unidades de medidas, métricas e métodos de análises coerentes com as respectivas linhas de pesquisas.

A Linguística Forense é uma linha de pesquisa interdisciplinar ligada à Linguística Aplicada que relaciona, principalmente, estudos da Linguística com estudos do Direito para evidenciar (i.e. apontar provas periciais) a que autor específico um texto suspeito pertence. Na resolução de conflitos jurídicos relacionados a autorias de textos, é o resultado do trabalho do perito em AA que servirá de evidência para o advogado apresentar no tribunal (CALDAS-COULTHARD, 2014).

A Estilística Computacional é uma linha de pesquisa que investiga, principalmente, os estilos nos textos. A análise estilística trabalha com a crença de que existem conjuntos de padrões que diferenciam textos jornalísticos de textos literários, assim como, existem conjuntos de padrões que diferenciam um escritor de outro. Esses conjuntos de padrões que diferenciam textos e escritores são chamados de *estilo*. Em algumas pesquisas (ROCHA *et al.*, 2017; ANDRADE, 2015; DAROONEH; SHARIATI, 2014), o estilo é estudado como uma característica observada na escrita de um indivíduo ou de um grupo; ele é composto por jargões, estruturações textuais, erros de ortografia e uso de variedades de uma língua.

Em outras pesquisas, o estilo foca as características dos textos para classificá-los conforme as escolas literárias (BARUFALDI *et al.*, 2010), conforme as características das línguas naturais (AMANCIO, 2013; CALDEIRA, 2005), com base em aspectos de formalidade e informalidade (SISOVIC; MARTINCIC-IPSIC; MESTROVIS, 2014) etc. Na análise estilística, o pesquisador deve estar ciente de que cada escritor, conscientemente ou subconscientemente, usa um repositório de palavras finitas em conformidade com diferentes contextos; cabe ao pesquisador organizar e descrever, quantitativamente, o funcionamento desses repertórios de modo que possam ser usados para diferenciar os escritores (HOLMES, 1985).

Na Linguística Computacional, o principal objeto de pesquisa é a língua natural. Como recurso para analisá-la, matemática e estatisticamente, é comum que discursos orais ou conversas sejam transcritas. Essa estratégia de transcrição auxilia a metrificação de elementos relacionados à semântica, à sintaxe, ao estilo etc. A metrificação das características dos textos produzidos pelos falantes auxilia, por exemplo, o pesquisador na identificação dos falares regionais (LAGUTINA *et al.*, 2019).

3.3 Premissas teóricas da análise de autoria em textos digitais

Nesta seção, estão descritas as principais vantagens, desvantagens e limitações observadas nas métricas e nos métodos de pesquisas. Esses tópicos fazem parte do conjunto teórico

da AA e delineiam os caminhos já percorridos pelas pesquisas.

O primeiro destaque a ser feito diz respeito às medições por n-gramas; tais medições, ainda que apresentem certas limitações, estão entre as métricas mais precisas[12] da análise de autoria de textos digitais (LAGUTINA *et al.*, 2019; JAMIL; MUSTAFA, 2018; MARKOV; BAPTISTA; PICHARDO-LAGUNAS, 2017; LUYCKX, 2010; GRIEVE, 2007). As vantagens dos n-gramas são: a facilidade de manuseio; aplicam-se a qualquer língua; não é necessário conhecimento linguístico prévio; e são flexíveis em relação ao número de caracteres trabalhados (LAGUTINA *et al.*, 2019; VENCKAUSKAS *et al.*, 2015). As desvantagens ou limitações são: os n-gramas têm piores resultados quando o número de autores analisados passa de cinquenta (LUYCKX, 2010); n-gramas de menores comprimentos são mais precisos que os de maiores comprimentos (GRIEVE, 2007).

Em segundo lugar, é relevante observar que a combinação de diferentes unidades de medidas (e.g. letras e palavras) é uma estratégia que melhora a predição de autoria (SOUSA-SILVA *et al.*, 2010; MARGARIDO *et al.*, 2008; GRIEVE, 2007). Usar mais de uma métrica de quantificação melhora a precisão dos algoritmos de identificação de autoria (LAGUTINA *et al.*, 2019; LUYCKX; DAELEMANS, 2011; ZHANG *et al.*, 2006). Embora resulte em melhor precisão, a combinação de métricas implica ocupar mais espaço em disco e maior tempo de processamento (ROCHA *et al.*, 2017).

Em terceiro lugar, destaca-se que a passagem do tempo deixa marcas no modo como as pessoas escrevem (HOLMES, 1985). Can & Patton (2004) observaram que o comprimento das palavras de dois escritores turcos aumentou com o passar dos anos. As influências dos contextos e das épocas podem ser percebidas no trabalho de Barufaldi *et al.* (2010) que conseguiram relacionar um conjunto de textos de diferentes autores às suas respectivas escolas literárias.

Em quarto, ressalta-se o fato de que as línguas dos textos também devem ser objeto de atenção. A precisão dos algoritmos usados para atribuir autoria mudam conforme a língua dos textos analisados (QUINTANILLA, 2020; ADAMOVIC *et al.*, 2019; HALVANI; WINTER; PFLUG, 2016; ZHANG *et al.*, 2006). Um algoritmo que é preciso na atribuição de autoria para textos em inglês pode não ser para textos em espanhol, por exemplo. Uma possível razão para isso acontecer é que as métricas escolhidas para a análise das autorias geram características mais marcantes[13] em uma língua do que em outras (VARELA, 2017; HALVANI; WINTER; PFLUG, 2016).

Em quinto, é relevante considerar que as traduções de documentos também requerem especial atenção na análise de autoria. Os tradutores usam estilos próprios e, dessa forma,

[12] Métrica que mais vezes acertou o real autor de um texto suspeito.
[13] Características que distinguem os autores com eficiência.

um texto traduzido por diferentes tradutores terá diferentes estilos (ANDRADE, 2015). Várias traduções de obras de um autor não têm semelhanças entre si e nem é possível classificá-las como pertencentes a um mesmo autor (ARUN; SURESH; MADHAVAN, 2009). As características que ligam os textos a seus respectivos autores são obstruídas quando o texto é traduzido para outra língua (ARUN; SURESH; MADHAVAN, 2009; ANDRADE, 2015). As marcas que um tradutor deixa nos textos que traduz podem ser detectadas com métodos tradicionais de análise de autoria ou com o suporte da teoria de redes complexas (EL-FIQI; PETRAKI; ABBASS, 2019).

Em sexto lugar, é necessário encontrar qual é o tamanho mínimo de um texto (i.e. total de palavras) para que seja possível realizar uma análise de autoria. Esse é um tema bastante estudado na AA (VARELA; ALBONICO; ASSIS, 2019; LUYCKX, 2010; HIRST; FEIGUINA, 2007). Conhecer esse valor é importante por diferentes motivos; saber o tamanho mínimo permite ao pesquisador, antes mesmo de começar a análise, saber se o texto suspeito tem as características mínimas de tamanho para que uma atribuição de autoria seja feita. Conhecer esse limite também evita que uma grande quantidade de dados seja coletada desnecessariamente. Diferentes valores mínimos já foram sugeridos: próximos de trezentas palavras (HIRST; FEIGUINA, 2007); quinhentas (VARELA; ALBONICO; ASSIS, 2019); mil (HONORIO *et al.*, 2007; MOSTELLER; WALLACE, 1964; MENDENHALL, 1887). Esses limites mínimos variam conforme os contextos. Em geral, a AA de documentos provenientes de contextos informais requer um número menor de palavras se comparada à AA de textos provenientes de contextos formais.

Em sétimo, deve-se observar que nas línguas naturais é possível organizar as palavras em dois grupos: as palavras de natureza lexical e as palavras de natureza gramatical/funcional. As palavras funcionais apresentam boa precisão quando usadas na verificação de autoria. Conjuntos específicos de palavras funcionais distinguem bem os padrões de pares de escritores (LUYCKX; DAELEMANS, 2011; KOPPEL; SCHLER; ARGAMON, 2009). Hollingsworth (2012), ao comparar a performance da identificação de autoria entre n-gramas e palavras funcionais, observou que as análises com palavras funcionais obtiveram mais acertos. A desvantagem desse tipo de análise é que um conjunto de palavras funcionais que é relevante para distinguir um par de autores, pode ser imprecisa na distinção de outros pares (LUYCKX; DAELEMANS, 2011).

Por fim, os algoritmos de tomada de decisão, em geral, têm enfrentado um problema: quanto mais autores comparados na análise de autoria, menor é a sua precisão (LUYCKX; DAELEMANS, 2011; LUYCKX, 2010; ZHANG *et al.*, 2006). No entanto, entre esses algoritmos, o SVM destacou-se nas pesquisas que o compararam com outros algoritmos por ter atingido um maior número de predições corretas (JAZILAH, 2019; MARKOV; BAPTISTA; PICHARDO-LAGUNAS, 2017; BOUANANI; KASSOU, 2014; LUYCKX, 2010).

Dessa forma, a partir dos pontos expostos acima, esta tese assumirá como parâmetros: usar mais de uma métrica de quantificação; considerar conjuntos de textos produzidos em um mesmo contexto; considerar textos de mesma língua e que não sejam traduções; comparar textos com quantidades semelhantes de palavras; considerar a natureza das palavras (gramaticais e lexicais); considerar algoritmos de predição que, em outras pesquisas, obtiveram altos índices de acertos nas autorias.

Capítulo Quatro

Análises de autorias com redes

Uma rede é uma representação de um sistema, em uma estrutura abstrata e simplificada, que possui entidades que se conectam segundo algum critério (NEWMAN, 2010). Nas redes, os relacionamentos entre entidades individuais são usados para extrair informações do conjunto dos relacionamentos e de relações entre as entidades. A AA, ao transformar um texto em uma rede, explora interações entre palavras e observa padrões que são mais difíceis de manipular e falsificar (MACHICAO *et al.*, 2018).

Este capítulo apresenta, em três seções, uma breve descrição da Teoria e Ciência das Redes e sua aplicação nas análises de autorias. Na Seção 4.1, estão descritos, brevemente, conceitos, métricas e características das redes. Na Seção 4.2, apresentam-se as relações existentes entre as análises de autorias e as redes. Na Seção 4.3, estão descritos conceitos e premissas teóricas das redes nas análises de autoria considerados relevantes para esta tese.

4.1 A Teoria e Ciência das Redes

Os estudos sobre redes e grafos têm origem atribuída ao matemático Leonhard Euler que encontrou uma solução para o problema de Konigsberg[1]. Graficamente, uma rede (e.g. Figura 4.1(a)) contém objetos conectados por linhas. Os objetos são chamados de vértices e, geralmente, são representados por círculos; as conexões entre os vértices podem ser denominadas arestas (graficamente, representadas por linhas) ou arcos (quando representadas por setas). Usam-se arestas quando não importa a direção dos relacionamentos entre os vértices e arcos quando os relacionamentos têm origens e destinos definidos.

A representação de um sistema por meio de vértices e conexões causa a redução de muitas informações do sistema completo. Entretanto, ao atribuir valores aos vértices e às suas conexões, as redes criam condições para que as relações entre os objetos do sistema sejam analisadas de diferentes maneiras (NEWMAN, 2010).

Historicamente, as redes já foram usadas para: modelagem matemática de dispersões em redes (e.g. doenças) onde os vértices conectavam-se aleatoriamente (SOLOMONOFF; RAPOPORT, 1951); análise de características determinísticas de redes aleatórias (ERDOS; RENYI, 1959); na teoria de um *mundo pequeno*, isto é, estudos relacionados às

[1] Encontrar um percurso na cidade de Konigsberg de modo que uma pessoa passe pelas sete pontes locais sem ter que atravessar a mesma ponte mais de uma vez.

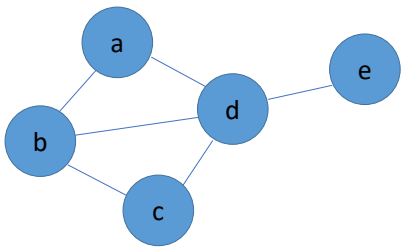
distâncias entre as pessoas em uma grande rede de contatos pessoais (TRAVERS; MILGRAM, 1969); na identificação de blocos (i.e. conjuntos) de pessoas com interesses em comum (WHITE; BOORMAN; BREIGER, 1976); na verificação das relevâncias de periódicos por meio de redes de citação (DOREIAN, 1985); na caracterização topológica de redes aleatórias, *mundos pequenos* e regulares (WATTS; STROGATZ, 1998); na descrição de comportamentos de redes reais (e.g. transmissão de doença, internet etc.) que dinamicamente atribuem mais conexões para os vértices preferenciais e mais importantes do que para os vértices com pouca importância (BARABASI, 2009); estudos sobre a caracterização de línguas (SOLE *et al.*, 2010; MASUCCI; RODGERS, 2006; CONG; LIU, 2014); análise de discursos por meio de redes semânticas (LIMA-NETO; CUNHA; PEREIRA, 2018) etc. Ao longo dos anos, os estudos sobre redes acumularam conhecimentos técnicos capazes de extrair diferentes tipos de informações de diferentes tipos de sistemas.

4.1.1 *Estruturas, propriedades e atributos das redes*

A representação matemática básica de uma rede ou um grafo é $G = (N, E)$, sendo N o conjunto de vértices (e.g. $N = \{v_1, v_2, ..., v_n\}$) e E o conjunto de arestas (e.g. $E = \{e_1, e_2, ..., e_m\}$, sendo $e_k = (i, j), v_i \neq v_j$). Uma maneira de armazenar e trabalhar com uma rede é organizar os vértices e as conexões em uma matriz de adjacência ou uma lista. Em uma matriz de adjacência (e.g. Figura 4.1 (b)), as linhas e as colunas representam os vértices da rede; as intersecções entre as linhas e as colunas representam a existência, ou não, de conexões entre os vértices. Em uma matriz A de tamanho N x N, se houver um relacionamento entre os vértices v_i e v_j, a célula a_{ij} da matriz será igual a um; em caso contrário, será igual a zero.

Figura 4.1: Representação gráfica de uma rede e respectiva matriz de adjacência

(a) Representação gráfica de uma rede



(b) Matriz de adjacência

	a	*b*	*c*	*d*	*e*
a	0	1	0	1	0
b	1	0	1	1	0
c	0	1	0	1	0
d	1	1	1	0	1
e	0	0	0	1	0

Fonte: Elaboração própria

A Figura 4.1 (a) exibe uma rede com cinco vértices ($n = |N| = 5$), rotulados de a a e

(N=*{a,b,c,d,e}*) e seis arestas (m=$|E|$=*6*; M=*{{a,b}, {b,c}, {b,d}, {c,d}, {a,d}, {d,e}}*). Para extrair informações das redes, as equações matemáticas e os algoritmos percorrem as células das matrizes e extraem diferentes métricas. O grau, por exemplo, é uma métrica que registra o total de arestas conectadas a um dos vértices de uma rede não dirigida[2]. O grau de um vértice é calculado pela Equação 4.1.

$$k_i = \sum_{j=1}^{n} a_{ij} \tag{4.1}$$

onde os graus dos vértices *{a, b, c, d, e}* são, respectivamente, k = *{2, 3, 2, 4, 1}*. O grau médio da rede é calculado por $<k>$ = $2m/n$.

Em redes dirigidas[3], o grau de entrada e o grau de saída contabilizam, respectivamente, os totais de arcos que chegam a um vértice e os totais de arcos que saem de um vértice (Equações 4.2 e 4.3).

$$k_i^{saida} = \sum_{j=1}^{n} a_{ij} \tag{4.2}$$

$$k_j^{entrada} = \sum_{i=1}^{n} a_{ij} \tag{4.3}$$

Nas redes ponderadas, os arcos e arestas registram os pesos das relações entre os vértices. Matematicamente, a rede ponderada é representada por $G^w = (N, E, W)$, onde $W = \{w_1, w_2, ..., w_E\}$ representa o conjunto pesos das conexões. Quando a rede ponderada também é dirigida, a_{ij} representa a relação entre os vértices $v_i \rightarrow v_j$ e w_{ij} registra o peso da conexão entre v_i e v_j. O grau ponderado de um vértice (Equação 4.4) mede a soma dos pesos dos arcos conectados ao vértice, o grau ponderado de entrada mede a soma dos pesos dos arcos que chegam a um vértice e o grau ponderado de saída mede a soma dos pesos dos arcos que saem de um vértice.

$$G_p(i) = \sum_{j=1}^{n} w_{ij} \tag{4.4}$$

As métricas de graus são úteis para comparar os vértices dentro de um cenário específico.

[2] Nas redes não dirigidas, as conexões entre os vértices são feitas por arestas.
[3] Nas redes dirigidas as conexões entre os vértices são feitas por arcos.

Em redes sociais, por exemplo, o vértice com maior número de arestas terá mais acesso a informações do que os vértices com menores números de arestas (NEWMAN, 2010).

Nas redes, as distâncias entre pares de vértices de um mesmo componente[4] são calculadas contando-se a quantidade de conexões que os separam. Na Figura 4.1(a), a distância entre os vértices a e b é igual a 1 e a distância entre os vértices a e e é igual a 2. Quando muitos caminhos conectam dois vértices, a distância a considerar entre eles deve ser a com menor número de conexões.

A centralidade de intermediação (ou só intermediação) mede o potencial de um vértice em controlar a troca de informações em uma rede (FREEMAN, 1978). A Equação 4.5 $C_B(i)$ representa a centralidade de intermediação do vértice i em uma rede não dirigida e apresenta seu valor relativo (i.e. normalizado):

$$C_B(i) = \frac{2}{(n-1)(n-2)} \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}} \tag{4.5}$$

onde $g_{jk}(i)$ é o número de caminhos entre j e k que passam por i e g_{jk} é o número de caminhos que ligam j e k, ou seja, a centralidade de intermediação é a proporção entre os menores caminhos entre j e k que passam por i e todos os caminhos entre j e k. Os vértices que aparecem muitas vezes no caminho entre outros pares de vértices possuem relevante influência nas redes de informações, mensagens e notícias porque eles controlam as informações que são repassadas (NEWMAN, 2010).

A excentricidade (Equação 4.6) afere qual é a máxima distância entre um vértice u a qualquer outro vértice v da rede. Ela mede a eficiência com que um vértice pode disseminar informações na rede; vértices com excentricidades de baixos valores denotam melhor eficiência de disseminação (VAZIRIAN; ZAHEDI, 2014).

$$E_u = max\{d(u,v)|u \neq v, v \in N\} \tag{4.6}$$

A centralidade de Laplace é uma métrica que atende às redes ponderadas. Nas redes não dirigidas, a importância de um vértice (i.e. a centralidade) é percebida ao se analisar a queda da energia Laplaciana ($E_L(G) = \sum_i \lambda_i^2$, sendo λ_i autovalores da matriz Laplaciana da rede ponderada G) da rede em resposta à exclusão de vértices (QI *et al.*, 2012).

As métricas das redes ajudam a evidenciar e formalizar comportamentos nos sistemas que, muitas vezes, não são facilmente visíveis quando se observa o sistema completo. Os

[4] Conjunto de vértices conectados de modo que sempre existe um caminho entre qualquer par de vértices.

estudos científicos das redes têm natureza interdisciplinar e ao combinar sistemas e ideias de diferentes áreas, como a biologia ou a sociologia, propiciam que novos métodos de análises e novos pontos de vista sejam descobertos (NEWMAN, 2010).

4.2 Análises de autorias e redes

Estruturalmente, as línguas podem ser estudadas em diferentes níveis. A combinação de sons para a formação das palavras é explicada por regras fonéticas e fonológicas; a estruturação interna das palavras ocorre por meio de regras morfológicas; a organização das palavras em sequências sentenciais ocorre por meio de regras sintáticas; e a construção dos significados passa pela semântica e pela pragmática. Ao considerar que as línguas podem ser pesquisadas a partir de diferentes níveis e que a Teoria e Ciência das Redes apresenta expertise na análise e métricas de sistemas do mundo cotidiano (e.g. internet, redes sociais), parece plausível usar a Teoria e Ciência das Redes na análise da linguagem humana (CONG; LIU, 2014).

As análises de textos por meio de redes começam com a transformação de um conjunto de documentos (e.g. livros, mensagens de textos, discursos orais transcritos) em redes. Essas transformações seguem diferentes critérios de construções e geram diferentes tipos de redes (e.g. as redes de coocorrência de palavras, redes de palavras adjacentes, redes de palavras funcionais, redes semânticas etc) e diferentes métodos de construção (e.g. adição de cliques[5] ou adição de linhas).

Na construção de redes com adição de linhas (e.g. redes de coocorrência de palavas, redes de palavras adjacentes), as relações de antecedência entre as palavras dos textos são preservadas pela direção dos arcos nas redes. Na construção de rede com a adição de cliques, as relações de antecedência entre as palavras não são preservadas; nessas redes, as palavras (i.e. os vértices) de cada frase do texto são agrupadas em um subconjunto de vértices de modo que cada par de vértice tenha uma aresta os conectando. As redes de linhas podem ser melhores para os estudos de redes de palavras adjacentes.

Em redes de n-gramas, os vértices representam sequências de n caracteres e, geralmente, as conexões entre os vértices são feitas linearmente a partir da ordem das palavras nos textos; por exemplo, em um n-grama de tamanho três (n=3 ou tri-grama), a frase *"hoje é terça"* torna-se o conjunto de vértices $N = \{$*"hoj", "oje", "ter", "erç", "rça"*$\}$[6] (JAZILAH, 2019; EDER; RYBICKI; KESTEMONT, 2016). Nas análises de autorias propostas por Chakraborty & Choudhury (2016), as redes baseadas em n-gramas obtiveram o segundo

[5] Subconjunto de vértices de uma rede não dirigida no qual cada par de vértices se conecta por uma aresta.

[6] São considerados como vértices apenas as sequências de três caracteres que cabem dentro das palavras; por esse motivo o verbo "é"foi ignorado.

melhor percentual de acertos; perderam somente para as redes baseadas em estruturas sintáticas. Apesar de não estarem relacionadas com a morfologia, as análises de n-gramas tentam encontrar padrões comuns em um conjunto de palavras.

As redes sintáticas são construídas após um determinado pré-processamento do texto. Um algoritmo (e.g. analisador sintático, etiquetador gramatical) lê as palavras e frases do texto, identifica sintaticamente cada palavra e as substitui por uma classificação sintática (e.g. $N = \{SN, SV, V\}$[7]). Embora as redes sintáticas não resultem em árvores sintáticas[8], elas revelam, por meio de suas métricas e propriedades, padrões que auxiliam na classificação de línguas (LIU; XU, 2011) e em análises autorais (DAKS; CLARK, 2016; GOMEZ-ADORNO *et al.*, 2016).

O processo de construção de redes semânticas (AMANCIO; OLIVEIRA-JR; COSTA, 2012b; TEIXEIRA *et al.*, 2010; FADIGAS *et al.*, 2009; AGUIAR, 2009; STEYVERSA; TENENBAUM, 2005) prioriza a seleção de palavras com maiores valores semânticos; por esse motivo, é comum que *stopwords* e palavras gramaticais não entrem nas redes. Com métricas tradicionais (i.e. métricas que não são extraídas das redes), as análises de autorias com características sintático-semânticas são mais confiáveis que as análises com características baseadas em caracteres (RAMEZANI; SHEYDAEI; KAHANI, 2013). No reconhecimento de autoria por meio de redes semânticas, as características topológicas das redes apresentam potencial para distinguir estilos (AMANCIO; OLIVEIRA-JR; COSTA, 2012b; AGUIAR, 2009); e, quando autores apresentam estilos de escrita semelhantes, seus estilos podem ser diferenciados semanticamente (AMANCIO; OLIVEIRA-JR; COSTA, 2012b).

Nas redes baseadas em características gramaticais (e.g. redes de palavras funcionais, redes de *stopwords*), os vértices são usados para representar conjuntos de palavras e caracteres com características gramaticais semelhantes (e.g. palavras gramaticais, pontuações). Em redes de *stopwords* (ARUN; SURESH; MADHAVAN, 2009), os vértices são criados a partir de uma lista de palavras pré-definidas; a crença é que cada escritor, inconscientemente, forma as frases dos textos com base em um padrão único de *stopwords* (ARUN; SURESH; MADHAVAN, 2009). Em Stanisz, Kwapien & Drozdz (2019), nas análises de autorias, os vértices das redes incluíram pontuações como marcadores gramaticais de estilo. Nas *redes de palavras funcionais adjacentes*[9], os vértices representam palavras funcionais (i.e. palavras que auxiliam o funcionamento das línguas como preposições, por exemplo). Nessas redes, palavras com maiores valores semânticos são ignoradas. Segarra, Eisen & Ribeiro (2014) observaram, nas redes de palavras funcionais adjacentes, que autores contemporâneos são mais semelhantes entre si do que entre autores de escolas literárias mais distantes

[7] Sintagma Nominal (SN), Sintagma verbal (SV), Verbo (V) são classificações sintáticas dos elementos que constituem uma sentença (frase).

[8] Diagrama que representa as classificações sintáticas de uma frase em uma estrutura hierárquica do tipo árvore.

[9] Em inglês, *Function Word Adjacency Networks*.

no tempo.

A construção de rede de palavras (e.g. redes de coocorrência de palavras, redes de palavras adjacentes) é feita de diferentes maneiras, aplicando processos de lematização nos textos, removendo *stopwords* etc. A distinção entre os diferentes tipos de redes de palavras está nas escolhas dos métodos de extração e de construção das redes. Para Lahiri & Mihalcea (2013), há duas maneiras de se construir uma rede de palavras; em ambas, as conexões entre os vértices das redes ocorrem linearmente conforme as palavras aparecem nos textos. A diferença entre elas é que, no primeiro modo, os limites das frases são respeitados (i.e. a última palavra de uma frase não se conecta a primeira da frase seguinte); e, no segundo, a última palavra de cada frase se conecta com a primeira palavra da frase seguinte.

Em Aguiar (2009) e Caldeira (2005), a construção de redes de palavras ocorre transformando as palavras das frases em cliques. Em redes de palavras adjacentes (STANISZ; KWAPIEN; DROZDZ, 2019) e redes de coocorrência de palavras[10] (ROZZ; MENEZES, 2018; MARINHO; HIRST; AMANCIO, 2016; AMANCIO; OLIVEIRA-JR; COSTA, 2012b; MEHRI; DAROONEH; SHARIATI, 2012; AMANCIO *et al.*, 2011), as conexões entre as palavras (i.e. vértices) ocorrem linearmente como elas aparecem nos textos.

As diferenças entre esses estudos estão no emprego de pesos nas relações, uso de redes direcionadas, métodos particulares de transformação das palavras (e.g. lematização) etc. Há, ainda, os estudos que analisam as autorias a partir de: agrupamentos nas redes (DUQUE; CARVALHO; VIMIEIRO, 2019; SINA *et al.*, 2015; AREFIN *et al.*, 2014); comportamentos relacionados à teoria dos *automatos celulares*[11] (MACHICAO *et al.*, 2018) etc.

Cada tipo de rede avalia as linguagens, os estilos e os autores sob um ponto de vista específico. Essas avaliações são relevantes para as análises de autoria quando revelam padrões comuns para textos de mesmos autores ou textos com características semelhantes (e.g. mesma língua, mesmo tipo de documento). Explorar a autoria sob diferentes tipos de redes ajuda a análise de autoria, por exemplo, a encontrar critérios de desempates quando dois autores têm características semelhantes.

4.3 Premissas teóricas das redes nas análises de autorias

Nesta seção, são apresentados, brevemente, trabalhos que evidenciam a relevância das pesquisas interdisciplinares no desenvolvimento da análise de autoria. Para categorizar e

[10] Em inglês, *Co-occurrence networks*.

[11] Sistemas dinâmicos discretos de simples construção e com um comportamento auto-organizado complexo (WOLFRAM, 1984).

delimitar as métricas e propriedades de redes em conformidade com as pesquisas sobre línguas, estilos e autorias, esta seção apresenta algumas premissas construídas a partir de alguns estudos desenvolvidos em diferentes correntes teóricas da linguística, estatística e matemática. A relação entre essas teorias delineia a proposta apresentada neste trabalho.

Primeiramente, assume-se que existem propriedades que são universais a todas as línguas do mundo (e.g. capacidade de nomear as coisas; usar conceitos relacionados à presente, passado e futuro etc.). Existem também propriedades que são próprias de cada língua (e.g. o conjunto de sons da língua, certas estruturas gramaticais) (CHOMSKY, 2019; CHOMSKY, 1994; MIOTO; SILVA; LOPES, 2007).

Em segundo lugar, a partir do funcionamento de cada língua, os falantes fazem suas próprias escolhas e organizações de palavras (e.g. o repertório de palavras que um falante usa durante o dia é diferente do repertório de outros falantes) e deixam marcas nas suas produções. Em terceiro lugar, nas redes, os estudos sobre estilos e línguas contribuem para a análise de autoria apontando os atributos que são relevantes para distinguir os textos dos autores. Além disso, as análises textuais, por meio de redes, identificam características que são semelhantes nos textos de vários autores (e.g. língua, tipo de texto, escolas literárias). As redes baseadas em textos, assim como outras redes do mundo real, apresentam o comportamento de redes do tipo mundo pequeno[12] (FADIGAS *et al.*, 2009; CANCHO; SOLÉ, 2001; STEYVERSA; TENENBAUM, 2005).

Nas redes, as línguas têm apresentado comportamentos em comum; redes com palavras em inglês, português, basco e russo apresentam distribuições de graus que se aproximam de uma lei de potência[13] (TEIXEIRA *et al.*, 2010; SOLE *et al.*, 2010; FADIGAS *et al.*, 2009; MASUCCI; RODGERS, 2006; ANTIQUEIRA *et al.*, 2005; CALDEIRA, 2005). A lei de potência também aparece em redes sociais, redes de informações, redes tecnológicas e redes biológicas (NEWMAN, 2010). Essa proximidade que existe entre a Teoria e Ciência das Redes e as diferentes áreas de pesquisas (e.g. biologia, humanas e exatas) permite que métricas, propriedades e conceitos de redes desenvolvidas por uma área possam ser compartilhados com as demais.

Com foco na distinção de línguas e autorias, Stanisz, Kwapien & Drozdz (2019), ao analisarem textos em inglês e polonês, identificaram métricas nas redes (caminho mínimo médio[14], coeficiente de aglomeração[15], modularidade[16] e coeficiente de assortatividade[17])

[12] O conceito de mundo pequeno trabalha com a ideia de que duas pessoas, selecionadas aleatoriamente em uma grande população, estão provavelmente conectadas por meio de poucos contatos em uma rede social (TRAVERS; MILGRAM, 1969).

[13] A lei de potência gera uma fórmula que pode ser descrita por $y = ax^k$; sendo a e k constantes; e x e y as variáveis.

[14] Média das distâncias entre todos os pares de vértices da rede.

[15] Mede a relação entre vértices vizinhos de modo a formar um grupo.

[16] Avalia a força da divisão da rede em módulos (grupos).

[17] A assortatividade mede a tendência de os vértices se conectarem a outros vértices.

com forte potencial para agrupar textos de mesma língua. Para os autores, métricas (coeficiente de agrupamento ponderado e o grau ponderado) auxiliam na identificação da autoria dos respectivos textos. A eficiência de métodos de análise de autoria com redes variam a depender da língua do texto analisado (QUINTANILLA, 2020; STANISZ; KWAPIEN; DROZDZ, 2019).

Para distinguir textos formais de informais, Sisovic, Martincic-Ipsic & Mestrovis (2014) usaram a razão entre o *strength*[18] e o grau da rede para diferenciar textos de *blogs* (i.e. textos geralmente escritos de maneira informal, sem o rigor de regras ortográficas ou gramaticais) de textos de literários (i.e. textos com maior rigor ortográfico e gramatical). Além das marcas de formalidade e informalidade, os textos também carregam características literárias. Amancio, Oliveira-Jr & Costa (2012a), ao analisarem as distribuições dos caminhos mínimos médios nas redes, encontraram padrões que detectam as mudanças de movimentos literários nos textos ao longo de um determinado período de tempo.

Em terceiro lugar, a qualidade de um texto pode ser medida por diferentes critérios de classificação (e.g. bom, regular, ruim) e pode estar relacionada a diferentes aspectos (e.g. ortografia, coerência, gramática). Ao analisar a qualidade dos textos[19] por meio de redes, Antiqueira *et al.* (2005) observaram que textos de boa qualidade apresentam coeficientes de aglomerações semelhantes. Antiqueira *et al.* (2007) observaram que os graus de saídas se relacionam com as qualidades dos textos. Em traduções, Amancio (2013) observou que caminhos mínimos e entropias nas redes apresentam resultados promissores para a avaliação da qualidade de traduções.

Em quarto lugar, em relação as métricas de redes e predições de autoria, o coeficiente de aglomeração, o grau e cálculos de caminhos estão entre as métricas de redes com mais acertos nas análises de autorias (STANISZ; KWAPIEN; DROZDZ, 2019; MACHICAO *et al.*, 2018; MARINHO; HIRST; AMANCIO, 2016; AMANCIO, 2015b; DAROONEH; SHARIATI, 2014; LAHIRI; MIHALCEA, 2013; MEHRI; DAROONEH; SHARIATI, 2012; AMANCIO *et al.*, 2011). O SVM, as redes neurais e as análises de agrupamento destacam-se como métodos de sugestão de autoria por apresentarem maiores índices de acertos (STANISZ; KWAPIEN; DROZDZ, 2019; MACHICAO *et al.*, 2018; ROZZ; MENEZES, 2018; AMANCIO, 2015b; SEGARRA; EISEN; RIBEIRO, 2014).

Por fim, a análise de autoria com apoio da Teoria e Ciência das Redes tem: modelos mistos (i.e. mesclam atributos de redes com métricas tradicionais da análise de autoria) (ROZZ; MENEZES, 2018); modelos baseados puramente em conceitos de redes (MARINHO; HIRST; AMANCIO, 2016); e modelos que mesclam conceitos de redes com conceitos de outras áreas (DUQUE; CARVALHO; VIMIEIRO, 2019; MACHICAO *et al.*, 2018).

[18] Mede a força de um vértice a partir do peso total de suas conexões (BARRAT *et al.*, 2004).
[19] Textos previamente avaliados por professores de redação.

Essa flexibilidade na criação de modelos acrescenta à AA formas diversas para validar o real autor de um texto. Apesar de a Teoria e Ciência das Redes ter acrescentado novos métodos na AA, algumas limitações ainda permanecem (e.g. a precisão de algoritmos de atribuição de autoria com métricas de redes diminui quando o número de autores analisados aumenta (STANISZ; KWAPIEN; DROZDZ, 2019; SEGARRA; EISEN; RIBEIRO, 2014)).

Dessa forma, a partir dos pontos apresentados acima, esta tese assumirá que: as línguas naturais partilham certas propriedades e, ao mesmo tempo, apresentam características que as particularizam; ao colocar em uso os sistemas linguísticos, os autores deixam marcas próprias em suas produções (e.g. escolhas de palavras, período da produção, variedade da língua etc.); a análise de autoria, por meio das redes, permite verificar, por exemplo, as relações de proximidade ou distanciamento entre as línguas e as marcas linguísticas deixadas pelos autores; as redes podem auxiliar na avaliação da qualidade de um texto; a aproximação de diferentes áreas de estudo é uma contribuição importante para a AA; métricas de rede e métodos de predição podem ser bons parâmetros para a verificação de autoria.

Capítulo Cinco

Métodos e materiais

Neste capítulo, apresenta-se um método de análise de autoria que: (i) identifica conjuntos de palavras recorrentes nas frases de diferentes autores e os organiza em redes; (ii) coleta dados das redes e os estrutura em tabelas com os autores e os seus respectivos atributos de redes; (iii) avalia a importância das métricas para a análise de autoria com algoritmos de predições.

Na Seção 5.1, estão descritos métodos e conceitos relacionados ao processo de descoberta de itens recorrentes em subconjuntos de dados (i.e. descobertas de regras de associação); na Seção 5.2, apresenta-se o método criado para transformar textos em redes; e, na Seção 5.3, são descritos os métodos de análise e a organização das métricas e dos métodos de predição.

5.1 Regras de associações

O método de análise proposto neste trabalho usou o algoritmo *Apriori* (AGRAWAL; SRIKANT, 1994; HAHSLER; GRUEN; HORNIK, 2005) para encontrar conjuntos de palavras recorrentes nas frases de diferentes autores. O *Apriori* foi formulado, inicialmente, para descobrir regras de associações em transações comerciais. As descobertas de regras de associações tornaram-se populares como método de análise para grandes volumes de dados comerciais, especialmente, nas análises das cestas de compras dos clientes de uma loja (HASTIE; TIBSHIRANI; FRIEDMAN, 2017; TAN; STEINBACH; KUMAR, 2006).

As análises de cestas de compras (*Market Basket Analysis*) usam métodos de associações para encontrar grupos de itens que, regularmente, são vendidos juntos em uma mesma transação (HASTIE; TIBSHIRANI; FRIEDMAN, 2017). A intenção é encontrar, nas transações comerciais, relações comuns entre os itens vendidos e usar essas informações para aumentar o lucro da loja seja, por exemplo, por meio de um sistema de recomendações de produtos, seja pela disposição dos produtos dentro das lojas (i.e. aproximando ou afastando itens que, regularmente, são vendidos juntos). As análises das cestas de compra podem ser realizadas por meio de duas abordagens: a mineração sequencial frequente (*Frequent Sequential Mining* (FSM)) e a mineração de itens frequentes (*Frequent Itemset Mining* (FIM)).

A mineração sequencial frequente trabalha com no mínimo três variáveis, a saber, *consumidor*, *data* e *item* (CAVIQUE, 2007a). O intuito é encontrar, em grandes bases de

dados, itens (i.e. produtos) que, constantemente, aparecem juntos em uma mesma cesta de compras ao longo do tempo. Para analisar esses tipos de dados, os algoritmos *SPADE* (ZAKI, 2001) e *RAMEX* (CAVIQUE, 2007a) reestruturam, respectivamente, as regras de associações em estruturas reticuladas e de redes.

A mineração de itens frequentes trabalha com duas variáveis, a saber, *identificação das transações*[1] e *itens* (CAVIQUE, 2007a). Nesse caso, os algoritmos de descoberta de regras de associação analisam as coocorrências entre os itens nas transações e formalizam regras como *"Quando os itens A e B aparecem, então o item D também aparece"* (WILLIAMS, 2011), *"muitos consumidores que compram fraldas também compram cerveja"* (TAN; STEINBACH; KUMAR, 2006) etc. Semanticamente, as regras de associação podem ser escritas de diferentes formas. No entanto, na síntese, elas são compostas por itens antecedentes e consequentes que se conectam em uma relação do tipo *Se {itens antecedentes}, então {itens consequentes}*[2].

A descoberta de regras de associação proposta por Agrawal & Srikant (1994), por meio do algoritmo *Apriori*, organiza os dados das transações comerciais em dois conjuntos; um conjunto I com todos itens $I = \{i_1, i_2, ..., i_n\}$ e um conjunto T com todas as transações $T = \{T_1, T_2, _3, ..., T_n\}$, sendo que, em cada transação T, existe um subconjunto de itens[3] de I (e.g. Tabela 5.1).

Tabela 5.1: Exemplo ilustrativo de transações e subconjuntos de itens

Transações	**Itens na transação**
T_1	$\{i_2, i_3\}$
T_2	$\{i_1, i_2, i_3\}$
T_3	$\{i_2, i_3, i_4\}$
T_4	$\{i_1, i_3, i_4\}$

Fonte: Elaboração própria

A regra de associação é, formalmente, representada por A $\rightarrow$ B, sendo A e B subconjuntos de itens distintos ($A \cap B = \varnothing$). Na regra de associação $\{i_1, i_3\} \rightarrow \{i_2\}$, A = $\{i_1, i_3\}$ e B = $\{i_2\}$. É a partir das regras de associações que métricas como *suporte* e *confiança* são calculadas. O *suporte* calcula a frequência na qual os itens de uma regra A $\rightarrow$ B (i.e. A $\cup$ B) aparecem no conjunto de transações analisadas. Por exemplo, com base nas transações da Tabela 5.1, o suporte da regra $\{i_2\} \rightarrow \{i_3\}$ é a frequência com que os itens $\{i_2, i_3\}$ aparecem juntos no conjunto de transações analisadas, no caso, 75% (os itens $\{i_2, i_3\}$ aparecem juntos em T_1, T_2 e T_3). A métrica de *suporte* está representada

[1] Código que identifica cada transação gravada no banco de dados.
[2] Ou então: Se *premissa*, então *conclusão*.
[3] Por definição, os subconjuntos de itens são chamados de *itemset*.

na Equação 5.1.

$$suporte_{regra}(A \cup B) = \frac{total(A \cup B)}{totaldeT} \tag{5.1}$$

onde *total* (A ∪ B) representa a quantidade total de transações em que A e B aparecem juntos e $totaldeT$ representa a quantidade total de transações analisadas.

Geralmente, o *suporte* é representado em porcentagem. A *confiança* é uma métrica baseada na razão entre o *suporte* da regra (A ∪ B) e o *suporte* de A (premissa da regra) (Equação 5.2). A *confiança* da regra $\{i_1, i_4\} \rightarrow \{i_3\}$ é 100%, dado que o *suporte* de $\{i_1, i_4\} \rightarrow \{i_3\}$ é 25% e o *suporte* da premissa $\{i_1, i_4\}$ é 25%. Semanticamente, isso significa dizer que "sempre que o conjunto $\{i_1, i_4\}$ aparece, o conjunto $\{i_3\}$ também aparece".

$$confianca_{regra}(A \cup B) = \frac{suporte_{regra}(A \cup B)}{suporte_{regra}(A)} \tag{5.2}$$

O *Apriori* está entre os algoritmos mais usados no trabalho de descobrimento de regras de associação (HERNANDEZ-GONZALEZ *et al.*, 2019). Esse algoritmo tem a vantagem de gerar regras de associações de fácil compreensão, além de formalizar regras que podem ser usadas para diferentes níveis de análise. No entanto apresenta algumas limitações ao criar um elevado número de regras de associações (WU *et al.*, 2021; FOURNIER-VIGER *et al.*, 2017; RODRIGUES; GAMA; FERREIRA, 2012; CAVIQUE, 2004); por exemplo, torna o trabalho de análise das informações uma atividade difícil e onerosa (FOURNIER-VIGER *et al.*, 2017; SANTOS; CARVALHO, 2017; KARIMI-MAJD; MAHOOTCHI, 2014; CAVIQUE, 2007b).

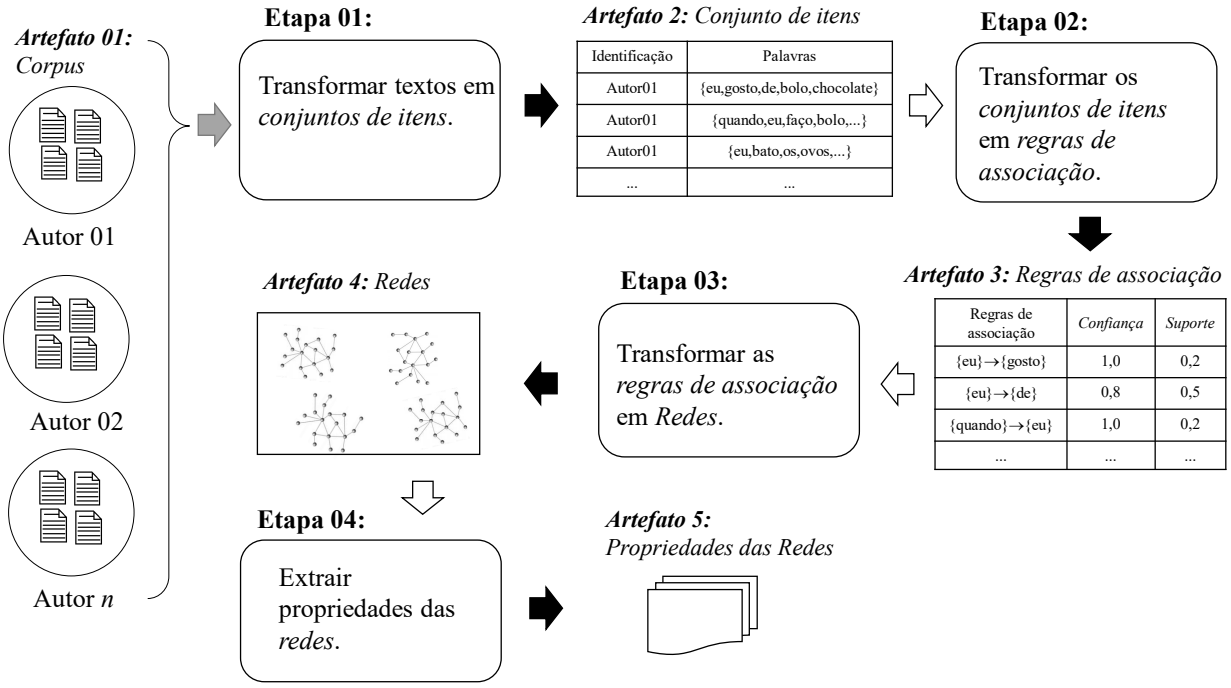
Além das aplicações comerciais, os algoritmos de *descoberta de regras de associação* já foram aplicados em sistemas de recomendação de livros (ALHARTHI; INKPEN; SZPAKOWICZ, 2018), sugestões jurídicas (LIU; HO, 2015), atribuição de autoria em *e-mails* (DEVI; RAVI, 2015) etc. Neste trabalho, as regras de associação foram usadas para encontrar conjuntos de palavras recorrentes nas frases de diferentes autores.

5.2 Método para construção de redes a partir de regras de associações

O pressuposto deste trabalho é que todos os autores têm diferentes preferências em relação à combinação das palavras que usam ao escreverem (i.e. uma identidade própria) e que

essas preferências podem ser percebidas em redes de regras de associações. O método ilustrado na Figura 5.1 foi criado para investigar esse aspecto. A Figura 5.1 apresenta a visão geral do processo de transformação de textos em redes de regras de associação. Os algoritmos desenvolvidos foram escritos na linguagem R[4] (TEAM, 2019) e estão nos apêndices deste trabalho. O *corpus* da pesquisa foi construído com o auxílio do *algoritmo Extrair Corpus* (Figura 2.1 - Apêndice C). A título de demonstração, foram usados textos de dois autores fictícios (Figura 5.2).

Figura 5.1: Visão geral do processo de transformação de textos em redes



Fonte: Elaboração própria

A diferença entre os estilos de escrita dos Autores 01 e 02 (Figura 5.2) é que o Autor01 prefere usar pronomes de primeira pessoa do singular e o Autor02 prefere usar pronomes de primeira pessoa do plural. Essas diferenças entre os estilos dos autores simulam as preferências de cada um e servem para demonstrar como o método aqui proposto detectará as diferenças entre os autores.

[4]O R é uma linguagem de programação usada em cálculos matemáticos e estatísticos.

Figura 5.2: Textos dos Autores 01 e 02

Autor01:

Texto 01	Texto 02	Texto 03	Texto 04
Eu gosto de bolo de chocolate. Quando eu faço bolo de chocolate, eu uso farinha, fermento e ovos. Eu bato os ovos em uma batedeira.	Eu corro regularmente. Gosto de correr de manhã e, depois de correr, tomo um bom banho. Não sei quantas horas eu corro.	Eu jogo xadrez duas horas por dia. Gosto de estudar as jogadas mais difíceis e exercito cada uma delas. No final do dia, eu corro dois quilometros para relaxar.	Quando eu lavo meu carro, eu limpo os vidros e as rodas por último. Eu gosto de limpar primeiro o painel e os estofados. Gosto de limpar o carro no domingo de manhã.

Autor02:

Texto 01	Texto 02	Texto 03	Texto 04
Nós gostamos de assar bolos em formas quadradas. Nós usamos chocolate, farinha e fermento. Assamos o bolo à temperatura de 180 graus.	Nós corremos em duplas. Entendemos que é importante um ajudar o outro durante as provas. No final, gostamos de alongar e conversar sobre a corrida.	Nós jogamos xadrez sempre que possível. Gostamos de praticar jogadas fáceis e difíceis. Depois dos treinos, nós conversamos um pouco.	Nós lavamos os carros em dias diferentes. Eu lavo na segunda-feira e ele lava na terça-feira. Gostamos de usar cera durante as lavagens.

Fonte: Elaboração própria

Na Etapa 01 (Figura 5.1), as frases de cada texto no *corpus* são transformadas em conjuntos de itens (i.e. palavras). A Tabela 5.2 ilustra o resultado da transformação do Texto 01 do Autor01 em um conjunto de palavras. O algoritmo da Etapa 01 (Apêndice D - *Algoritmo: Etapa 01 - Transformar textos em conjuntos de itens*) identifica as frases do texto[5], transforma os caracteres maiúsculos em minúsculos, remove números e caracteres especiais e organiza as palavras de cada frase em conjunto de itens, conforme a coluna

[5]Para identificar as frases, o algoritmo usa os caracteres de parada ".", "?", "!".

Itens da Tabela 5.2. Na construção de um conjunto de itens, não importa a ordem ou a quantidade dos produtos que apareceram nas cestas de compras pois, as criações das regras de associações serão baseadas nas aparições dos itens entre as transações e não na ordem ou nas quantidades que eles aparecem; dessa forma, não haverá repetição de palavras e também não importa a ordem que as palavras aparecem entre os conjuntos de itens. Depois de as palavras estarem organizadas em conjuntos itens, o algoritmo oferece as seguintes opções de processamento: (i) excluir, nos conjuntos de itens, as *stopwords* (i.e.palavras gramaticais); (ii) excluir, nos conjuntos de itens, as palavras que não estão nas *stopwords*; (iii) manter os conjuntos de itens como estão.

Tabela 5.2: Conjunto de palavras extraídas do Texto01 do Autor01

Identificador	**Itens**
Autor01	{eu, gosto, de, bolo, chocolate}
Autor01	{quando, eu, faço, bolo, de, chocolate, uso, farinha, fermento, e, ovos}
Autor01	{eu, bato, os, ovos, em, uma, batedeira}

Fonte: Elaboração própria

Os conjuntos de palavras criados na Etapa 01 servem de subsídio para que o algoritmo da Etapa 02 (Apêndice E - *Algoritmo: Etapa 02 - Transformar conjuntos de itens em regras de associação*) possa encontrar as respectivas regras de associação. Na Etapa 02, o algoritmo *Apriori* (AGRAWAL; SRIKANT, 1994; HAHSLER; GRUEN; HORNIK, 2005) foi configurado para encontrar regras de associação com *confianças* maiores que zero, *suportes* maiores que dois porcento e regras que combinem, no máximo, três palavras.

Ao considerar *confianças* maiores que zero, entram, para a análise, todas as regras de associação geradas pela Equação 5.2. Ao considerar *suportes* maiores que dois porcento, somente combinações de palavras com aparições maiores que esse valor serão consideradas nas análises. Configurar o *Apriori* dessa forma foi uma solução encontrada para evitar a sobrecarga da memória do computador com milhões de regras de associação e o consequente travamento. A Tabela 5.3 exibe parte das regras de associações encontradas com os dados da Tabela 5.2.

Na Etapa 03 (Apêndice F - *Algoritmo: Etapa 03 - Transformar regras de associação em redes*), ocorre a transformação das regras de associações em redes. A intenção, ao transformar regras de associação em redes, é reduzir a dimensão dos dados e aproveitar a estrutura da rede para extrair novos tipos de informações. No Texto 01 do Autor01, por exemplo, foram encontradas mais de dez mil regras de associação; esses dados, depois de transformados, geraram uma rede dirigida e ponderada com 17 vértices e 154 arcos. O

processo de transformação de um conjunto de regras de associação em uma rede segue o modelo proposto por Karimi-Majd & Mahootchi (2014). A Figura 5.3 ilustra o processo de construção da rede.

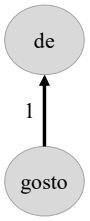
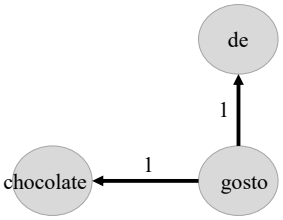
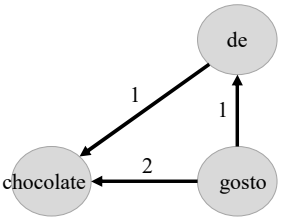
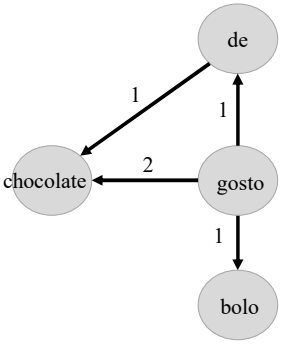
Tabela 5.3: Excerto das regras de associação criadas a partir do Texto 01 do Autor01

Identificação	**A → B**	**Suporte**	**Confiança**
Autor01	{gosto} → {de}	0,33	1,0
Autor01	{gosto} → {chocolate}	0,33	1,0
Autor01	{gosto} → {bolo}	0,33	1,0
Autor01	{gosto} → {eu}	0,33	1,0
Autor01	{os} → {em}	0,33	1,0
Autor01	{em} → {os}	0,33	1,0
Autor01	{os} → {bato}	0,33	1,0
Autor01	{bato} → {os}	0,33	1,0
Autor01	{os} → {uma}	0,33	1,0
Autor01	{uma} → {os}	0,33	1,0
Autor01	{os} → {ovos}	0,33	1,0
Autor01	{ovos} → {os}	0,33	0.5
Autor01	{os} → {eu}	0,33	1,0
Autor01	{em} → {ovos}	0,33	1,0
Autor01	{ovos} → {em}	0,33	0,5
...	...	...	...

Fonte: Elaboração própria

O processo de transformação de regras de associação em redes ocorre da seguinte forma: o algoritmo da Etapa 03 recebe um conjunto de regras de associação; lê a primeira regra de associação; transforma os itens (i.e. as palavras) em vértices; conecta, por meio de arcos, os vértices com base nos relacionamentos observados entre os itens da premissa e os itens da conclusão de cada regra de associação (e.g. linha nº1 da Figura 5.3). A cada novo relacionamento observado entre os itens das premissas e os itens das conclusões, o algoritmo soma mais um ao peso do arco envolvido. Na terceira regra de associação da Figura 5.3, por exemplo, o arco entre os vértices "gosto" e "chocolate" já existia, foi criado pela regra da linha nº2; com a terceira regra, o peso do arco recebeu mais uma conexão (i.e. passou de peso igual a um para peso igual a dois). Nesse tipo de construção, os pesos dos arcos registram os pares de palavras que mais vezes se relacionaram entre as premissas e conclusões das regras de associação e preservam, nas direções dos arcos, a informação sobre as prevalências das palavras.

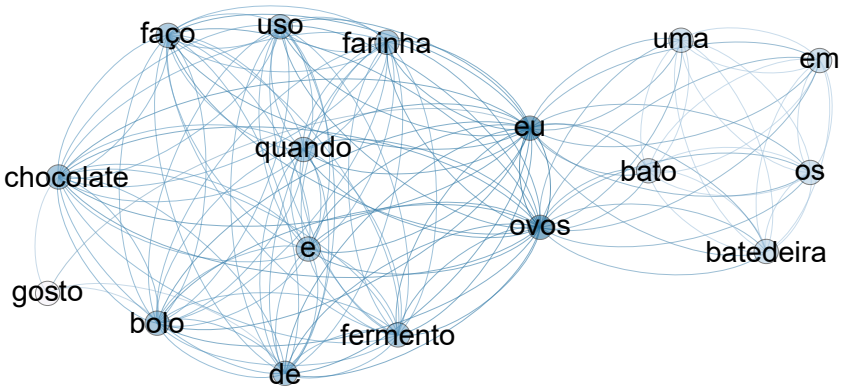
Figura 5.3: Transformação de regras de associação em uma rede

Nº	Regras de associação	Rede resultante
1	{gosto} → {de}	
2	{gosto} → {chocolate}	
3	{gosto, de} → {chocolate}	
4	{gosto} → {bolo}	
....	...	...

Fonte: Elaboração própria

O algoritmo da Etapa 03 cria redes dirigidas e ponderadas. No método proposto (Figura 5.1), cada texto no *corpus* será transformado em uma rede de regras de associação. A Figura 5.4 apresenta a rede de regras de associação criada a partir do Texto 01 do Autor01.

Figura 5.4: Redes de regras de associação gerada a partir do Texto 01 do Autor01



Fonte: Elaboração própria

Depois que todas as redes estão prontas, têm início os processos de extração e organização das métricas das redes. Na Etapa 4, são utilizados dois pacotes, o *igraph*[6] e o *centiserve*. O pacote *igraph* (CSARDI; NEPUSZ, 2006) é usado para extrair as métricas de centralidade de intermediação, grau ponderado, grau ponderado de entrada, grau ponderado de saída, grau, grau de entrada, grau de saída, excentricidade, excentricidade de entrada, excentricidade de saída, grau mais pesos[7], grau de entrada mais pesos de entrada e grau de saída mais pesos de saída.

O pacote *centiserve* (JALILI, 2017) é usado para extrair as centralidades de Laplace, centralidades de Laplace de entrada e centralidades de Laplace de saída. A centralidade de Laplace, se comparada com outras centralidades usadas em redes dirigidas (e.g. grau, centralidade de intermediação), é uma métrica que avalia as centralidades intermediárias entre as características globais e locais dos vértices (QI *et al.*, 2012). Com as centralidades de Laplace, pretende-se explorar a análise de autoria com uma métrica intermediária de centralidade e com as demais métricas, que já mostraram suas relevâncias em outras análises de autoria.

O algoritmo da Etapa 04 (Apêndice G - *Algoritmo: Etapa 04 - Extrair métricas das redes*) organiza as métricas das redes em tabelas cujas linhas representam as redes. As primeiras colunas registram os nomes dos autores dos textos que deram origem às redes e as demais colunas representam as métricas extraídas dos vértices das redes (e.g. Centralidade de intermediação apresentada na Tabela 5.4).

[6] O *igraph* é um pacote do programa estatístico R que trabalha com métricas e criações de redes.
[7] Soma das Equações 4.1 e 4.4.

Tabela 5.4: Excerto das centralidades de intermediações do vértices das redes de regras de associação criadas a partir dos textos dos Autores 01 e 02

Nome dos autores	**nós**	**de**	**eu**	...
Autor01	0	0	57	...
Autor01	0	0	2	...
Autor01	0	0	0	...
Autor01	0	0	62	...
Autor02	130	144	0	...
Autor02	0	0	0	...
Autor02	48	0	0	...
Autor02	0	240	0	...

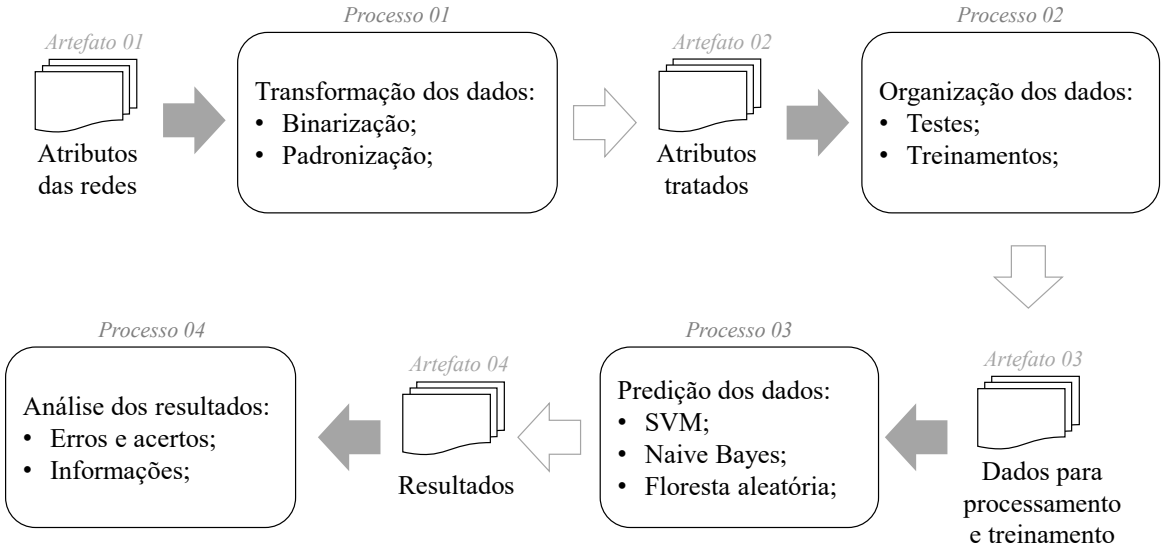
Fonte: Elaboração própria

O formato sugerido na Tabela 5.4 é compatível com os formatos solicitados pelos algoritmos de predição e, por esse motivo, a tabela não exibe o nome dos textos de origem. Nas colunas, estão os nomes dos vértices encontrados em todas as redes; se em alguma rede não tiver um dos vértices listado na tabela de comparação, o valor zero será atribuído para esse vértice.

5.3 Predições das autorias

A análise preditiva investiga o que existe de semelhante entre diferentes subconjuntos de dados e usa o resultado das investigações para inferir a que categoria um novo objeto pertence. Um algoritmo de predição meteorológica, por exemplo, analisa os atributos históricos do clima de uma região (e.g. velocidade dos ventos, umidade etc.) e, quando questionado se hoje choverá, o algoritmo, com base no clima de hoje, infere uma resposta do tipo sim, não ou a probabilidade de o evento ocorrer. O processo de predição desta pesquisa está esquematizado na Figura 5.5.

Figura 5.5: Modelo de análise de predição



Fonte: Elaboração própria

Na Figura 5.5, os processos 01, 02 e 03 estão descritos no algoritmo Processos de Predições (Apêndice H - *Algoritmo: Processos de Predições*). O artefato 01 representa as tabelas de atributos das redes geradas na Etapa 04 (Figura 5.1). No Processo 01, os atributos das redes passam pela binarização e normalização dos dados.

Na binarização de uma tabela, os valores iguais a zero permanecem inalterados e os diferentes de zero são substituídos por um. Nesse tipo de transformação, os objetos são caracterizados pela ocorrência (i.e. quando o valor é igual a um) ou pela ausência (i.e. quando o valor é igual a zero) de atributo. Na normalização, os dados numéricos são transformados em valores que variam entre 0 e 1 conforme a Equação 5.3.

$$v_{norm} = \frac{v - min(v)}{max(v) - min(v)} \tag{5.3}$$

onde v representa o valor a normalizar, a função $max(v)$ retorna o valor máximo do conjunto de dados e $min(v)$ retorna o valor mínimo.

A normalização coloca os dados em um intervalo de valor comum e permite a comparação

de variáveis com escalas de valores muito diferentes. A binarização e a normalização são técnicas usadas pelas análises de predição para aprimorar a precisão dos algoritmos.

O Processo 02 (Figura 5.5) recebe como parâmetro de entrada os dados transformados (binarizados ou normalizados) e os organiza de modo a facilitar os testes e verificações dos algoritmos de predições. A estratégia usada para organizar os dados de treinamento deste trabalho é conhecida por *leave-one-out* e está representada na Tabela 5.5.

Tabela 5.5: Estratégia *leave-one-out*

Iteração	Autor01 - Texto 01	Autor01 - Texto 02	Autor01 - Texto 03	Autor01 - Texto 04	Autor02 - Texto 01	Autor02 - Texto 02	Autor02 - Texto 03	Autor02 - Texto 04
1	0	1	1	1	1	1	1	1
2	1	0	1	1	1	1	1	1
3	1	1	0	1	1	1	1	1
4	1	1	1	0	1	1	1	1
5	1	1	1	1	0	1	1	1
6	1	1	1	1	1	0	1	1
7	1	1	1	1	1	1	0	1
8	1	1	1	1	1	1	1	0

0 - Dado usado no teste

1 - Dados usado no treinamento

Fonte: Elaboração própria

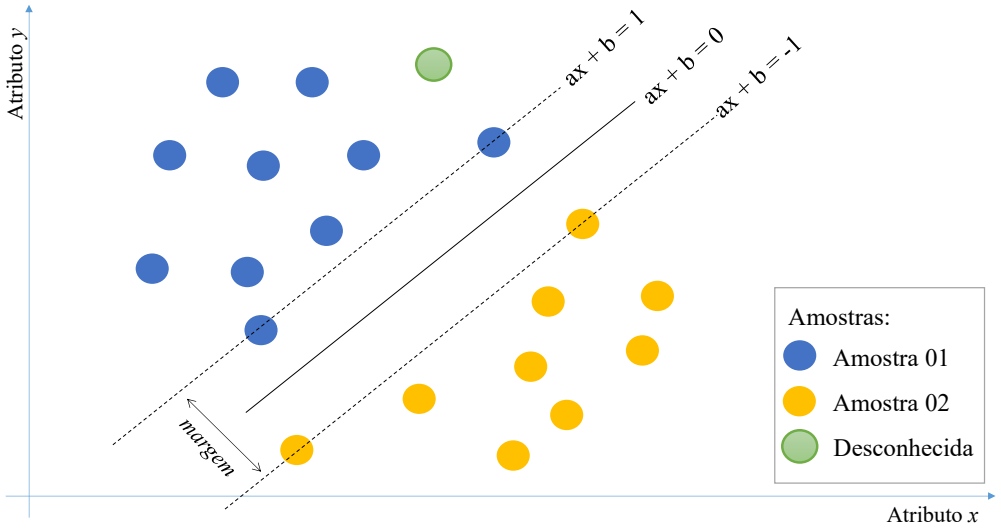
Em cada iteração da Tabela 5.5, os dados são divididos em dois subconjuntos, um com dados para o treinamento e o outro com dados para o teste do algoritmo de predição. Na estratégia *leave-one-out*, o algoritmo de teste, a cada iteração, oculta o nome de um dos objetos (valor com zero nas linhas da Tabela 5.5) e treina o algoritmo de predição com os demais exemplares (valores com um). No final do processo, todos os exemplares são testados pelo menos uma vez. A exigência mínima dessa estratégia é que exista um mínimo de dois objetos (e.g. nome do autor) e que cada objeto tenha dois conjuntos de atributos (e.g. livros), desse modo, um dos objetos é reservado para o teste do algoritmo de predição e os outros são usados para o treinamento do algoritmo. A estratégia *leave-one-out* é recomendada para a predição de *corpus* com poucos objetos (SILVA; PERES; BOSCARIOLI, 2016).

No Processo 03 (Figura 5.5), os algoritmos de predição SVM (MEYER *et al.*, 2019), Floresta aleatória (LIAW; WIENER, 2002) e Naive Bayes (MEYER *et al.*, 2019) são usados para realizar as análises de autorias. Esses três métodos de predição foram selecionados por apresentarem bons percentuais de acertos em outras pesquisas sobre análise de autoria

tradicionais e com redes (MACHICAO *et al.*, 2018; ROZZ; MENEZES, 2018; MARTINS *et al.*, 2019; MARINHO; HIRST; AMANCIO, 2018).

O SVM (*Support Vector Machines*) é um algoritmo de aprendizado supervisionado que, para classificar um objeto não identificado, inicialmente projeta todos os objetos conhecidos em uma representação espacial (e.g. gráfico de dispersão) e busca os objetos (i.e. vetores de suporte) que estão nas fronteiras entre os grupos de objetos conhecidos (e.g. as amostras 01 e 02 da Figura 5.6).

Figura 5.6: Representação gráfica da classificação linear do SVM

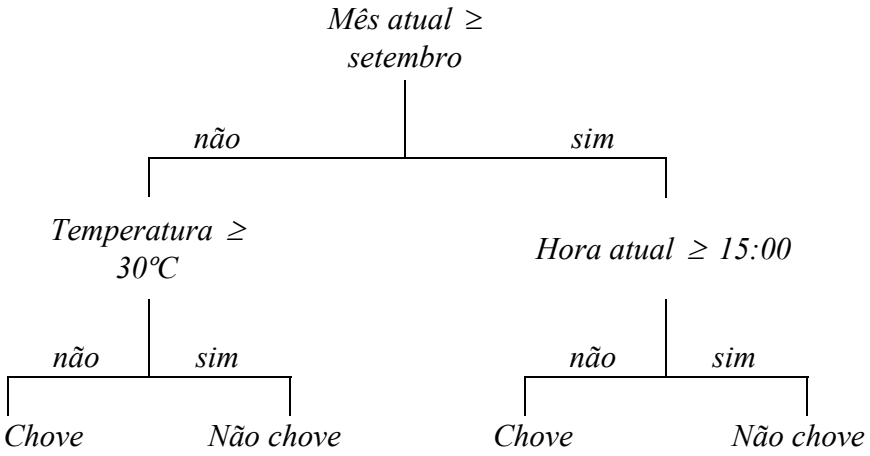


Fonte: Elaboração própria

Ao usar uma estratégia linear, o SVM projeta linhas retas próximas aos vetores de suporte de modo a identificar as fronteiras entre os grupos. A distância entre essas duas linhas é chamada de margem e a linha no meio, que separa os dados em duas dimensões, é chamada de hiperplano. O algoritmo busca classificar os objetos considerando a maior margem entre os grupos. Para classificar o objeto não identificado, o algoritmo o projeta na representação espacial e verifica qual ajuste linear melhor o categoriza.

O algoritmo de Floresta aleatória gera resultados a partir da análise de um conjunto de árvores de decisões. Uma árvore de decisão é um algoritmo que, hierarquicamente, organiza os dados em diferentes subconjuntos. A árvore na Figura 5.7 representa um modelo básico de estruturação dos dados que começa a ser organizado a partir da questão "hoje chove?".

Figura 5.7: Representação básica de uma árvore de decisão



Fonte: Elaboração própria

A previsão na árvore de decisão é interpretada a partir dos ramos e dos nós do modelo. A resposta para a questão "hoje chove?" depende dos resultados observados nas variáveis *mês atual*, *hora atual* e *temperatura*. O algoritmo de Floresta aleatória, ao trabalhar com um conjunto de árvores, toma uma decisão ao estudar diferentes arranjos de variáveis entre as árvores. Pelo fato de remover algumas variáveis durante o treinamento do algoritmo, o Floresta aleatória tende a ser mais preciso do que o algoritmo de árvore de decisão (WILLIAMS, 2011). Para este trabalho, o algoritmo de Floresta aleatória foi parametrizado para trabalhar com quinhentas árvores e com classificação de dados; esses parâmetros vêm pré-configurados no algoritmo e, por convenção, neste trabalho, optou-se por não alterar os valores desses parâmetros.

Baseado no Teorema de Bayes (Equação 5.4), o algoritmo de predição Naive Bayes trabalha com probabilidades condicionais (e.g. dado que um evento B ocorreu, qual é a probabilidade do evento A ocorrer?). Na Equação 5.4, A e B são eventos; P(A) e P(B) as probabilidades a *priori* de A e B; P(A|B) a probabilidade a *posteriori* de A condicionada a B; e P(B|A) a probabilidade a *posteriori* de B condicionada a A.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{5.4}$$

No Processo 04 (Figura 5.5), os parâmetros usados e os resultados das predições são organizados em uma tabela (e.g. Tabela 5.6) com os seguintes dados: o total de palavras

extraídas dos textos; os tipos de palavras[8]; os intervalos das *confianças*[9]; as métricas das redes; os métodos de predição; os tipos de transformações dos dados (i.e. binarização ou normalização); o número de predições corretas; e o total de textos analisados nos testes de predição.

Cada linha da Tabela 5.6 representa um teste de predição. Na primeira linha, a coluna *Total de palavras por texto* ilustra quantas foram as palavras extraídas dos textos; a coluna *Tipo de corpus* mostra quais conjuntos de palavras foram usadas nos testes; em *Intervalos das confianças (c)*, estão apresentados os intervalos de *confianças* usados; em *Métricas das redes*, está disposta a métrica de rede avaliada; em *Métodos de predição*, está nome o do algoritmo de predição usado pelo teste; em *Predições corretas*, está o total de predições corretas encontrada com os parâmetros informados nas colunas anteriores; e, em *Total de Textos*, está a quantidade de textos testados.

O intuito, ao explorar esses diferentes cenários de testes, é identificar métricas e informações que sejam relevantes para a caracterização deste novo método de pesquisa e encontrar características entre os parâmetros que sejam pertinentes para a AA.

Tabela 5.6: Excerto dos parâmetros e predições feitas com os textos do Autor01 e do Autor02

Total de palavras por texto	**Tipo de *corpus***	**Transformações dos dados**	**Métodos de predição**	**Métricas das redes**	**Intervalos de *confiança* (c)**	**Predições corretas**	**Total de Textos**
100	Original	Normalizados	Naive Bayes	Excentricidade de entrada	c =1.0	3	8
100	Sem *Stopwords*	Binarizados	Floresta aleatória	Excentricidade	0.8 ≤ c < 1.0	2	8
100	*Stopwords*	Binarizados	Floresta aleatória	Grau entrada	0.4 ≤ c < 0.6	4	8
100	Original	Binarizados	Floresta aleatória	Grau ponderado de entrada	c = 1.0	8	8
100	Original	Normalizados	Naive Bayes	Grau ponderado	0.2 ≤ c < 0.4	7	8
...		...	...	...	...	...	...

Fonte: Elaboração própria

Na Tabela 5.6, são apresentados os resultados dos testes de predição feitos com os textos do Autor01 e do Autor02 (Figura 5.2). No total, as predições realizaram testes em quinhentas e quarenta e seis arranjos de parâmetros (i.e. total de palavras por texto, tipo de *corpus*, transformações de dados, métodos de predição, métricas de redes e intervalos de *confiança*). A Tabela 5.7 sintetiza, entre tipo de *corpus*, tipo de transformação de dados

[8] Conforme exposto na Seção 2.2, os textos foram transformados em três conjuntos de itens: aquele com palavras lexicais e gramaticais (denominado de "*Original*"; aquele sem palavras lexicais (denominado de "*Stopwords*"); e aquele com o menor número possível de palavras gramaticais (denominado de "Sem *stopwords*")

[9] Foram criados seis intervalos diferentes com as métricas de *confiança* (c) {0.0≤c<0.2; 0.2≤c<0.4; 0.4≤c<0.6; 0.6≤c<0.8; 0.8≤c<1.0; c=1.0;}.

e métodos de predição, os testes que acertaram os autores dos textos no *corpus*.

Tabela 5.7: Síntese dos parâmetros que acertaram as predições feitas com os textos do Autor01 e do Autor02

Tipo de *corpus*	Transformação dos dados	Métodos de predição			Total Geral
		Floresta aleatória	Naives Bayes	SVM	
Original					
	Binarizados	188	201	**206**	595
	Normalizados	134	181	192	507
	Sub-total	***322***	***382***	***398***	***1.102***
Sem stopwords					
	Binarizados	124	120	150	394
	Normalizados	58	125	88	271
	Sub-total	***182***	***245***	***238***	***665***
Stopwords					
	Binarizados	178	193	190	561
	Normalizados	148	168	157	473
	Sub-total	***326***	***361***	***347***	***1.034***
Total Geral		**830**	**988**	**983**	**2.801**

Fonte: Elaboração própria

As predições feitas com o tipo de *corpus Original* foram as que obtiveram o maior número de acertos entre os testes realizados com os textos do Autor01 e Autor02 (Tabela 5.7). As análises preditivas dos três tipos de agrupamento do *corpus* serão usadas para investigar a relevância dos tipos *Original*, Sem *stopwords* e *Stopwords* na caracterização dos autores e estilos literários. Na análise dos textos de Autor01 e Autor02, o conjunto de palavras *Original* (i.e. com palavras gramaticais e lexicais) gerou mais elementos para caracterizar as combinações de palavras de cada autor do que os conjuntos de palavras dos outros tipos.

Em relação ao tipo de transformação de dados (Tabela 5.7), a binarização apresenta maior número de predições corretas; isso significa que, entre as métricas de redes analisadas, a presença (ou a ausência) de métricas entre os vértices nas redes é mais relevante para a caracterização dos marcadores de autoria de um escritor do que os valores proporcionais atribuídos aos vértices pelas métricas das redes.

Entre as redes formadas com o tipo de *corpus Original*, o método de *predição SVM* foi o que mais vezes acertou o nome dos autores testados. A Tabela 5.8 detalha, entre as métricas de redes e *confianças*, as predições dos dados da Tabela 5.7 para o tipo de *corpus Original*, tipo de *transformação binarizada* e método de *predição SVM*.

Tabela 5.8: Métricas das redes e escalas de *confianças* entre as *predições por SVM*, tipo de *corpus Original* e transformação de *dados binarizadas*

Métricas das redes	***Confianças***						**Total Geral**
	0.0<=c<0.2	**0.2<=c<0.4**	**0.4<=c<0.6**	**0.6<=c<0.8**	**0.8<=c<1.0**	**c=1.0**	
Excentricidade			7			8	15
Excentricidade de entrada			4			7	11
Excentricidade de saída			6			8	14
Grau			7			8	15
Grau de entrada			4			7	11
Grau de entrada mais peso			4			7	11
Grau mais peso			7			8	15
Grau ponderado			7			8	15
Grau ponderado de entrada			4			7	11
Grau ponderado de saída			6			8	14
Grau de saída			6			8	14
Grau de saída mais peso			6			8	14
Intermediacao			0			6	6
Laplace			7			8	15
Laplace de entrada			4			7	11
Laplace de saída			6			8	14
Total Geral			**85**			**121**	**206**

Fonte: Elaboração própria

A Tabela 5.8 ilustra onde, entre as métricas de redes e escalas de *confiança*, as predições feitas pelo *SVM* ocorreram. Para avaliar as métricas de rede Excentricidade, Grau, Grau mais peso, Intermediação e Laplace foram construídas redes ponderadas e não dirigidas; e, para avaliar as métricas de rede restantes, foram construídas redes ponderadas e dirigidas. Na Tabela 5.8, os valores variam de zero a oito; zero significa que, ou o método de predição não encontrou os autores dos documentos testados, ou não foram encontradas combinações de palavras entre as escalas de *confiança* avaliadas; quando o valor é igual a oito, o teste identificou os autores dos oito textos avaliados. Para o *corpus* composto pelos textos do Autor01 e do Autor02, a melhor parametrização do método de análise de autoria proposto neste trabalho consiste em usar o *corpus* do tipo *Original*, transformação de dados binarizada e o método de *predição SVM* porque, entre eles, ocorreu o maior número de predições corretas.

Na Seção 5.2, foi sugerido que o Autor01 prefere usar pronomes de primeira pessoa do singular e o Autor02 prefere usar pronomes de primeira pessoa do plural. Esses estilos dos Autores 01 e 02 foram criados com o propósito didático de apresentar e testar o método de análise deste trabalho. As marcas que um autor imprime em um texto envolvem escolhas de palavras que vão além da preferência por pronomes de primeiras pessoas do singular ou do plural.

O processo de análise proposto neste trabalho consiste em estudar as combinações de palavras partindo de um conjunto de dados maior para um menor. Os textos da pesquisa foram organizados e subdivididos em conjuntos de combinações de palavras (i.e. tipos de

corpus: *Original*, Sem *stopwords* e *Stopwords*); cada um desses conjuntos tem seus dados avaliados por um tipo de transformação de dados (i.e. binarização e normalização); cada tipo de transformação de dados é usado para predizer com os algoritmos Floresta aleatória, Naive Bayes e SVM, entre as métricas de redes e escalas de confiança, autores e elementos do contexto da produção dos textos. Essa estruturação dos dados foi proposta com o intuito de avaliar as questões e hipóteses da Seção 1.2. Na Tabela 5.9, são apresentados os parâmetros e valores considerados neste método.

Tabela 5.9: Descrição dos parâmetros

Parâmetros	**Descrição**	**Intervalo de valores**
Total de palavras por texto	Total de palavras extraídas dos textos. Analisar os valores mínimos e máximos de palavras em um texto ajuda a determinar as quantidades necessárias de palavras que o método precisa para trabalhar.	{100 palavras; 1.000 palavras}
Tipo de *corpus*	Palavras gramaticais e lexicais organizadas em três conjuntos. A análise de predição será baseada na organização das palavras nesses três conjuntos.	{Original, Sem *Stopwords, Stopwords}*
Transformações de dados	Método usado para padronização dos dados. Será usado para determinar qual dessas duas formas de transformação é relevante para as análises de predição.	{Binarização, Normalização}
Métodos de predição	Algoritmos usados na predição de dados. Serão usados para determinar a relevância desses algoritmos na predição do período de edição, da variedades linguística, da escola literária e autoria	{Floresta aleatória, Naive Bayes, SVM}
Métricas de redes	Métricas usadas para avaliar a importância dos vértices da rede. Essas métricas serão usadas para organizar e valorar as palavras.	{Grau de entrada, Grau de entrada mais peso de entrada, Grau de saída, Grau de saída mais peso saída, Grau mais peso, Grau ponderado, Grau ponderado de entrada, Intermediação, Laplace, Laplace de entrada, Laplace de saída,}
Intervalos de *confiança*	Intervalos que categorizam seis diferentes conjuntos de valores de *confiança*. As combinações de palavras aparecem entre os textos com diferentes números de ocorrências. Organizar as *confianças* em diferentes intervalos ajuda a diferenciar as combinações de palavras e a categorizar as preferências de cada autor.	{0.0<=c<0.2; 0.2<=c<0.4; 0.4<=c<0.6; 0.6<=c<0.8; 0.8<=c<1.0; c = 1.0}

Fonte: Elaboração própria

O método proposto analisa as predições de autoria sob diferentes parâmetros (cf. disposto nas colunas da Tabela 5.6). Na Tabela 5.6, os testes de predições são registrados considerando três tipos de *corpus*, seis intervalos diferentes de *confianças*, dezesseis métricas de redes, três métodos de predição e dois tipos de transformações de dados. Serão testados, aproximadamente, mil e setecentos cenários diferentes por *corpus*; esse valor pode variar porque, nem sempre, o texto de um autor gera regras de associação para os seis intervalos de confianças definidos. No Capítulo 6, o método aqui proposto será aplicado na

análise do *corpus* da pesquisa em diferentes contextos (i.e. períodos de edição, variedade linguística, escolas literárias e autoria).

Capítulo Seis

Resultados e discussões

Neste capítulo, o método descrito no Capítulo 5 foi testado a partir da extração de diferentes subconjuntos do *corpus* apresentado no Apêndice A. Por se tratar de um novo processo de trabalho, a proposta foi explorar o máximo de combinações de métricas possível e configurar os parâmetros (i.e. as métricas) que mais bem predizem os objetos das pesquisas. Nas próximas seções, estão descritos os testes realizados.

6.1 Os textos do corpus

Na primeira análise, o método proposto no Capítulo 5 explorou todos os textos no Apêndice A como um *corpus* único, aqui chamado de *Corpus* A. O conjunto de textos do *Corpus* A reúne obras de seis escolas literárias, publicadas em diferentes épocas e escritas por autores brasileiros e portugueses. Embora seja um problema para a análise de autoria misturar textos com diferentes características em um mesmo *corpus*, o propósito desta primeira análise foi explorar o método apresentado no Capítulo 5 e identificar vantagens e limites no processamento de grandes volumes de dados.

Durante essa primeira análise, identificou-se que um só texto pode gerar milhões de regras de associação e, por conta disso, o processamento dos dados fica lento e oneroso em termos de ocupação da memória volátil do computador (i.e. memória RAM). Para reduzir a quantidade de regras de associação e tornar o processo de análise menos custoso, configurou-se o algoritmo *Apriori* para descobrir regras de associação com no mínimo dois e no máximo três itens entre premissas e conclusões das regras de associação. Dessa forma, o método proposto analisa a frequência de combinações de no mínimo duas e no máximo três palavras recorrentes em cada texto do *corpus*.

Os textos no *Corpus* A foram organizados em subconjuntos categorizados por total de palavras por texto (*cem* e *mil palavras*); tipo de conjunto de texto (*Original*, Sem *stopwords* e *Stopwords*); e escalas de *confianças* (c) entre $\{0.0 <= c < 0.2; 0.2 <= c < 0.4; 0.4 <= c < 0.6; 0.6 <= c < 0.8; 0.8 <= c < 1.0; c = 1.0\}$. Depois de organizados, cada conjunto passou pelo processo de construção de redes de regras de associação e predição apresentados no Capítulo 5. O método proposto foi preparado para testar as combinações de três tipos de *corpus*, seis escalas de *confianças*, dois totais de palavras por texto, dezesseis métricas de redes, três métodos de predição e dois tipos de transformação de dados.

A combinação dessas variáveis gera *três mil quatrocentos e cinquenta e seis* parâmetros

(i.e. cenários) para teste; no entanto, com o *Corpus* A só *duas mil setecentas e setenta e duas* combinações de parâmetros foram criadas. Essa diferença aconteceu porque nem todos os escritores formalizam, em seus textos, combinações de três palavras que aparecem entre as seis escalas de *confianças*. A Tabela 6.1 exibe os parâmetros que obtiveram a maior quantidade de predições corretas entre os testes.

Tabela 6.1: Parâmetros dos testes que acertaram o maior número de nomes dos autores nos textos avaliados

Total de palavras por texto	Tipo de *corpus*	Transformação dos dados	Métodos de predição	Métricas das redes	Confianças	Total de predições corretas	Total de textos avaliados
1.000	Stopwords	Binarizados	Floresta aleatória	Excentricidade	0.2>c>=0.0	29	86
1.000	Stopwords	Binarizados	Floresta aleatória	Laplace	0.2>c>=0.0	29	86

Fonte: Elaboração própria

Por conta da mistura de contextos (edições em diferentes anos, diferentes escolas literárias etc.) no *corpus*, já era esperado um baixo valor de predições corretas em relação ao total de documentos testados. Na Tabela 6.2, estão sintetizadas a predições feitas com o autores do *Corpus* A.

Tabela 6.2: Síntese dos parâmetros e respectivos totais de predições corretas

Total de palavras por texto	Tipo de *corpus*	Transformação dos dados	Métodos de predição			Total Geral
			Floresta aleatória	Naives Bayes	SVM	
100						
	Original					
		Binarizados	999	287	991	2.277
		Normalizados	915	263	860	2.038
		Sub-total	***1.914***	***550***	***1.851***	***4.315***
	Sem Stopwords					
		Binarizados	311	68	425	804
		Normalizados	281	80	343	704
		Sub-total	***592***	***148***	***768***	***1.508***
	Stopwords					
		Binarizados	791	321	727	1.839
		Normalizados	740	390	713	1.843
		Sub-total	***1.531***	***711***	***1.440***	***3.682***
	Sub-totais de 100 palavras por texto		***4.037***	***1.409***	***4.059***	***9.505***
1.000						
	Original					
		Binarizados	1.765	671	1.919	4.355
		Normalizados	1.739	1.039	1.485	4.263
		Sub-total	***3.504***	***1.710***	***3.404***	***8.618***
	Sem Stopwords					
		Binarizados	1.070	323	1.177	2.570
		Normalizados	975	483	1.005	2.463
		Sub-total	***2.045***	***806***	***2.182***	***5.033***
	Stopwords					
		Binarizados	**1.987**	706	1.887	4.580
		Normalizados	1.928	854	1.766	4.548
		Sub-total	***3.915***	***1.560***	***3.653***	***9.128***
	Sub-totais de 1.000 palavras por texto		***9.464***	***4.076***	***9.239***	***22.779***
Total Geral			**13.501**	**5.485**	**13.298**	**32.284**

Fonte: Elaboração própria

Nesse contexto do *Corpus* A Tabela (6.2), os parâmetros com maiores valores foram: textos com *mil palavras*; predição por *Floresta aleatória*; *dados binarizados*; e *Stopwords*. Na Tabela 6.3, estão, entre as métricas de redes e de *confianças*, os detalhes das predições realizadas com *mil palavras*, *Stopwords* e binarização dos dados.

Tabela 6.3: Predições de autorias com *mil palavras*, predição por *Floresta aleatória*, *dados binarizados* e *stopwords*

Métricas das redes	***Confianças***						**Total Geral**
	0.0<=c<0.2	**0.2<=c<0.4**	**0.4<=c<0.6**	**0.6<=c<0.8**	**0.8<=c<1.0**	**c=1.0**	
Excentricidade	**29**	25	23	20	10	24	131
Excentricidade de entrada	28	24	20	23	16	21	132
Excentricidade de saída	10	21	18	27	11	28	115
Grau	28	26	24	20	12	28	138
Grau de entrada	27	23	22	20	16	23	131
Grau de entrada mais peso de entrada	26	22	22	22	16	22	130
Grau de saída	12	21	17	24	10	26	110
Grau de saída mais peso saída	12	20	17	26	11	26	112
Grau mais peso	26	25	24	20	11	27	133
Grau ponderado	27	24	23	20	11	24	129
Grau ponderado de entrada	27	24	24	18	17	20	130
Grau ponderado de saída	13	19	19	26	10	26	113
Intermediacao	12	17	20	20	11	24	104
Laplace	**29**	24	23	22	10	26	134
Laplace de entrada	27	25	22	23	16	23	136
Laplace de saida	13	19	16	24	10	27	109
Total Geral	**346**	**359**	**334**	**355**	**198**	**395**	**1987**

Fonte: Elaboração própria

Na Tabela 6.3, estão, entre métricas de redes e *confianças*, os totais das predições de autoria realizadas. Caso o método tivesse identificado corretamente todos os autores, o valor máximo, entre métricas de redes e *confianças*, deveria ser oitenta e seis (i.e. o total de obras analisadas no *corpus*). Na Seção 6.5, será demonstrado que as escalas de *confianças* estão relacionadas com o repertório de palavras próprias de cada autor e que, por essa razão, nas comparações entre obras de dois autores, as predições entre as escalas de *confianças* variam. Para que uma comparação possa distinguir entre dois autores, por exemplo, seus textos devem gerar métricas de *confianças* em uma mesma escala e ter combinações de palavras distintas.

Embora o método proposto não tenha previsto corretamente a autoria de todos os textos do *Corpus* A, as análises desses textos foram relevantes para tornar o processamento dos dados do método mais econômico em termos de volume de dados gerado na memória do computador. Quando o método proposto não estava parametrizado para gerar regras de associação com três itens, o processamento dos dados mostrou-se muito lento e sobrecarregou a memória volátil do computador.

6.2 Diferentes edições de uma mesma obra

O pressuposto na AA é que o conjunto de palavras que um escritor usa ao escrever é diferente do conjunto de palavras de outros escritores. No método proposto, a métrica de *confiança* é usada para encontrar combinações de até três palavras que são recorrentes em um texto e, depois que as regras de associações são categorizadas em escalas de *confianças* e organizadas em redes, o método proposto infere, ao analisar as métricas das redes, se os textos têm características em comum ao ponto de pertencerem a uma mesma pessoa. Nesta seção, os textos analisados pertencem a um mesmo autor, então, ao usar o método proposto para inferir os séculos das edições, o propósito é encontrar combinações de palavras que demonstrem que as alterações realizadas nos textos são condizentes com uma estrutura textual que aparece em um determinado período do tempo.

As reedições de uma obra literária são motivadas por diferentes aspectos (e.g. adequação ortográfica, acréscimo de notas etc.). Nem sempre as alterações feitas nas reedições são realizadas pelos autores. Nesta seção, o método proposto foi usado para analisar as reedições de uma mesma obra. Pretende-se mostrar que as alterações e os acréscimos nas reedições deixam, nos textos, marcas que caracterizam a época na qual foram reeditados. Para a primeira análise, foram consideradas duas obras de Machado de Assis, editadas em séculos diferentes (Tabela 6.4).

Tabela 6.4: Obras de Machado de Assis publicadas em diferentes anos

Nome do autor	Título	Escola Literária	Ano da edição	Século da edição
Machado de Assis	Memórias Póstumas de Brás Cubas	Realismo	1881	XIX
Machado de Assis	Memórias Póstumas de Brás Cubas	Realismo	**1994**	**XX**
Machado de Assis	Quincas Borba	Realismo	**1994**	**XX**
Machado de Assis	Quincas Borba	Realismo	1891	XIX

Fonte: Elaboração própria

Na Tabela 6.4, os textos pertencem a um mesmo autor e a mesma escola literária, com variação dos anos das edições. O método proposto, ao invés de predizer um autor, foi usado para inferir o século das edições.

Tabela 6.5: Síntese e resultados das predições dos séculos nos textos de Machado de Assis

Total de palavras por texto	**Tipo de *corpus***	**Transformação dos dados**	**Métodos de predição**			**Total Geral**
			Floresta aleatória	**Naives Bayes**	**SVM**	
100						
	Original					
		Binarizados	2	0	**72**	74
		Normalizados	4	0	7	11
		Sub-total	***6***	***0***	***79***	***85***
	Sem stopword					
		Binarizados	0	0	0	0
		Normalizados	3	0	0	3
		Sub-total	***3***	***0***	***0***	***3***
	Stopword					
		Binarizados	0	0	21	21
		Normalizados	14	0	1	15
		Sub-total	***14***	***0***	***22***	***36***
	Sub-totais de 100 palavras por texto		***23***	***0***	***101***	***124***
1.000						
	Original					
		Binarizados	2	0	**72**	74
		Normalizados	4	0	7	11
		Sub-total	***6***	***0***	***79***	***85***
	Sem stopword					
		Binarizados	0	0	0	0
		Normalizados	3	0	0	3
		Sub-total	***3***	***0***	***0***	***3***
	Stopword					
		Binarizados	0	0	21	21
		Normalizados	14	0	1	15
		Sub-total	***14***	***0***	***22***	***36***
	Sub-totais de 1.000 palavras por texto		***23***	***0***	***101***	***124***
Total Geral			**46**	**0**	**202**	**248**

Fonte: Elaboração própria

Na Tabela 6.5, estão registrados os resultados e os parâmetros dos testes de predição dos séculos das obras de Machado de Assis. Os parâmetros tipo de *corpus Original*, método de *predição SVM* e tipo de transformação de *dados binarizados* obtiveram mais acertos ao inferir os séculos dos textos. Entre esses parâmetros, não há diferenças nos totais de acertos entre as predições dos textos com *cem palavras* e textos com *mil*. Na Tabela 6.6, estão, entre as métricas de redes e *confianças*, os resultados dos testes de predição.

Tabela 6.6: Predições das edições de Machado de Assis parametrizadas com *mil palavras*, tipo de *corpus Original*, *dados binarizados* e predição por *SVM*

Métricas das redes	***Confianças***						**Total Geral**
	0.0<=c<0.2	**0.2<=c<0.4**	**0.4<=c<0.6**	**0.6<=c<0.8**	**0.8<=c<1.0**	**c=1.0**	
Excentricidade		2		**4**			6
Excentricidade de entrada		2			2		4
Excentricidade de saída		1		3			4
Grau		2		**4**			6
Grau de entrada		2			2		4
Grau de entrada mais peso		2			2		4
Grau de saída		1		3			4
Grau de saída mais peso		1		3			4
Grau mais peso		2		**4**			6
Grau ponderado		2		**4**			6
Grau ponderado de entrada		2			2		4
Grau ponderado de saída		1		3			4
Intermediacao		1			1		2
Laplace		2		**4**			6
Laplace de entrada		2			2		4
Laplace de saída		1		3			4
Total Geral		**26**		**35**	**11**		**72**

Fonte: Elaboração própria

Na Tabela 6.6, entre as métricas de redes e *confianças*, os testes de predição que acertaram os séculos das edições de todos os textos de Machado de Assis aparecem com valor igual a quatro (i.e. dos quatro textos analisados, o algoritmo acertou a predição do *período de edição* de todos). Se os parâmetros (i.e. *mil palavras*, tipo de *corpus Original*, *dados binarizados* e *predição por SVM*) observados na Tabela 6.5 apresentam potencial relevância para a predição de edições publicadas em diferentes séculos de uma mesma obra, é esperado que esses parâmetros apareçam também na análise de edições de outros autores. No *Corpus* A, outro autor que também tem textos com diferentes edições é João Carlos Medeiros Pardal Mallet (Tabela 6.7).

Tabela 6.7: Diferentes edições da obra Hóspede de João Carlos Medeiros Pardal Mallet

Nome do autor	**Título da obra**	**Capítulos**	**Escola Literária**	**Ano da edição**	**Século da edição**
João Carlos de Medeiros Pardal Mallet	Hóspede	I-II	Romantismo	1887	XIX
João Carlos de Medeiros Pardal Mallet	Hóspede	III-IV	Romantismo	1887	XIX
João Carlos de Medeiros Pardal Mallet	Hóspede	I-II	Romantismo	**2008**	**XXI**
João Carlos de Medeiros Pardal Mallet	Hóspede	III-IV	Romantismo	**2008**	**XXI**

Fonte: Elaboração própria

Na Tabela 6.7, há duas edições de uma mesma obra; uma publicada no século XIX e a outra no século XXI. Para o teste nas edições da obra Hóspede, de João Carlos de Medeiros Pardal Mallet, os textos foram divididos em duas partes, uma com os capítulos um e dois; e outra com os capítulos três e quatros. Os textos da Tabela 6.7 foram processados pelo

método proposto e estão sintetizados na Tabela 6.8.

Tabela 6.8: Síntese e resultados das predições dos séculos nos textos de João Carlos de Medeiros Pardal Mallet

Total de palavras por texto	Tipo de *corpus*	Transformação dos dados	Métodos de predição			Total Geral
			Floresta aleatória	Naives Bayes	SVM	
1.000						
	Original					
		Binarizados	1	0	**204**	205
		Normalizados	10	0	24	34
		Sub-total	***11***	***0***	***228***	***239***
	Sem Stopwords					
		Binarizados	0	0	135	135
		Normalizados	29	0	11	40
		Sub-total	***29***	***0***	***146***	***175***
	Stopwords					
		Binarizados	2	0	165	167
		Normalizados	14	0	9	23
		Sub-total	***16***	***0***	***174***	***190***
Total Geral			**56**	**0**	**548**	**604**

Fonte: Elaboração própria

Em comparação com os dados observados na Tabela 6.5, os parâmetros *mil palavras* por texto, tipo de *corpus Original*, *dados binarizados* e método de *predição SVM* apresentaram a mesma relevância na predição por século de edição. Na Tabela 6.9, estão detalhadas, entre métricas das redes e *confianças*, as predições parametrizadas por *mil palavras*, *corpus Original*, *dados binarizados* e tipo de *predição por SVM*.

Tabela 6.9: Resultado das predições com *mil palavras* por texto, o tipo de *corpus Original*, *dados binarizados* e *predição por SVM* nos textos de João Carlos de Medeiros Pardal Mallet

Métricas das redes	*Confianças*						Total Geral
	0.0<=c<0.2	0.2<=c<0.4	0.4<=c<0.6	0.6<=c<0.8	0.8<=c<1.0	*c=1.0*	
Excentricidade		4	4	4		4	16
Excentricidade de entrada		4	4	4		4	16
Excentricidade de saída				4		4	8
Grau		4	4	4		4	16
Grau de entrada		4	4	4		4	16
Grau de entrada mais peso		4	4	4		4	16
Grau de saída				4		4	8
Grau de saída mais peso				4		4	8
Grau mais peso		4	4	4		4	16
Grau ponderado		4	4	4		4	16
Grau ponderado de entrada		4	4	4		4	16
Grau ponderado de saída				4		4	8
Intermediacao						4	4
Laplace		4	4	4		4	16
Laplace de entrada		4	4	4		4	16
Laplace de saída				4		4	8
Total Geral		**40**	**40**	**60**		**64**	**204**

Fonte: Elaboração própria

Nas obras de João Carlos de Medeiros Pardal Mallet (Tabela 6.9), as distinções textuais entre os séculos são identificadas em diferentes escalas de *confianças* e em diferentes métricas das redes. Os valores iguais a quatro, na Tabela 6.9, ilustram que o método inferiu, corretamente, os *períodos de edição* dos quatro textos de João Carlos de Medeiros Pardal Mallet.

O modo como as modificações são feitas entre as edições de uma mesma obra varia quando essas modificações alteram combinações de até três palavras recorrentes nos textos. O método proposto, por meio das métricas de rede e *confiança*, consegue identificar a mudança e inferir a que século a modificação pertence; no caso dos textos de Machado de Assis, essas mudanças foram melhor caracterizadas em uma escala específica de *confianças*; por sua vez, nas redes de João Carlos de Medeiros Parda Mallet, as mudanças foram bem marcadas em diferentes escalas.

Os parâmetros que melhor identificaram os séculos das edições foram encontrados nos textos com *mil palavras*, tipo de *corpus Original*, *dados binarizados* e com o método de *predição SVM*; isso ocorreu tanto no conjunto de textos de Machado de Assis, quanto no conjunto de textos de João Carlos de Medeiros Pardal Mallet. Apesar de *mil palavras* por texto representarem um valor que se aplica, relevantemente, a diferentes cenários de testes, há, entre os autores, contextos de produção onde, com *cem palavras* por texto (e.g. textos de Machado de Assis (Tabela 6.5)), é possível inferir, de maneira relevante, o contexto de produção textual.

Investigar as distinções de *período de edição* com um número maior de textos é relevante, pois ajuda a fortalecer a relevância do parâmetro sugerido (i.e. *mil palavras*, tipo de *corpus Original*, *dados binarizados* e com o método de *predição SVM*); e, no que se refere à relação entre métricas de rede e escalas de *confianças*, auxilia na indicação dos locais onde as predições são mais precisas entre as obras de um mesmo autor.

6.3 Variedade de língua portuguesa dos autores

A língua de origem do autor é um elemento que influencia a análise de autoria. Conforme apresentado anteriormente, os algoritmos de AA variam de precisão quando analisam *corpus* com línguas diferentes; além disso, as traduções são um problema para a AA, porque o tradutor deixa marcas no texto que traduz. Existe uma lacuna na AA no que diz respeito ao modo como as análises são afetadas quando se trata de *corpus* contendo variedades de uma mesma língua.

No *Corpus* A, há textos escritos em duas variedades de uma mesma língua, no caso, o português brasileiro e o português europeu. Para investigar essas duas variedades do

português, usou-se o método proposto para processar o conjunto de textos listados na Tabela 6.10.

Tabela 6.10: Autores brasileiros e portugueses que publicaram entre 1870 e 1880

Nome do Autor	Título da obra	Ano da edição	Escola literária	País de origem da autor
Pinheiro Chagas	Os Guerrilheiros da Morte	1872	Romantismo	Portugal
Pinheiro Chagas	O terramoto de Lisboa	1874	Romantismo	Portugal
Camilo Castelo Branco	Livro de Consolação	1872	Romantismo	Portugal
Camilo Castelo Branco	Carlota Ângela	1874	Romantismo	Portugal
Machado de Assis	A mão e a luva	1874	Romantismo	Brasil
Machado de Assis	Helena	1876	Romantismo	Brasil
Bernardo Guimarães	O índio Afonso	1873	Romantismo	Brasil
Bernardo Guimarães	Maurício ou Os Paulistas em São João del-Rei	1877	Romantismo	Brasil

Fonte: Elaboração própria

Tabela 6.11: Resultados e parâmetros dos testes de predição feitos para inferir a variedade linguística dos autores

Total de palavras por texto	Tipo de *corpus*	Transformação dos dados	Métodos de predição			Total Geral
			Floresta aleatória	Naive Bayes	SVM	
100						
	Original					
		Binarizados	170	264	116	550
		Normalizados	132	206	70	408
		Sub-total	***302***	***470***	***186***	***958***
	Sem stopwords					
		Binarizados	0	0	0	0
		Normalizados	3	30	24	57
		Sub-total	***3***	***30***	***24***	***57***
	Stopwords					
		Binarizados	238	308	214	760
		Normalizados	182	220	127	529
		Sub-total	***420***	***528***	***341***	***1.289***
	Sub-totais de 100 palavras por texto		725	1.028	551	2.304
1.000						
	Original					
		Binarizados	275	**543**	345	1.163
		Normalizados	195	318	176	689
		Sub-total	***470***	***861***	***521***	***1.852***
	Sem stopwords					
		Binarizados	229	394	269	892
		Normalizados	196	328	191	715
		Sub-total	***425***	***722***	***460***	***1.607***
	Stopwords					
		Binarizados	249	443	279	971
		Normalizados	196	328	172	696
		Sub-total	***445***	***771***	***451***	***1.667***
	Sub-totais de 1.000 palavras por texto		***1.340***	***2.354***	***1.432***	***5.126***
Total Geral			**2.065**	**3.382**	**1.983**	**7.430**

Fonte: Elaboração própria

Na Tabela 6.10, há oito obras literárias, quatro autores (dois brasileiros e dois portugueses)

e textos que pertencem à mesma escola literária, editados na mesma década. Para analisar esses textos, o método proposto usou como objeto de predição o nome dos países de origem dos autores.

Na Tabela 6.11, estão os resultados e parâmetros das predições feitas para inferir o país de origem dos autores [1] dos textos da Tabela 6.10. Na Tabela 6.11, o parâmetro que inferiu, corretamente, o país de origem dos autores (i.e. a variedade linguística) foi: *mil palavras*, tipo de *corpus Original*, *dados binarizados* e método de predição por *Naive Bayes*. A partir desse parâmetro, as predições foram detalhas, entre métricas de rede e *confiança*, na Tabela 6.12.

Tabela 6.12: Resultados das predições da variedade linguística dos escritores com os parâmetros *mil palavras*, tipo de *corpus Original*, *dados binarizados* e predição por *Naive Bayes*

Métricas das redes	***Confianças***						**Total Geral**
	0.0<=c<0.2	**0.2<=c<0.4**	**0.4<=c<0.6**	**0.6<=c<0.8**	**0.8<=c<1.0**	**c=1.0**	
Excentricidade	5	6	7	7	7	4	36
Excentricidade de entrada	6	6	6	7	4	4	33
Excentricidade de saída	5	5	6	7	7	5	35
Grau	7	7	6	7	7	5	39
Grau de entrada	7	5	7	7	4	5	35
Grau de entrada mais peso	6	**8**	7	7	4	6	38
Grau de saída	3	2	5	7	7	5	29
Grau de saída mais peso	5	0	6	7	7	4	29
Grau mais peso	7	7	7	7	7	3	38
Grau ponderado	6	6	5	7	7	4	35
Grau ponderado de entrada	7	5	6	7	4	5	34
Grau ponderado de saída	4	4	4	7	7	4	30
Intermediacao	0	2	7	7	4	5	25
Laplace	7	7	7	7	7	5	40
Laplace de entrada	7	7	7	7	4	5	37
Laplace de saída	4	3	5	7	7	4	30
Total Geral	**86**	**80**	**98**	**112**	**94**	**73**	**543**

Fonte: Elaboração própria

Na Tabela 6.12, o valor igual a oito indica que o método inferiu, corretamente, a variedade linguística dos escritores nos oito textos analisados. Observa-se, na Tabela 6.12, que as combinações de palavras que diferenciam o português brasileiro do português europeu aparecem em todas escalas de *confiança*, com mais precisão entre algumas métricas de rede e com menos entre outra. As marcas (i.e. combinações de palavras) que as variedades do português (brasileiro e europeu) deixam nos textos, entre as métricas de rede e *confiança*, são mais frequentes nos textos do que, por exemplo, as marcas que diferenciam os *períodos de edição* de uma mesma obra (Tabelas 6.6 e 6.9).

[1] Nesta tese, a expressão "país de origem do autor" é utilizada para fazer referência à variedade linguística.

6.4 *Escolas literárias*

Na história da literatura brasileira e portuguesa, algumas escolas literárias (e.g. romantismo, realismo, parnasianismo e simbolismo) coexistiram em um determinado período (Tabela 2.3). Nesse contexto de proximidade entre as escolas literárias, diferentes autores mudaram de estilo para produzir textos em conformidade com uma nova tendência literária; por exemplo, Machado de Assis foi um desses autores. Na Tabela 6.13, estão listadas duas obras que o autor escreveu durante o romantismo e duas que escreveu durante o realismo.

Tabela 6.13: Obras de Machado de Assis escritas durante o romantismo e o realismo

Nome do autor	Título da obra	Ano da edição	País de origem do autor	Escolas literárias
Machado de Assis	A mão e a luva	1874	Brasil	Romantismo
Machado de Assis	Helena	1876	Brasil	Romantismo
Machado de Assis	Memórias Póstumas de Brás Cubas	1881	Brasil	Realismo
Machado de Assis	Quincas Borba	1891	Brasil	Realismo

Fonte: Elaboração própria

Na Tabela 6.13, estão reunidas obras de um autor específico que foram editadas em intervalos máximos de dez anos entre elas e que pertencem a duas escolas literárias diferentes. Com esse conjunto de textos, o método proposto foi usado para predizer as escolas literárias dessas obras de Machado de Assis.

Na Tabela 6.14, estão sintetizados as predições referentes às escolas literárias, bem como, os parâmetros considerados. Nessa tabela, os parâmetros com maior número de predições corretas foram: textos com *mil palavras*, tipo de *corpus Original*, *dados binarizados* e predição por *Floresta aleatória*. Na Tabela 6.15, está detalhado, entre as métricas de redes e de *confianças*, onde as predições realizaram-se.

Na Tabela 6.15, os valores iguais a quatro são, entre métricas de rede e *confiança*, resultados das predições que acertaram as escolas literárias das obras de Machado de Assis. As combinações de palavras que diferem as escolas literárias de Machado de Assis aparecem em todas as escalas de *confiança* e em diferentes métricas de redes. As combinações de palavras, para esse autor, criam um padrão nos textos e tornam possível diferenciar as escolas (i.e. romantismo e realismo).

Tabela 6.14: Resultados das predição das escolas literárias nos textos de Machado de Assis

Total de palavras por texto	Tipo de *corpus*	Transformação dos dados	Métodos de predição			Total Geral
			Floresta aleatória	Naives Bayes	SVM	
100						
	Original					
		Binarizados	5			5
		Normalizados	21			21
		Sub-total	***26***			***26***
	Sem *stopwords*					
		Binarizados				0
		Normalizados	2			2
		Sub-total	***2***			***2***
	Stopwords					
		Binarizados	86		44	130
		Normalizados	102		30	132
		Sub-total	***188***		***74***	***262***
	Sub-totais de 100 palavras por texto		***216***		***74***	***290***
1.000						
	Original					
		Binarizados	**329**		131	460
		Normalizados	229		47	276
		Sub-total	***558***		***178***	***736***
	Sem *stopwords*					
		Binarizados	169		107	276
		Normalizados	122		12	134
		Sub-total	**291**		**119**	**410**
	Stopwords					
		Binarizados	255		202	457
		Normalizados	191		68	259
		Sub-total	***446***		***270***	***716***
	Sub-totais de 1.000 palavras por texto		***1.295***		***567***	***1.862***
Total Geral			**1.511**		**641**	**2.152**

Fonte: Elaboração própria

Tabela 6.15: Predições das escolas literárias com *mil palavras* por texto, tipo de *corpus Original*, *dados binarizados* e predição por *Floresta aleatória* em textos de Machados de Assis

Métricas de rede	Confianças						Total Geral
	0.0<=c<0.2	0.2<=c<0.4	0.4<=c<0.6	0.6<=c<0.8	0.8<=c<1.0	c=1.0	
Excentricidade	4	4	4	4	3	4	23
Excentricidade de entrada	4	4	4	4	4	4	24
Excentricidade de saída	1	1	4	4	3	4	17
Grau	4	4	4	4	3	4	23
Grau de entrada	4	4	4	3	4	4	23
Grau de entrada mais peso	4	4	4	3	4	4	23
Grau de saída	2	1	4	3	3	4	17
Grau de saída mais peso	1	1	4	3	3	4	16
Grau mais peso	4	4	4	4	3	4	23
Grau ponderado	4	4	4	4	3	4	23
Grau ponderado de entrada	4	4	4	4	4	4	24
Grau ponderado de saída	2	1	4	4	3	4	18
Intermediação		1	2	3	4	2	12
Laplace	4	4	4	4	3	4	23
Laplace de entrada	4	4	4	3	4	4	23
Laplace de saída	2	1	4	3	3	4	17
Total Geral	**48**	**46**	**62**	**57**	**54**	**62**	**329**

Fonte: Elaboração própria

Aluísio Azevedo, dentro do conjunto de escritores considerados, é outro autor que tem publicações em duas escolas literárias distintas. Na Tabela 6.16, há dois livros de Aluísio Azevedo editados com um ano de diferença entre eles e que pertencem a duas escolas literárias diferentes.

Tabela 6.16: Obras de Aluísio Azevedo escritas durante o romantismo e o naturalismo

Nome do autor	Título da obra	Partes	Ano de edição do texto	País de origem do autor	Escolas literárias
Aluísio Azevedo	A Condessa Vésper	I	1882	Brasil	Romantismo
Aluísio Azevedo	A Condessa Vésper	II	1882	Brasil	Romantismo
Aluísio Azevedo	O Mulato	I	1881	Brasil	Naturalismo
Aluísio Azevedo	O Mulato	II	1881	Brasil	Naturalismo

Fonte: Elaboração própria

O método proposto foi usado para analisar as predições das escolas literárias da Tabela 6.16. Os resultados e os parâmetros dos testes de predição das escolas literárias referentes às obras de Aluísio de Azevedo estão resumidos na Tabela 6.17.

Tabela 6.17: Resultados e parâmetros dos testes de predição das escolas literárias nos textos de Aluísio Azevedo

Total de palavras por texto	Tipo de *corpus*	Transformação dos dados	Métodos de predição			Total Geral
			Floresta aleatória	Naives Bayes	SVM	
100						
	Original					
		Binarizados	94		26	120
		Normalizados	76		40	116
		Sub-total	***170***		***66***	***236***
	Sem *stopwords*					
		Binarizados	60		30	90
		Normalizados	47		42	89
		Sub-total	***107***		***72***	***179***
	Stopwords					
		Binarizados	91		34	125
		Normalizados	82		32	114
		Sub-total	***173***		***66***	***239***
	Sub-totais de 100 palavras por texto		***450***		***204***	***654***
1.000						
	Original					
		Binarizados	**203**		92	295
		Normalizados	106		18	124
		Sub-total	***309***		***110***	***419***
	Sem *stopwords*					
		Binarizados	112		19	131
		Normalizados	81		11	92
		Sub-total	***193***		***30***	***223***
	Stopwords					
		Binarizados	172		109	281
		Normalizados	67		26	93
		Sub-total	***239***		***135***	***374***
	Sub-totais de 1.000 palavras por texto		***741***		***275***	***1.016***
Total Geral			**1.191**		**479**	**1.670**

Fonte: Elaboração própria

Na Tabela 6.17, os parâmetros com os maiores números de predições corretas são: *mil palavras* por texto, tipo de *corpus Original*, *dados binarizados* e método de predição *Floresta aleatória*. Na Tabela 6.18, estão, entre as métricas de redes e *confianças*, os resultados das predições das escolas literárias a partir dos textos de Aluísio de Azevedo.

Tabela 6.18: Resultados e predições das escolas literárias com parâmetros de *mil palavras* por texto, tipo de *corpus* Original, *dados binarizados* e predição por *Floresta aleatória* nos textos de Aluísio Azevedo

Métricas das redes	***Confianças***						**Total Geral**
	0.0<=c<0.2	**0.2<=c<0.4**	**0.4<=c<0.6**	**0.6<=c<0.8**	**0.8<=c<1.0**	**c=1.0**	
Excentricidade	3	4	1	4		3	15
Excentricidade de entrada	3	1	2	4		3	13
Excentricidade de saída		2		4		2	8
Grau	4	3	2	4		2	15
Grau de entrada	3	4		4		4	15
Grau de entrada mais peso	4		1	4		4	13
Grau de saída	1		1	4		4	10
Grau de saída mais peso			2	4		2	8
Grau mais peso	3	4	1	4		3	15
Grau ponderado	2	4	1	4		3	14
Grau ponderado de entrada	3	4	4	4		4	19
Grau ponderado de saída			2	4		4	10
Intermediacao			1	2			3
Laplace	2	4	2	4		3	15
Laplace de entrada	4	4	4	4		3	19
Laplace de saída		1	2	4		4	11
Total Geral	**32**	**35**	**26**	**62**	**0**	**48**	**203**

Fonte: Elaboração própria

Nesta seção, o método proposto foi usado para analisar a predição de escolas literárias em dois conjuntos de textos diferentes (i.e. textos de Machado de Assis e Aluísio de Azevedo). Entre as análises desses dois conjuntos e dentro de um contexto de proximidade entre nome dos autores e anos de edições de cada conjunto, observa-se que os parâmetros *mil palavras* por texto, tipo de *corpus Original*, *dados binarizados* e predição por *Floresta aleatória* (Tabela 6.19) apresentam características importantes para predição de escolas literárias, pois foram aqueles que resultaram em maior número de acertos entre os testes.

Tabela 6.19: Comparação entre os parâmetros com mais acertos entre as predições de escolas literárias nos conjuntos de textos de Machado de Assis e Aluísio Azevedo

Escolas literárias	**Parâmetros**				**Total de predições corretas**
	Tipo de corpus	**Método de predição**	**Transformação**	**Total de palavras por texto**	
Romantismo e Realismo	Original	Floresta aleatória	Binarização	1.000	329
Romantismo e Naturalismo	Original	Floresta aleatória	Binarização	1.000	203

Fonte: Elaboração própria

As diferenças entre os totais de predições corretas na Tabela 6.19 são resultados de características específicas de cada escritor. Dentro do método de análise proposto, as combinações de três palavras que cada autor usou são detectadas pelas regras de associação, formalizadas nas redes e avaliadas pelas métricas de redes e métodos de predição. Os repertórios de palavras que os autores usam são diferentes; e as combinações que os autores fazem de seu repertório de palavras variam em conformidade com contextos específicos.

O padrão de combinações de palavras que Machado de Assis faz em seus textos românticos são distintas daquelas presentes em seus textos realistas; o mesmo ocorre para Aluísio de Azevedo para seus textos românticos e naturalistas. As diferenças de combinações nos textos de Machado de Assis são mais vezes detectadas pelo método proposto do que as combinações de palavras feitas por Aluísio Azevedo em suas obras.

As combinações de palavras particulares a cada autor são detectadas pelas métricas de *confiança* das regras de associação. Na comparação entre as predições baseadas nas métricas de *confiança* de Machado de Assis e Aluísio Azevedo (Tabelas 6.15 e 6.18), as combinações de palavras que Machado de Assis faz (*confianças* (c) com intervalos {0.4>c>=0.6; c=1.0}, Tabela 6.15) são mais relevantes para distinguir as escolas literárias a que suas obras estão filiadas do que as escolhas de palavras que Aluísio Azevedo faz com as mesmas confianças (Tabela 6.18) para distinguir as escolas literárias a que seus textos pertencem. Esses dados ilustram que, para além da identificação das marcas dos autores, é possível identificar as marcas que podem caracterizar diferentes escolas literárias (quando comparadas obras do mesmo autor associadas a períodos literários distintos).

6.5 Autores do romantismo

O método proposto caracteriza os autores a partir de combinações de três palavras que aparecem em suas obras. Em um *corpus* com obras que compartilham um mesmo contexto (e.g. período de edição, língua e escola literária), é esperado que entre os textos dos autores existam algumas combinações de palavras que são comuns a todos, por trabalharem com uma mesma língua e uma mesma escola literária. É esperado também que existam combinações de palavras que são escolhas próprias de cada autor. Nos textos de um escritor, essas escolhas de palavras (i.e. combinações de palavras) são encontradas em diferentes proporções e quantidades. No método aqui proposto, as regras de associação identificam, entre premissas e conclusões, por meio da métrica de *confiança*, as combinações de palavras recorrentes nos textos dos autores. O método proposto, ao comparar as seis escalas de *confiança* (c) {0.0=<c<0.2; 0.2=<c<0.4; 0.4=<c<0.6; 0.6=<c<0.8; 0.8=<c<1.0; c=1.0} entre os autores, busca conjuntos de palavras em cada uma dessas escalas para inferir a que autor pertence um texto sem identificação de autoria.

Na Tabela 6.20, há cinco autores brasileiros do romantismo que publicaram suas obras antes de 1900. Com o método proposto, não é viável analisar as autorias do conjunto de texto da Tabela 6.20, pois o método não consegue predizer com exatidão as autorias de todos os textos quando o *corpus* tem mais que dois autores.

Para comparar as combinações de palavras entre os autores, o método busca, entre as métricas de redes e *confianças*, valores que os diferenciem; no entanto essa comparação só ocorre quando os autores geram combinações de palavras em uma mesma escala de *confiança* e, algumas vezes, isso não acontece. Depois de processar as obras apresentadas na Tabela 6.20, o método proposto, no melhor cenário, só acertou o nome de quatro autores dos dez textos testados. A Tabela 6.21 exibe os parâmetros com os maiores números de predições corretas encontradas.

Tabela 6.20: Obras com diferentes autores do romantismo brasileiro escritas antes de 1.900

Nome dos autores	**Título das obras**	**Ano da edição**	**País de origem**	**Escola literária**
Bernardo Guimarães	O índio Afonso	1873	Brasil	Romantismo
Bernardo Guimarães	A Ilha Maldita	1879	Brasil	Romantismo
Joaquim Manuel de Macedo	A misteriosa	1872	Brasil	Romantismo
Joaquim Manuel de Macedo	Os quatro pontos cardeais	1872	Brasil	Romantismo
Machado de Assis	A mão e a luva	1874	Brasil	Romantismo
Machado de Assis	Helena	1876	Brasil	Romantismo
Teixeira e Sousa	O filho do pescador	1843	Brasil	Romantismo
Teixeira e Sousa	Gonzaga ou a conjuração do Tiradentes	1848	Brasil	Romantismo
Visconde de Taunay	A mocidade de Trajano	1871	Brasil	Romantismo
Visconde de Taunay	No declínio	1898	Brasil	Romantismo

Fonte: Elaboração própria

Tabela 6.21: Parâmetros com maior número de predições corretas entre as obras literárias do romantismo brasileiro da Tabela 6.20

Total de palavras	**Tipo de corpus**	**Transformação dos dados**	**Métodos de predição**	**Métricas das redes**	**Confianças**	**Predições corretas**	**Total de textos**
1.000	Original	Binarizados	SVM	Intermediação	1.0>c>=0.8	4	10
1.000	Original	Binarizados	SVM	Laplace de saída	0.2>c>=0.0	4	10

Fonte: Elaboração própria

Por não ser viável processar, concomitantemente, os textos de mais de dois autores, foram analisados, inicialmente, os textos de Machado de Assis e de Bernardo Guimarães[2]. Ao avaliar, com o método proposto, os textos desses autores, observou-se (Tabela 6.22) que o maior número de predições corretas aconteceu com os parâmetros *mil palavras* por texto,

[2]O critério para a escolha desses dois autores ocorreu de modo aleatório.

tipo de *corpus Original*, *binarização de dado*s e método de predição por *Floresta aleatória*.

Tabela 6.22: Resultados e parâmetros dos testes de predição entre os textos de Machado de Assis e Bernardo Guimarães

Total de palavras por texto	Tipo de *corpus*	Transformação dos dados	Métodos de predição			Total Geral
			Floresta aleatória	Naives Bayes	SVM	
100						
	Original					
		Binarizados	220		207	427
		Normalizados	171		89	260
		Sub-total	***391***		***296***	***687***
	Stopwords					
		Binarizados	218		156	374
		Normalizados	162		49	211
		Sub-total	***380***	***0***	***205***	***585***
	Sub-totais de 100 palavras por texto		***771***	***0***	***501***	***1.272***
1.000						
	Original					
		Binarizados	**358**		354	712
		Normalizados	318		131	449
		Sub-total	***676***	***0***	***485***	***1.161***
	Sem *Stopwords*					
		Binarizados	244		237	481
		Normalizados	225		147	372
		Sub-total	***469***	***0***	***384***	***853***
	Stopwords					
		Binarizados	344		280	624
		Normalizados	301		228	529
		Sub-total	***645***	***0***	***508***	***1.153***
	Sub-totais de 1.000 palavras por texto		***1.790***	***0***	***1.377***	***3.167***
Total Geral			**2.561**	**0**	**1.878**	**4.439**

Fonte: Elaboração própria

Na Tabela 6.23, estão, entre métricas de redes e *confianças*, as predições de autoria dos textos de Machado de Assis e Bernado de Guimarães. Nessa tabela, o valor quatro ilustra que o método inferiu, corretamente, os autores de todos os textos. Ainda nessa tabela, observa-se que, entre as métricas de rede e *confiança*, o método, de um máximo de trezentos e oitenta e quatro acertos possíveis, acertou as predições de trezentas e cinquenta e quatro autorias.

Tabela 6.23: Resultados e predições das autoria dos textos de Machado de Assis e Bernardo de Guimarães com parâmetros de *mil palavras* por texto, tipo de *corpus Original*, *dados binarizados* e predição por *Floresta aleatória*

Métricas de rede	Confianças						Total Geral
	0.0<=c<0.2	0.2<=c<0.4	0.4<=c<0.6	0.6<=c<0.8	0.8<=c<1.0	c=1.0	
Excentricidade	4	4	4	4	4	4	24
Excentricidade de entrada	4	4	4	4	1	4	21
Excentricidade de saída	4	4	4	4	4	4	24
Grau	4	4	4	4	4	4	24
Grau de entrada	4	4	4	4	1	4	21
Grau de entrada mais peso	4	4	4	4	1	4	21
Grau de saída	4	4	4	4	4	4	24
Grau de saída mais peso	4	4	4	4	4	4	24
Grau mais peso	4	4	4	4	4	4	24
Grau ponderado	4	4	4	4	4	4	24
Grau ponderado de entrada	4	4	4	4	1	4	21
Grau ponderado de saída		3	4	4	4	4	19
Intermediacao		4	4	4	1	1	14
Laplace	4	4	4	4	4	4	24
Laplace de entrada	4	4	4	4	1	4	21
Laplace de saída	4	4	4	4	4	4	24
Total Geral	**56**	**63**	**64**	**64**	**46**	**61**	**354**

Fonte: Elaboração própria

Ao considerar que os parâmetros *mil palavras* por texto, tipo de *corpus Original*, *binarização de dados* e método de predição por *Floresta aleatória* foram relevantes para a verificação de autoria de Machado de Assis e Bernardo Guimarães, esses mesmos parâmetros foram, então, usados para avaliar as autorias dos textos na Tabela 6.20. Os textos dos autores na Tabela 6.20 foram organizados considerando todas as combinações de pares de autores; a cada par, as predições de autoria foram realizadas. A Tabela 6.24 apresenta as somatórias das predições de autoria que o método inferiu corretamente.

Tabela 6.24: Total de predições corretas inferidas pelo método proposto para verificar a autoria dos pares de autores brasileiros

	Bernardo Guimarães	Joaquim Manuel de Macedo	Machado de Assis	Teixeira e Sousa	Visconde de Taunay
Bernardo Guimarães		144	344	196	75
Joaquim Manuel de Macedo			351	187	179
Machado de Assis				363	311
Teixeira e Sousa					125
Visconde de Taunay					

Percentual de acertos (p)

(total de acerto / número de acertos possíveis)

75% < p <= 100%

50% < p <= 75%

25% < p <= 50%

0 <= p <= 25%

Fonte: Elaboração própria

Tabela 6.25: Resultados das análises de autoria entre pares de autores e a soma das predições encontradas

Combinação de autores em pares		Excentricidade	Excentricidade de entrada	Excentricidade de saída	Grau	Grau de entrada	Grau de entrada mais peso	Grau de saída	Grau de saída mais peso	Grau mais peso	Grau ponderado	Grau ponderado de entrada	Grau ponderado de saída	Intermediação	Laplace	Laplace de entrada	Laplace de saída	Total Geral
Autor 01	**Autor 02**																	
Bernardo Guimarães																		
	Joaquim Manuel de Macedo	10	7	8	7	12	9	2	4	16	12	10	5	4	17	11	10	**144**
	Machado de Assis	24	22	20	24	24	23	19	18	24	24	23	19	13	24	24	19	**344**
	Teixeira e Sousa	15	11	15	13	12	8	10	12	14	19	11	15	6	15	10	10	**196**
	Visconde de Taunay	4	4	7	3	3	4	9	4	3	6	3	5	8	2	3	7	**75**
Joaquim Manuel de Macedo																		
	Machado de Assis	24	22	20	24	22	22	20	22	24	24	22	22	15	24	22	22	**351**
	Teixeira e Sousa	14	6	15	11	8	4	10	17	18	15	7	14	10	13	9	16	**187**
	Visconde de Taunay	15	17	5	8	12	18	4	7	12	12	16	9	7	11	17	9	**179**
Machado de Assis																		
	Teixeira e Sousa	24	24	22	24	24	23	22	22	24	24	24	23	14	24	24	21	**363**
	Visconde de Taunay	22	20	16	22	21	20	16	16	22	24	22	16	16	20	22	16	**311**
Visconde de Taunay																		
	Teixeira e Sousa	8	9	8	7	9	7	10	11	6	10	8	8	4	7	7	6	**125**

Fonte: Elaboração própria

Na Tabela 6.25, entre as métricas de redes, estão os totais de predições corretas que o método realizou. Nessa tabela, o valor máximo de predições corretas que o método pode encontrar, em cada métrica de rede, é vinte e quatro; esse valor corresponde ao total de textos analisados multiplicado pelo total de escalas de *confiança*. Por meio do método proposto, observa-se que, entre os textos de Machados de Assis e Teixeira e Souza, existe o maior número de predições corretas (trezentas e sessenta e três) (Tabelas 6.24 e 6.25); esse valor está detalhado, entre regras de redes e *confianças*, na Tabela 6.26. Na Tabela 6.26, o número quatro indica que o método proposto acertou os autores de todos os textos analisados.

Neste método, encontrar um conjunto de parâmetros (i.e. total de palavras, tipo de *corpus*, etc.) que mais bem represente as características de um autor (i.e. combinações de palavras nas redes) implica localizar, entre diferentes arranjos de parâmetros, predições que infiram, relevantemente, mais vezes o real autor de um texto. Na Tabela 6.24, os parâmetros *mil palavras* por texto, tipo de *corpus Original*, *binarização de dados* e método de predição por *Floresta aleatória* mostram-se relevantes para analisar textos de Machado

de Assis. Esses parâmetros mostram-se relevantes ao diferenciarem, com mais de setenta e cinco porcento de acertos, os textos de Machado de Assis dos demais autores no *corpus* da análise.

Tabela 6.26: Resultados das predições entre textos de Teixeira e Souza e Machado de Assis

Métricas das redes	Confianças						Total Geral
	0.0>c>=0.2	0.2>c>=0.4	0.4>c>=0.6	0.6>c>=0.8	0.8>c>=1.0	c=1.0	
Excentricidade	4	4	4	4	4	4	24
Excentricidade de entrada	4	4	4	4	4	4	24
Excentricidade de saída	4	3	3	4	4	4	22
Grau	4	4	4	4	4	4	24
Grau de entrada	4	4	4	4	4	4	24
Grau de entrada mais peso	4	4	3	4	4	4	23
Grau de saída	4	2	4	4	4	4	22
Grau de saída mais peso	4	3	3	4	4	4	22
Grau mais peso	4	4	4	4	4	4	24
Grau ponderado	4	4	4	4	4	4	24
Grau ponderado de entrada	4	4	4	4	4	4	24
Grau ponderado de saída	4	3	4	4	4	4	23
Intermediacao		1	1	4	4	4	14
Laplace	4	4	4	4	4	4	24
Laplace de entrada	4	4	4	4	4	4	24
Laplace de saída	4	2	3	4	4	4	21
Total Geral	**60**	**54**	**57**	**64**	**64**	**64**	**363**

Fonte: Elaboração própria

Ao usar as redes para analisar as autorias dos textos, o método proposto avalia também as relações entre as palavras que compõem os marcadores de autoria de cada escritor; não é apenas inferir que, se um conjunto x de palavras aparecer em um texto, então o texto pertence ao Autor01 ou Autor02. O método proposto analisa, entre as dezesseis métricas de redes, as palavras dos marcadores de autoria de cada escritor e aponta em quais métricas esses marcadores melhor se manifestam.

6.6 Considerações sobre o capítulo

Na Seção 6.1, o método proposto foi usado para avaliar os limites (i.e. quantidades máximas de dados para que o algoritmo funcione bem) e os ajustes nos algoritmos de predição e construção de regras de associação (i.e. parâmetros internos desses algoritmos). Configurou-se o método de modo a reduzir o tempo de processamento dos dados e diminuir o volume de informações processadas. Esses ajustes foram necessários porque a geração de regras de associação cria grandes volumes de informações e torna a análise dos dados lenta. Foi nessa etapa do trabalho que os parâmetros iniciais do método proposto foram configurados.

Na Seção 6.2, os anos das edições de algumas obras literárias foram o objeto da pesquisa. Nessa etapa, foram comparadas obras literárias com anos de edição diferentes, mas com

mesmas autorias. Nessa análise, o método proposto localizou, entre os parâmetros, uma configuração relevante para a análise dos *períodos das edições*. Essa configuração mostrou-se relevante para a análise dos textos de Machado de Assis e de João Carlos de Medeiros Pardal Mallet.

Na Seção 6.3, o método proposto foi usado para investigar a predição de variedades da língua portuguesa. Nessa etapa, foram comparadas obras escritas por autores brasileiros e portugueses. Os resultados mostram que as variedades de português, brasileiro e europeu, são detectadas pelo método proposto e devem ser consideradas em uma análise de autoria.

Na Seção 6.4, investigou-se a predição de escolas literárias considerando obras escritas pelo mesmo autor em diferentes períodos. A configuração de parâmetros que caracterizam a distinção de escolas literárias foi a mesma nas comparações dos textos de Machado de Assis e de Aluísio Azevedo. No entanto, entre as métricas de redes e *confianças*, as marcas que esses autores fazem para diferenciar as escolas literárias se realizam em escalas de *confiança* diferentes.

Na última (Seção 6.5), o método proposto foi usado para analisar a autoria de alguns textos do romantismo escritos por autores brasileiros. Apesar de o método não ser capaz de analisar conjuntos com textos de mais de dois autores, a análise de autoria em pares de escritores apresenta relevância para a identificação de parâmetros que diferenciam um autor de outros.

Tabela 6.27: Relevância dos parâmetros para a verificação de autoria e os contextos de produções textuais

Parâmetros	**Seção 6.2**	**Seção 6.3**	**Seção 6.4**	**Seção 6.5**
	Diferentes edições de uma mesma obra	**Variedades de língua portuguesa dos autores**	**Escolas literárias**	**Autores do romantismo**
Total de palavras	1.000	1.000	1.000	1.000
Tipo de *corpus*	Original	Original	Original	Original
Transformação dos dados	Binarização	Binarização	Binarização	Binarização
Métodos de predição	SVM	Naive Bayes	Floresta aleatória	Floresta aleatória
Métricas de redes	Variam entre os autores	Todas	Variam entre os autores	Variam entre os autores
Escalas de *confiança*	Variam entre os autores	Todas	Variam entre os autores	Variam entre os autores

Fonte: Elaboração própria

Na Tabela 6.27, estão ilustradas as relevâncias dos parâmetros entre os testes de verificação de autoria e produção conforme discutido nas Seções 6.2, 6.3, 6.4 e 6.5. Nessa tabela, relevância representa maior número de predições inferidas, corretamente, entre os testes. Os resultados apresentados neste capítulo mostram que o método proposto no Capítulo 5 é

uma contribuição relevante para AA, pois esclarece que variedades de uma mesma língua, mais especificamente, o português brasileiro e o português europeu, deixam marcas nos textos; mostra, também, respeitando os respectivos contextos, ser capaz de inferir *período das edições*, escolas literárias e autorias. Nos textos, as marcas deixadas pelos autores e pelos contextos de produção podem ser percebidas nas relações que existem entre as métricas de redes e as escalas de *confianças*; quanto mais predições corretas existirem entre as métricas de rede e as escalas de *confianças*, maiores serão as dissemelhanças entre os objetos estudados.

Capítulo Sete

Considerações finais

A análise de autoria (AA) de textos investiga marcas presentes em um documento cuja autoria é desconhecida comparando-as com as marcas presentes em conjuntos de textos de autores conhecidos. A partir das evidências coletadas, a AA infere se a obra investigada pertence ou não ao autor conhecido. Identificar o escritor de uma obra é relevante, pois evidencia, por exemplo, falsas atribuições de textos a um autor. Esta tese abordou métodos de análise de autoria. Mais especificamente, esta pesquisa tratou da análise de verificação de autoria e de elementos do contexto de produção de uma obra (i.e. *período de edição, variedade linguística e escola literária*).

7.1 Conclusões

O objetivo desta tese foi elaborar um método de análise para verificação de autoria e contextos de produção (e.g. *período de edição, variedade linguística e escolas literárias*) de obras literárias. Propõe-se, com o auxílio da Teoria e Ciência das Redes, um algoritmo baseado em combinações de palavras (Hipótese 1.1). Consolidar, nas redes, as combinações de palavras extraídas dos textos foi uma estratégia relevante, pois foi possível encontrar, entre as métricas de redes e as escalas de *confiança*, marcas que caracterizam, de maneiras diferentes, os *períodos de edição* de uma mesma obra; as variedades de língua portuguesa (brasileira e europeia); as escolas literárias; e as autorias dos textos (e.g. autores do romantismo).

Organizar os textos considerando três formas de combinações de palavras (*Original*, Sem *stopwords* e *Stopwords*) foi importante para a análise dos resultados (Hipótese 1.2). As combinações dessas palavras mostraram diferentes relevâncias para as análises dos anos das edições; das variedades linguísticas; das escolas literárias e de autorias. De modo geral, o tipo de *corpus Original* teve o maior número de predições corretas entre todas as análises; ou seja, a partir do conjunto que combina palavras lexicais e gramaticais, identificaram-se mais vezes, entre os parâmetros analisados, as marcas que diferenciam as autorias e os elementos que caracterizam os contextos de produção. Apesar de a retirada de palavras (ora foram retiradas as palavras gramaticais, ora as lexicais constituindo tipos de *corpus* Sem *stopwords* e *Stopwords*, respectivamente) gerar menor número de predições corretas entre os parâmetros analisados, as marcas que os autores deixam nos textos ainda podem ser encontradas nesses conjuntos.

Ao explorar as predições com diferentes totais de palavras por texto (*cem* e *mil palavras*), o intuito foi avaliar o método em relação a alguns limites. Conhecer os limites mínimos de palavras por texto é relevante, por exemplo, para saber se o método proposto tem condições para avaliar conjuntos de textos com poucas palavras. No geral, a quantidade relevante de palavras foi mil por texto; com esse parâmetro foi possível encontrar marcas para diferenciar as autorias e contextos de produção nos testes realizados. No entanto, em alguns contextos, a marca que o autor deixa no texto é tão significativa que, com cem palavras, é possível extrair informações das redes. Na Tabela 6.5, por exemplo, com parâmetros parecidos (i.e. tipo de *corpus Original*, *dados binarizados* e predição por *Floresta aleatória*), as análises com *cem* e com *mil palavras* realizaram o mesmo número de predições corretas ao inferir os *períodos de edição* de obras de Machado de Assis; por outro lado, com esses mesmos parâmetros, na análise de predição de *períodos de edição* da obra de João Carlos de Medeiros Pardal Mallet, textos com *cem palavras* não foram relevantes para a análise de suas obras (Tabela 6.8).

Na Hipótese 2.1, assume-se que os valores atribuídos pelas métricas aos vértices das redes, quando próximos entre si, registram características que evidenciam o mesmo autor (e, consequentemente, diferenciam autores). Essa hipótese foi confirmada, pois, neste trabalho, a normalização dos dados transformou, em uma escala de zero a um, os valores atribuídos aos vértices nas redes e, por meio dos algoritmos de predição, o método buscou diferenciar os autores e os contextos de produção a partir da proximidade que existe entre os valores normalizados dos vértices. Embora entre os testes realizados a *normalização* tenha inferido autores e contextos, a *binarização* apresentou maiores números de predições corretas. Ao binarizar os dados, os algoritmos de predição buscam, entre as métricas dos vértices, subconjuntos de palavras que estão presentes ou ausentes nas métricas das redes. Neste método, por meio de algoritmos de predição, analisar a presença ou ausência de vértices (i.e. palavras) entre as métricas das redes é mais relevante para inferir um maior número de predições corretas entre os parâmetros analisados do que analisar as proporções de valores que existem entre as métricas dos vértices. As métricas de redes contribuem para a análise de método proposto neste trabalho, ao destacarem conjuntos de vértices que são semelhantes entre as redes de um mesmo autor, ou de um mesmo contexto de produção.

Os algoritmos de predição (SVM, Floresta aleatória e Naive Bayes) apresentaram diferentes relevâncias para o método aqui proposto. Na Tabela 6.27, observa-se que, a partir do algoritmo de predição, a parametrização do método proposto começa a relacionar os arranjos dos parâmetros às diferentes aplicações (i.e. à verificação de autoria e aos contextos de produções). Respeitando os parâmetros com maiores números de predições corretas (i.e. *mil palavras* por texto; tipo de *corpus Original* e *dados binarizados*), para a análise de edições de uma mesma obra (Seção 6.2), o SVM foi o algoritmo que mais vezes inferiu, corretamente, a que período pertenciam as edições dos textos analisados.

Para analisar as *variedades de língua portuguesa* dos autores (Seção 6.3), o *Naive Bayes* inferiu, corretamente, mais vezes as variedades linguísticas. Para a análise das *escolas literárias* e autores do romantismo (Seções 6.4 e 6.5, respectivamente), o *algoritmo Floresta aleatória* foi o que mais vezes inferiu, corretamente, de um lado, as escolas literárias e, de outro, os autores.

No modelo proposto nesta tese, as combinações de palavras que existem nos textos dos escritores são formalizadas em regras de associação; essas regras são agrupadas entre seis escalas de *confiança* e transformadas em redes (cf. Figura 5.3). Cada escala de *confiança* agrupa, nas redes, subconjuntos de combinações que são recorrentes nos textos dos autores. As combinações que os autores fazem de seus repertórios de palavras variam em conformidade com os contextos de produção textual.

Na verificação de autoria, quando as combinações de palavras que diferenciam dois autores são muito grandes, essas distinções aparecem em diferentes escalas de *confiança* e diferentes métricas de redes (Hipótese 2.2). Na Tabela 6.26, entre métricas de redes e *confiança*, as diferenças entre os textos de Machado de Assis e Teixeira e Souza foram encontradas, pelo método proposto, em todas as escalas de *confiança* e entre todas as métricas de rede. Quando as combinações de palavras que distinguem os autores são menores, as predições corretas de autoria manifestam-se com menores valores entre as métricas de redes (cf. comparação entre Bernardo de Guimarães e Visconde de Taunay na Tabela 6.25).

Na verificação de diferentes edições de uma mesma obra (Seção 6.2), as predições de *período de edição*, realizadas entre obras de Machado de Assis e Aluísio Azevedo, ilustram, entre as métricas de redes e *confianças* (Tabela 6.15 e 6.18), que as alterações feitas nas obras desses dois autores manifestam-se de modo mais significativo em diferentes escalas de *confiança*. Nas obras de Machado de Assis (Tabela 6.18), as alterações que mais vezes foram detectadas pelo método ocorreram entre *confianças* iguais a um (i.e. palavras que cem porcento das vezes aparecem relacionando os antecedentes aos consequentes nas regras de associação). Nessa verificação, o tipo de *corpus* relevante foi o *Original* (Hipótese 3.1). As análises de predições nos anos de edição das obras de um autor contribuem para a AA fornecendo informações que complementam a verificação de autoria inferindo o período que o texto foi modificado. Um texto original e inédito de um autor tende a apresentar, entre as combinações de palavras, um comportamento coerente com outros textos do mesmo autor publicados no mesmo período. O método proposto apresenta resultados que contribuem para esse tipo de verificação.

Na verificação de variedade de língua portuguesa (Seção 6.3), as diferenças entre o português brasileiro e europeu são representadas, entre métricas de redes e *confianças*, pelos totais dos testes que inferiram, corretamente, as variedades da língua. Na Tabela 6.12,

observa-se que as diferenças entre as variedades da língua manifestam-se mais com combinações de palavras que ocorrem com *confianças* entre 0.6 e 0.8 (Hipótese 3.2).

Nesta tese, a verificação de autoria e produção textual foi avaliada por um novo método de análise de autoria. Os resultados deste trabalho ilustram que, por meio de métricas de redes e *confianças*, o método infere parâmetros relevantes para verificação de autorias e produções textuais relacionadas a cada autor, *períodos de edição*, variedades da língua portuguesa e escolas literárias. Demonstra-se que, a partir das combinações de palavras nos textos, é possível extrair informações particulares dos autores e de suas obras.

7.2 Contribuições

A principal contribuição desta tese é apresentar um método de trabalho para auxiliar pesquisas na análise de autoria. Esse novo método, ao avaliar combinações de palavras em redes, encontra, entre diferentes parâmetros (i.e. total de palavras, tipo de *corpus*, transformação de dados e algoritmos de predição), marcas que caracterizam e diferenciam os autores e o contexto relativos à produção textual de cada um. Além disso, o *algoritmo Extrair Corpus* é uma contribuição para os estudos interessados em extrair conjuntos de textos que compartilhem um mesmo *períodos de edição*, variedade linguística (país de origem), escolas literárias e totais de palavras.

7.3 Oportunidades para pesquisas e desenvolvimentos futuros

A intenção, ao explorar o método proposto a partir de diferentes arranjos de parâmetros (i.e. total de palavras por texto, tipos de *corpus*, transformação de dados, métodos de predição, métricas de redes e escalas de *confiança*), foi encontrar um arranjo de parâmetros que mais bem inferissem as verificações de autoria e contextos da produção textual.

Em investigações futuras, pretende-se usar o método proposto para analisar, em um número maior de textos do que o considerado aqui, os parâmetros que mais bem distinguem: (i) *período de edição*; (ii) *variedades da língua portuguesa*; (iii) características de *escola literárias*; e (iv) marcas de autoria de cada escritor. Isso poderá reforçar a relevância dos parâmetros propostos (Tabela 6.27).

Referências

ABUHAMAD, Mohammed; RHIM, Ji su; ABUHMED, Tamer; ULLAH, Sana; KANG, Sanggil; HUNNYANG, Dae. Code authorship identification using convolutional neural networks. *Future Generation Computer Systems*, v. 95, p. 104–115, 2019.

ADAMOVIC, S.; MISKOVIC, V.; MILOSAVLJEVIC, M.; SARAC, M.; VEINOVIC, M. D. Automated language-independent authorship verification (for indo-european languages). *Journal of the Association for Information Science and Technology*, v. 70, n. 8, p. 858–871, 2019.

AGRAWAL, Rakesh; SRIKANT, Ramakrishnan. Fast algorithms for mining association rules. In: *VLDB Conference*. Santiago: IBM Reserch Report RJ9839, 1994.

AGUIAR, M. D. S. F. D. *Redes de palavras em textos escritos: uma análise da linguagem verbal utilizando redes complexas*. Dissertação (Mestre em Física) — Programa de Pós-Graduação em Física, Universidade Federal da Bahia, Salvador, 2009.

ALHARTHI, Haifa; INKPEN, Diana; SZPAKOWICZ, Stan. A survey of book recommender systems. *J Intell Inf Syst*, n. 51, p. 139–160, 2018.

AMANCIO, D. R. *Classificação de textos com redes complexas*. Tese (Doutor em Ciências) — Programa de Pós-Graduação em Física do Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2013.

AMANCIO, Diego R. Authorship recognition via fluctuation analysis of network topology and word intermittency. *Journal of Statistical Mechanics Theory and Experiment*, n. 1-20, 2015.

AMANCIO, Diego Raphael. Complex network approach to stylometry. *PLOS ONE*, p. 1–21, 2015.

AMANCIO, Diego R.; ALTMANN, Eduardo G; OLIVEIRA, Osvaldo N; COSTA, Luciano da F. Comparing intermittency and network measurements of words and their dependence on authorship. *New Journal of Physics*, v. 13, n. 12, p. 1–23, 2011.

AMANCIO, Diego Raphael; OLIVEIRA-JR, Osvaldo N.; COSTA, Luciano da Fontoura. Identification of literary movements using complex networks to represent texts. *New Journal of Physics*, v. 14, p. 1–15, 2012.

AMANCIO, Diego R.; OLIVEIRA-JR, Osvaldo N.; COSTA, Luciano da F. Structure-semantics interplay in complex networks and its effects on the predictability of similarity in texts. *Physica A: Statistical Mechanics and its Applications*, v. 391, n. 18, p. 4406–4419, 2012.

AMANCIO, D. R.; SILVA, F. N.; COSTA, L. D. F. Concentric network symmetry grasps authors' styles in word adjacency networks. *EPL (Europhysics Letters)*, v. 110, 2015.

ANDRADE, Francine de Souza. Linguística de *corpus* como ferramenta para identificar o estilo do tradutor: peculiaridade, autoria e formação de tradutores. *Revele*, n. 8, p. 121–137, 2015.

ANTIQUEIRA, Lucas; NUNES, Maria das Graças V.; OLIVEIRA-JR, Osvaldo N. de; COSTA, Luciano da F. Modelando textos como redes complexas. In: *XXV Congresso da Sociedade Brasileira de Computação*. São Leopoldo (RS): UNISINOS, 2005. p. 2089–2098.

ANTIQUEIRA, Lucas; NUNES, Maria das Gracas V.; OLIVEIRA, Osvaldo N.; COSTA, Luciano da F. Strong correlations between text quality and complex networks features. *Physica A*, v. 373, p. 811–820, 2007.

ANWAR, Waheed; BAJWA, Imran Sarwar; RAMZAN, Shabana. Design and implementation of a machine learning-based authorship identification model. *Scientific Programming*, Hindawi, p. 1–14, 2019.

AREFIN, Ahmed Shamsul; VIMIEIRO, Renato; RIVEROS, Carlos; CRAIG, Hugh; MOSCATO, Pablo. An information theoretic clustering approach for unveiling authorship affinities in shakespearean era plays and poems. *PLoS ONE*, v. 9, 2014.

ARUN, R.; SURESH, V.; MADHAVAN, C. V. Stopword graphs and authorship attribution in text corpora. In: *IEEE International Conference on Semantic Computing*. Berkeley, CA, USA: IEEE, 2009. p. 192–196.

AVELINO, Guilherme; PASSOS, Leonardo; HORA, Andre; VALENTE, Marco Tulio. Measuring and analyzing code authorship in 1 + 118 open source projects. *Science of Computer Programming*, v. 176, p. 14–32, 2019.

BAGNO, Marcos. *Nada na língua é por acaso: por uma pedagogia da variaćão linguística*. São Paulo: Parábola Editorial, 2007.

BARABASI, Albert-Laszlo. Scale-free networks: A decade and beyond. *Science*, v. 325, n. 5939, p. 412–413, 2009.

BARBON, Sylvio; CAMPOS, Gabriel F.C.; TAVARES, Gabriel M.; IGAWA, Rodrigo A.; PROENçA, Mario L.; GUIDO, Rodrigo Capobianco. Detection of human, legitimate bot, and malicious bot in online social networks based on wavelets. *ACM Trans. Multimedia Comput. Commun. Appl.*, v. 14, n. 1, p. 36–43, 2018.

BARRAT, Alain; BARTHELEMY, Marc; PASTOR-SATORRAS, Romualdo; VESPIGNANI, Alessandro. The architecture of complex weighted networks. *PNAS*, The National Academy of Sciences of the USA, v. 101, n. 11, p. 3747–3752, 2004.

BARUFALDI, Bruno; JUNIOR, Milton Marques; SANTANA, Eduardo Freire; POEL, JanKees van der; FILHO, Jose Rogerio Bezerra Barbosa; BATISTA, Leonardo Vidal. Classificação automática de textos por período literário utilizando compressão de dados através do ppm-c. *Linguamática*, v. 2, p. 35–43, 2010.

BELINE, Ronald. A variação linguística. In: *Introdução a linguística I - Objetos teóricos*. São Paulo: Contexto, 2010. p. 121–140.

BOUANANI, Sara El Manar El; KASSOU, Ismail. Authorship analysis studies: A survey. *International Journal of Computer Applications*, v. 86, n. 12, p. 22–29, 2014.

BRENNAN, Michael; AFROZ, Sadia; GREENSTADT, Rachel. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security*, v. 15, n. 3, p. 12–22, 2012.

BRENNAN, Michael; GREENSTADT, Rachel. Practical attacks against authorship recognition techniques. In: *21st Innovative Applications of Artificial Intelligence Conference*. Pasadena, CA, USA: IAAI-09, 2009.

CALDAS-COULTHARD, Carmen Rosa. O que é a linguística forense? *ReVEL*, v. 12, n. 23, p. 1–6, 2014.

CALDEIRA, S. G. M. *Caracterização da rede de signos lingüísticos: um modelo baseado no aparelho psíquico de Freud*. Dissertação (Mestre em Modelagem Computacional) — Centro de Pesquisa e Pós-Graduação da Faculdade Visconde de Cairu, Salvador, 2005.

CAN, Fazli; PATTON, Jon M. Change of writing style with time. *Computers and the Humanities*, v. 38, p. 61–82, 2004.

CANCHO, Ramon Ferrer; SOLÉ, Richard V. The small world of human language. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, The Royal Society, v. 268, n. 1482, p. 2261–2265, 2001.

CARDEI, Claudia; REBEDEA, Traian. Detecting sexual predators in chats using behavioral features and imbalanced learning. *Natural Language Engineering*, v. 23, n. 4, p. 586–616, 2017.

CAVALCANTI, Camillo Baptista Oliveira. Moderna perspectiva das classes de palavras. In: *Congresso Nacional de Linguística e Filologia*. [S.l.: s.n.], 2004. v. 8.

CAVIQUE, Luís. Graph-based structures for the market baskets analysis. *Investigação Operacional*, n. 24, p. 233–246, 2004.

CAVIQUE, Luis. A network algorithm to discover sequential patterns. In: *EPIA 2007*. Guimaraes: LNAI 4874, 2007. p. 406–414.

CAVIQUE, Luis. A scalable algorithm for the market basket analysis. *Journal of Retailing and Consumer Services*, v. 14, n. 6, p. 400–407, 2007.

CHAKRABORTY, T.; CHOUDHURY, P. Authorship identification in bengali language: A graph based approach. In: *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. San Francisco: IEEE/ACM, 2016. p. 443–446.

CHASKI, Carole E. Empirical evaluations of language-based author identification techniques. *The International Journal of Speech, Language and Law: Forensic Liguistics*, v. 8, n. 1, p. 1–65, 2001.

CHOMSKY, Avram Noam. *O conhecimento da língua: sua natureza, origem e uso*. Lisboa: Caminho, 1994.

CHOMSKY, Avram Noam. *Estruturas sintáticas*. [S.l.]: Editora Vozes, 2018. (De Linguística). ISBN 9788532659231.

CHOMSKY, Avram Noam. *Sobre natureza e linguagem*. São Paulo: WMF Martins Fontes, 2019.

CONG, Jin; LIU, Haitao. Approaching human language with complex networks. *Physics of Life Reviews*, v. 4, n. 598-618, 2014.

COYOTL-MORALES, Rosa María; VILLASEñOR-PINEDA, Luis; MONTES, Manuel; ROSSO, Paolo. Authorship attribution using word sequences. In: *11th Iberoamerican Congress in Pattern Recognition*. Cancun-Mexico: CIARP 2006, 2006. v. 4225, p. 844–853.

CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. *InterJournal*, Complex Systems, p. 1695, 2006.

CYRINO, Sonia. Objetos nulos em português brasileiro. *Cuadernos de la ALFAL*, n. 12, p. 387–410, nov. 2020.

DAKS, Alon; CLARK, Aidan. Unsupervised authorial clustering based on syntactic structure. In: *54th Annual Meeting of the Association for Computational Linguistics*. Berlin: Association for Computational Linguistics, 2016. p. 114–118.

DAROONEH, Amir H; SHARIATI, Ashrafalsadat. Metrics for evaluation of the author's writing styles: Who is the best? *Chaos*, v. 24, n. 33132, 2014.

DEVI, Khongbantabam Susila; RAVI, R. A mining algorithm to generate the candidate pattern for authorship attribution for filtering spam mail. *International Journal of Computer Science and Information Technologies*, v. 6, n. 2, p. 1917–1921, 2015.

DING, Steven H. H.; FUNG, Benjamin C. M.; IQBAL, Farkhund; CHEUNG, William K. Learning stylometric representations for authorship analysis. *IEEE Transactions on Cybernetics*, v. 49, n. 1, p. 107–121, 2016.

DOREIAN, P. A measure of standing of journals in stratified networks. *Journal of the American Society for Information Science*, v. 8, n. 5-6, p. 341–363, 1985.

DUQUE, A. B.; CARVALHO, F. D. A. T. D.; VIMIEIRO, R. A multiview clustering approach for mining authorial affinities in literary texts. In: *VII Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.]: IEEE, 2019. p. 818–823.

EDER, Maciej; RYBICKI, Jan; KESTEMONT, Mike. Stylometry with r: A package for computational text analysis. *The R Journal*, v. 8, 2016.

EGLIN, Veronique; BRES, Stephane; RIVERO, Carlos. Hermite and gabor transforms for noise reduction and handwriting classification in ancient manuscripts. *IJDAR*, Springer, n. 9, p. 101–122, 2007.

EL-FIQI, Heba; PETRAKI, Eleni; ABBASS, Hussein A. Network motifs for translator stylometry identification. *PLOS ONE*, p. 1–33, 2019.

ELEWA, Abdelhamid. Authorship verification of disputed hadiths in sahih al-bukhari and muslim. *Digital Scholarship in the Humanities*, Oxford University Press, v. 34, n. 2, p. 261–276, 2019.

ERDOS, P.; RENYI, A. On random graphs. *Publicationes Mathematicae*, v. 6, p. 290–297, 1959.

FADIGAS, Inácio de Sousa; HENRIQUE, Trazíbulo; SENNA, Valter de; MORET, Marcelo A.; PEREIRA, Hernane Borges de Barros. Análise de redes semânticas baseada em títulos de artigos de periódicos científicos: o caso dos periódicos de divulgação em educação matemática. *Educ. Mat. Pesqui.*, v. 11, n. 1, p. 167–193, 2009.

FIORIN, José Luiz. O acordo ortográfico - uma questão de política linguística. *Veredas - Atemática*, v. 1, n. 9, p. 07–19, 2009.

FOURNIER-VIGER, Philippe; LIN, Jerry Chun-Wei; VO, Bay; CHI, Tin Truong; ZHANG, Ji; LE, Hoai Bac. A survey of itemset mining. *WIREs Data Mining Knowl Discov*, John Wiley and Sons, v. 7, p. 1–18, 2017.

FRANCHI, Carlos; NEGRAO, Esmeralda Vailati; MULLER, Ana Lucia. Um exemplo de análise e argumentaćão em sintaxe. *Revista da ANPOLL*, n. 5, p. 37–63, jul./dez. 1998.

FREEMAN, L.C. Centrality in social networks conceptual clarification. social networks. *Elsevier*, v. 1, n. 3, p. 215–239, 1978.

G1. *É FAKE que poema sobre quarentena foi escrito de forma profética por romancista do século 19. Editora Globo - Extra.* 2020. Disponível em:https://extra.globo.com/fato-ou-fake/e-fake-que-poema-sobre-quarentena-foi-escrito-de-forma-profetica-por-romancista-do-seculo-19-24404489.html/. Acessado em: 03/05/2020.

GALINA, R. L.; FLORES, D. D. N. R.; KOMATI, K. S. Comparison of stylometric attributes for writing authorship identification: A case study of guimaraes rosa versus clarice lispector. In: *XVI Encontro Nacional de Inteligência Artificial e Computacional*. Salvador, BA, Brazil: ENIAC 2019, 2019. p. 353–364.

GOMEZ-ADORNO, Helena; SIDOROV, Grigori; PINTO, David; VILARINO, Darnes; GELBUKH, Alexander. Automatic authorship detection using textual patterns extracted from integrated syntactic graphs. *Sensors*, MPDI, v. 16, n. 1374, p. 1–9, 2016.

GRANT, T.; BAKER, K. Identifying reliable, valid markers of authorship: a response to chaski. *Forensic linguistics*, v. 8, n. 1, p. 66–79, 2001.

GRAY, A.; SALLIS, P.; MACDONELL, S. Software forensics: Extending authorship analysis techniques to computer programs. In: *3rd Biannual Conference of the International Association of Forensic Linguists (IAFL)*. Durham, NC, USA: International Association of Forensic Linguists (IAFL), 1997. p. 1–8.

GRIEVE, J. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, v. 22, n. 2, p. 251, 2007.

HAHSLER, Michael; GRUEN, Bettina; HORNIK, Kurt. arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, v. 14, n. 15, p. 1–25, October 2005. ISSN 1548-7660.

HALVANI, Oren; WINTER, Christian; PFLUG, Anika. Authorship verification for different languages, genres and topics. *Digital Investigation*, v. 16, p. s33–s34, 2016.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2017.

HERNANDEZ-GONZALEZ, Salvador; ESTRADA-OMANA, Alejandro; FLORES-DE-LA-MOTA, Idalia; JIMENEZ-GARCIA, Jose Alfredo; FIGUEROA-FERNANDEZ, Vicente. Analisis de los comprobantes de compra de un minorista aplicando redes complejas. *Ingenieria Industrial*, Universidad del Bio-Bio, v. 18, n. 1, p. 81–98, 2019.

HIRST, G.; FEIGUINA, O. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, v. 22, n. 4, p. 405–417, 2007.

HOLLINGSWORTH, C. Using dependency-based annotations for authorship identification. *Text, Speech and Dialogue*, v. 7499, 2012.

HOLMES, D. I. The analysis of literary style - a review. *Journal of the Royal Statistical Society*, v. 148, n. 4, p. 328–341, 1985.

HONORIO, Tatiane Cruz de Souza; NETO, Francisco Dantas Nobre; ALMEIDA, Thallys Pereira de; DUARTE, Rodrigo Cartaxo Marques; BARBOSA, Yuri de Almeida Malheiros; ROCHA, Vinicius de Melo. Atribuição de autoria com weka. In: *IX Encontro de Extensão e X Encontro de Iniciação*. Joao Pessoa: Editora Universitária/UFPB, 2007. v. 1, p. 42.

HOU, Renkui; HUANG, Chu-Ren. Robust stylometric analysis and author attribution based on tones and rimes. *Natural Language Engineering*, v. 26, p. 49–71, 2020.

JALILI, Mahdi. *centiserve: Find Graph Centrality Indices*. [S.l.], 2017. R package version 1.0.0. Disponível em: <https://CRAN.R-project.org/package=centiserve>.

JAMIL, M. T.; MUSTAFA, T. K. Ranking attribution: A novel method for stylometric authorship identification. In: *International Journal of Advanced Computer Science and Applications*. [S.l.]: IJACSA, 2018. v. 9, p. 54–61.

JAZILAH, N. I. Close and open task authorship attribution: a computational authorship analysis. *PARADIGM: Journal of Language and Literary Studies*, v. 2, n. 1, 2019.

JUOLA, P.; VESCOVI, D. Analyzing stylometric approaches to author obfuscation. In: *Advances in Digital Forensics VII*. Berlin: Springer, 2011. v. 361, p. 115–125.

KARIMI-MAJD, Amir-Mohsen; MAHOOTCHI, Masoud. A new data mining methodology for generating new service ideas. *Information Systems and e-Business Management*, Springer, v. 13, n. 3, p. 421–443, 2014.

KERNOT, David; BOSSOMAIER, Terry; BRADBURY, Roger. Using shakespeares sotto voce to determine true identity from text. *Frontiers in Psychology*, v. 9, p. 1–17, 2018.

KOPPEL, Moshe; SCHLER, Jonathan; ARGAMON, Shlomo. Computational methods in authorship attribution. *Journal of the americam society for information science and technology*, v. 60, n. 1, p. 9–26, 2009.

KUZU, R. S.; BALCI, Koray; SALAH, Albert Ali. Authorship recognition in a multiparty chat scenario. In: *4th International Conference on Biometrics and Forensics*. Limassol, Cyprus: IEEE 2016, 2016. p. 1–6.

KUZU, R. S.; SALAH, A. A. Chat biometrics. *IET Research Journals, The Institution of Engineering and Technology*, p. 1–12, 2015.

LAGUTINA, Ksenia; LAGUTINA, Nadezhda; BOYCHUK, Elena; VORONTSOVA, Inna; SHLIAKHTINA, Elena; BELYAEVA, Olga; PARAMONOV, Ilya; DEMIDOV, P.G. A survey on stylometric text features. In: *25th Conference of Open Innovations Association (FRUCT)*. Helsinki, Finland: IEEE Xplore, 2019. p. 184–195.

LAHIRI, S.; MIHALCEA, R. Authorship attribution using word network features. *ArXiv*, abs/1311.2978, 2013.

LE, H.; SAFAVI-NAINI, R. On de-anonymization of single tweet messages. In: *Fourth ACM International Workshop on Security and Privacy Analytics*. Tempe, AZ, USA: IWSPA 18, 2018. p. 8–14.

LIAW, A.; WIENER, M. Classification and regression by randomforest. *R News*, v. 2, n. 3, p. 18–22, 2002.

LIMA-NETO, José Lamartine de Andrade; CUNHA, Marcelo do Vale; PEREIRA, Hernane Borges de Barros. Redes semânticas de discursos orais de membros de grupos de ajuda mútua. *OBRA DIGITAL*, n. 14, p. 50–66, 2018.

LIU, H.; XU, C. Can syntactic networks indicate morphological complexity of a language? In: *Europhysics Letters Association EPL*. [S.l.]: Europhysics Letters, 2011. v. 93.

LIU, Yi-Hung; HO, Yen-Liang Chen aand Wu-Liang. Predicting associated statutes for legal problems. *Information Processing and Management*, n. 51, p. 194–211, 2015.

LUYCKX, K. *Scalability Issues in Authorship Attribution*. Tese (Doutor em Linguística) — Departement Taalkunde, Faculteit Letteren en Wijsbegeerte, Universiteit Antwerpen, Antwerpen, 2010.

LUYCKX, K.; DAELEMANS, W. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, v. 26, n. 1, 2011.

MACHICAO, Jeaneth; CORRêA-JR, Edilson A.; MIRANDA, Gisele H. B.; AMANCIO, Diego R.; BRUNO, Odemir M. Authorship attribution based on life-like network automata. *PLOS ONE*, p. 1–21, 2018.

MARGARIDO, Paulo R. A.; NUNES, Maria das Graças V.; PARDO, Thiago A. S.; OLIVEIRA-JR, Osvaldo N. de. Estabilização de métricas de redes complexas. In: *Anais do XXVIII Congresso da SBC*. Belém do Pará: SBC, 2008. p. 61–70.

MARINHO, Vanessa Queiroz. *Desenvolvimento de novos modelos para reconhecimento de autoria com a utilização de redes complexas*. Dissertação (Mestra em Ciências) — Instituto de Ciências Matemáticas e de Computação (ICMC-USP), São Carlos, 2017.

MARINHO, V. Q.; HIRST, G.; AMANCIO, D. R. Authorship attribution via network motifs identification. In: *5th Brazilian Conference on Intelligent Systems*. Recife: BRACIS, 2016. p. 355–360.

MARINHO, Vanessa Queiroz; HIRST, Graeme; AMANCIO, Diego Raphael. Labelled network subgraphs reveal stylistic subtleties in written texts. *Journal of Complex Networks*, Oxford University Press, n. 6, p. 620–638, 2018.

MARKOV, I.; BAPTISTA, J.; PICHARDO-LAGUNAS, O. Authorship attribution in portuguese using character n-grams. *Acta Polytechnica Hungarica*, v. 14, n. 3, p. 59–78, 2017.

MARTINS, Ricardo; ALMEIDA, José João; HENRIQUES, Pedro; NOVAIS, Paulo. A sentiment analysis approach to increase authorship identification. *Expert Systems*, p. 1–12, 2019.

MASUCCI, A.; RODGERS, G. Network properties of written human language. *Physical review. E, Statistical, nonlinear, and soft matter physics*, v. 74, 2006.

MATEUS, M. H. M.; BRITO, A.M.; DUARTE, I.; FARIA, I.H.; FROTA, S.; MATOS, G.; VILLALVA, A. *Gramática da língua portuguesa*. Lisboa: Caminho, 2003.

MEHRI, A.; DAROONEH, A. H.; SHARIATI, A. The complex networks approach for authorship attribution of books. *Physica A*, n. 391, p. 2429–2437, 2012.

MENDENHALL, T. C. The characteristic curves of composition. *Science*, n. 9, p. 237–249, 1887.

MENDONCA, M. *Assinaturas Conflitantes. Revista Epoca.* 2002. Disponível em:http://revistaepoca.globo.com/Epoca/0,6993,EPT435457-1664-1,00.html/. Acessado em: 04 nov 2018.

MEYER, David; DIMITRIADOU, Evgenia; HORNIK, Kurt; WEINGESSEL, Andreas; LEISCH, Friedrich; CHANG, Chih-Chung; LIN, Chih-Chen. *e1071: Misc Functions of the Department of Statistics, Probability. R package version 1.7-3.* 2019. Disponível em: https://CRAN.R-project.org/package=e1071. Acessado em: 12 mai 2020.

MIOTO, Carlos; SILVA, Maria Cristina Figueiredo; LOPES, Ruth Elisabeth Vasconcellos. *Novo manual de sintaxe.* [S.l.]: Insular, 2007.

MOSTELLER, F.; WALLACE, D. L. Inference and disputed authorship: The federalist. *Reading, MA: Addison-Wesley*, 1964.

MULLER, Ana; OLIVEIRA, Fátima. Bare nominals and number in brazilian and european portuguese. *Journal of Portuguese Linguistics*, v. 3, n. 1, p. 9–36, 2004.

NEVES, Maria Helena de Moura. Como as palavras se organizam em classes. *Portal da Língua Portuguesa*, 2006.

NEWMAN, M. E. J. *Networks: An Introduction.* New York: Oxford University Press, 2010.

NIRKHI, S.; DHARASKAR, R. V.; THAKARE, V. M. Authorship verifIcation of online messages for forensic investigation. *Procedia Computer Science*, v. 78, p. 640–645, 2016.

OLIVEIRA, A. L. C. *Análise de autoria em contexto forense: um estudo de caso.* Dissertação (Mestre em Linguística) — Linguística, Porto, 2019.

OLIVEIRA-JUNIOR, W. R. D. *Atribuição de autoria de documentos em língua portuguesa utilizando a distância normalizada de compressão.* Dissertação (Mestre em Informática Aplicada) — Programa de PósGraduação em Informática Aplicada, Curitiba, 2011.

PAVELEC, D.; OLIVEIRA, L. S.; JUSTINO, E.; NETO, F. D. Nobre; BATISTA, L. V. Compression and stylometry for author identification. In: *Proceedings of International Joint Conference on Neural Networks.* Atlanta (USA): IEEE, 2009. p. 2445–2450.

QI, Xingqin; FULLER, Eddie; WU, Qin; WU, Yezhou; ZHANG, Cun-Quan. Laplacian centrality: A new centrality measure for weighted networks. *Information Sciences*, Elsevier, n. 194, p. 240–253, 2012.

QUINTANILLA, Pamela Revuelta. *Comparing vector document representation methods for authorship identification.* Dissertação (Mestre em Ciências) — Institulo de Matemática e Estatística da Universidade de São Paulo, São Paulo, 2020.

RAMEZANI, Reza; SHEYDAEI, Navid; KAHANI, Mohsen. Evaluating the effects of textual features on authorship attribution accuracy. In: *3rd International Conference on Computer and Knowledge Engineering.* Ferdowsi: IEEE, 2013.

REXHA, Andi; KROLL, Mark; ZIAK, Hermann; KERN, Roman. Authorship identification of documents with high content similarity. *Scientometrics*, Springer, n. 115, p. 223–237, 2007.

ROCHA, A.; SCHEIRER, W. J.; FORSTALL, C. W.; CAVALCANTE, T.; THEOPHILO, A.; SHEN, B.; CARVALHO, A. R. B.; STAMATATOS, E. Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, v. 12, n. 1, p. 5–33, 2017.

ROCHA, M. A. D. *Um texto tão singular quanto a impressão digital: O uso de sistemas inteligentes para reconhecimento de autoria*. Dissertação (Mestre em Ciências) — Pós-Graduação em Engenharia Elétrica e de Computação da UFRN, Universidade Federal Do Rio Grande Do Norte, Natal, 2019.

RODRIGUES, Melissa; GAMA, João; FERREIRA, Carlos Abreu. Identifying relationships in transactional data. In: *IBERAMIA 2012, LNAI 7637*. [S.l.]: Springer, 2012. p. 81–90.

ROZZ, Y.; MENEZES, R. Author attribution using network motifs. In: *Complex Networks IX*. Boston: Springer, 2018. v. 9, p. 199–207.

RUDMAN, J. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, v. 31, n. 4, p. 351–365, 1998.

SANTOS, Bruno R. M.; CARVALHO, Deborah Ribeiro. Visualization of association rules. *Iberoamerican Journal of Applied Computing*, v. 7, n. 1, p. 1–13, 2017.

SAVOY, J. Is starnone really the author behind ferrante? *Digital Scholarship in the Humanities*, v. 33, n. 4, 2018.

SCHOMAKER, Lambert; FRANKE, Katrin; BULACU, Marius. Using codebooks of fragmented connected-component contours in forensic and historic writer identification. *Pattern Recognition Letters*, Elsevier, n. 28, p. 719–727, 2007.

SEGARRA, S.; EISEN, M.; RIBEIRO, A. Authorship attribution using function words adjacency networks. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver: IEEE, 2013. p. 5563–5567.

SEGARRA, S.; EISEN, M.; RIBEIRO, A. Authorship attribution through function word adjacency networks. *arXiv.org*, v. 63, n. 20, p. 5464–5478, 2014.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. *Introdução à mineração de dados - com aplicações em R*. Rio de Janeiro: Elsevier Brasil, 2016.

SILVA, Rosa V. Mattos e. Diversidade e unidade - a aventura linguística do português. *ICALP*, v. 11, p. 60–72, mar. 1988.

SINA, Sigal; ROSENFELD, Avi; KRAUS, Sarit; AKIVA, Navot. A hybrid approach of classifier and clustering for solving the missing node problem. *AAAI*, 2015.

SISOVIC, S.; MARTINCIC-IPSIC, S.; MESTROVIS, A. Comparison of the language networks from literature and blogs. In: *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. Opatija, Croatia: IEEE, 2014. p. 1603–1608.

SOLE, Ricard V.; MURTRA, Bernat Corominas; VALVERDE, Sergi; STEELS, Luc. Language networks: their structure, function and evolution. *Complexity*, v. 15, p. 20–26, 2010.

SOLOMONOFF, R.; RAPOPORT, A. Connectivity of random nets. *Bulletín of Mathematical Biophysics*, v. 13, 1951.

SOLORIO, T.; HASAN, R.; MIZAN, M. Sockpuppet detection in wikipedia: A *Corpus* of real-world deceptive writing for linking identities. In: *Ninth International Conference on Language Resources and Evaluation*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. p. 1355–1358.

SOUSA-SILVA, Rui; SARMENTO, Luís; GRANT, Tim; OLIVEIRA, Eugénio; MAIA, Belinda. Comparing sentence-level features for authorship analysis in portuguese. In: *PROPOR 2010*. Porto Alegre, RS, Brazil: LNAI 6001, 2010. p. 51–54.

STAMATATOS, E. A survey of modern authorship attribution methods. *Jornal of the american society for information science and technology*, v. 60, n. 3, p. 538–556, 2009.

STANISZ, T.; KWAPIEN, J.; DROZDZ, S. Linguistic data mining with complex networks: A stylometric-oriented approach. *Information Sciences*, n. 482, p. 301–320, 2019.

STEYVERSA, M.; TENENBAUM, J. B. The large-scale structure of semantic networks statistical analyses and a model of semantic growth. *Cognitive Science, v. 29, p. 41-78, 2005.*, v. 29, p. 41–78, 2005.

TAMBOLI, M. S.; PRASAD, R. S. Authorship analysis and identification techniques: A review. *International Journal of Computer Applications*, v. 77, p. 11–15, 2013.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. *Introduction to data mining*. Boston: PearsonEducation, 2006.

TEAM, R. C. R: A language and environment for statistical computing. *R Foundation for Statistical Computing. Vienna, Austria*, 2019.

TEIXEIRA, Gesiane Miranda; AGUIAR, Madaya dos Santos Figueredo de; CARVALHO, Chrissie Ferreira de; DANTAS, Douglas Ramos; CUNHA, Marcelo do Vale; MORAIS, José Henrique Miranda de; PEREIRA, Hernane Barros de Borges; MIRANDA, José Garcia Vivas. Complex semantic networks. *International Journal of Modern - Physics C*, v. 21, p. 333–347, 2010.

THISTED, R.; EFRON, B. Did shakespeare write a newly-discovered poem? *Biometrika*, v. 74, n. 3, p. 445, 1987.

TRAVERS, J.; MILGRAM, S. An experimental study of the small world problem. *Sociometry*, v. 32, n. 4, p. 425–443, 1969.

VALENCIA, C. A. *Propriedades de redes aplicadas à atribuição de autoria*. Tese (Doutor em Ciências) — Programa de Pós-Graduação em Física do Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2017.

VARELA, P. J. *O uso de atributos estilométricos na identificação da autoria de textos*. Dissertação (Mestre em Informática) — Pós Graduação em Informática Aplicada, Curitiba, 2010.

VARELA, P. J. *Uma abordagem computacional Baseada Em Análise Sintática Multilíngue Na Atribuição Da Autoria De Documentos Digitais*. Tese (Doutor Em Informática Aplicada) — Programa De Pós-Graduação em Informática Aplicada, Pontifícia Universidade Católica Do Paraná, Curitiba, 2017.

VARELA, Paulo Junior; ALBONICO, Michel; ASSIS, João Lucas Varela de. Authorship attribution based on syntactic attributes of the portuguese language. In: *XLV - The Latin American Information Technology Conference*. Cidade do Panamá: [s.n.], 2019. p. 1–10.

VAZIRIAN, S.; ZAHEDI, M. Path-based and whole-network. In: *ALHAJJ, Reda; ROKNE, Jon (org.). Encyclopedia of Social Network Analysis and Mining*. New York: Springer, 2014. p. 1256–1268.

VAZIRIAN, S.; ZAHEDI, M. A modified language modeling method for authorship attribution. In: *2016 Eighth International Conference on Information and Knowledge Technology (IKT)*. Hamedan, Iran: IEEE, 2016. p. 32–37.

VEL, O. D. Mining e-mail authorship. paper presented at the workshop on text mining. In: *ACM International Conference on Knowledge*. [S.l.: s.n.], 2000.

VENCKAUSKAS, Algimantas; DAMASEVICIUS, Robertas; MARCINKEVICIUS, Romas; KARPAVICIUS, Arnas. Problems of authorship identification of the national language electronic discourse. In: *Information and Software Technologies: 21st International Conference*. Druskininkai, Lithuania: Springer, 2015. v. 538.

VERDU, Daniel. *O mistério de Elena Ferrante continua fascinando a Itália*. 2019. Disponível em: https://brasil.elpais.com/brasil/2019/11/06/cultura/1573046745_374458.html/. Acessado em: 23 abr 2020.

VILLAR-RODRIGUEZ, Esther; SER, Javier Del; BILBAO, Miren Nekane; SALCEDO-SANZ, Sancho. A feature selection method for author identification in interactive communications based on supervised learning and language typicality. *Engineering Applications of Artificial Intelligence*, n. 56, p. 175–184, 2016.

VOROBEVA, A. A. Forensic linguistics: automatic web author identification. *Scientific and technical bulletin of information technologies, mechanics and optics*, n. 2, p. 295–302, 2016.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of small-world networks. *Nature*, v. 393, p. 440–442, 1998.

WHITE, H. C.; BOORMAN, S. A.; BREIGER, R. L. Social structure from multiple nerworks. i. blockmodels of roles and positions. *American Journal of Sociology*, v. 81, n. 4, p. 730–780, 1976.

WILLIAMS, G. J. *Data Mining with Rattle and R: The art of excavating data for knowledge discovery*. [S.l.]: Springer, 2011.

WISSE, W.; VEENMAN, C. Scripting dna: Identifying the javascript programmer. *Digital Investigation*, n. 15, p. 61–71, 2015.

WOLFRAM, Stephen. Universality and complexity in cellular automata. *Physica D: Nonlinear Phenomena*, ScienceDirect, v. 10, n. 1, p. 1–35, 1984.

WU, Jiacong; WANG, Yu; SHAFIEE, Sara; ZHANG, Dongsong. Discovery of associated consumer demands: Construction of a co-demanded product network with community detection. *Expert Systems With Applications*, Elsevier, n. 178, p. 1–20, 2021.

YULE, G. U. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, v. 30, n. 3-4, p. 363–390, 1939.

ZAKI, Mohammed J. Spade: An efficient algorithm for mining frequent sequences. *Machine Learning*, v. 42, p. 31–60, 2001.

ZHANG, Chunxia; WU, Xindong; NIU, Zhendong; DING, Wei. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, v. 57, n. 3, p. 378–393, 2006.

ZHANG, Chunxia; WU, Xindong; NIU, Zhendong; DING, Wei. Authorship identification from unstructured texts. *Knowledge-Based Systems*, v. 66, p. 99–111, 2014.

ZHANG, Y.; LIU, Y.; CHEN, G. A solution of anonymous email identification based on writing structural pattern. In: *12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. Zhangjiajie, China: IEEE, 2015. p. 1525–1531.

Apêndice A

Obras que compõem o corpus da pesquisa

Tabela com os detalhes das obras que compõem o corpus da pesquisa

Autor	País de origem	Título	1ª publicação	Ano de edição	Fonte do Texto
Aluísio Azevedo	Brasil	O Mulato	1881	1881	BLPL
Aluísio Azevedo	Brasil	A Condessa Vésper	1882	1882	BLPL
Aluísio Azevedo	Brasil	O Coruja	1889	1889	BLPL
Aluísio Azevedo	Brasil	O Cortiço	1890	1890	BLPL
Ana Luísa de Azevedo Castro	Brasil	D. Narcisa De Villar: legenda do tempo colonial	1859	1859	BLPL
Antônio Joaquim da Rosa	Brasil	A Cruz de Cedro	1854	1924	BLPL
Antônio Joaquim da Rosa	Brasil	A Feiticeira	1854	1919	BLPL
Barão de Teffé	Brasil	A corveta Diana	1873	1873	BLPL
Bernardo Guimarães	Brasil	O índio Afonso	1872	1873	BLPL
Bernardo Guimarães	Brasil	Maurício ou Os Paulistas em São João del-Rei	1877	1877	BLPL
Bernardo Guimarães	Brasil	A Ilha Maldita	1879	1879	BLPL
Bernardo Guimarães	Brasil	Rosaura, a enjeitada	1882	1914	BLPL
Camilo Castelo Branco	Portugal	A Filha do Arcediago	1854	1868	Gutenberg
Camilo Castelo Branco	Portugal	A Neta do Arcediago	1854	1860	Gutenberg
Camilo Castelo Branco	Portugal	Lágrimas Abençoadas	1857	1906	Gutenberg
Camilo Castelo Branco	Portugal	Cenas da Foz	1857	1860	Gutenberg
Camilo Castelo Branco	Portugal	Carlota Ângela	1858	1874	Gutenberg
Camilo Castelo Branco	Portugal	O Que Fazem Mulheres	1858	1907	Gutenberg
Camilo Castelo Branco	Portugal	Estrelas Funestas	1862	1906	Gutenberg
Camilo Castelo Branco	Portugal	Agulha em Palheiro	1863	1904	Gutenberg
Camilo Castelo Branco	Portugal	Estrelas Propícias	1863	1863	Gutenberg
Camilo Castelo Branco	Portugal	O Olho de Vidro	1866	1918	Gutenberg
Camilo Castelo Branco	Portugal	Livro de Consolação	1872	1872	Gutenberg
Dionísia Gonçalves Pinto	Brasil	Dedicação de uma amiga	1850	1850	BLPL
Eça de Queiroz	Portugal	O Crime do Padre Amaro	1875	1998	Portal Domínio Público
Eça de Queiroz	Portugal	A relíquia	1887	1997	Portal Domínio Público
Eça de Queiroz	Portugal	Os Maias	1888	1888	Portal Domínio Público
Eça de Queiroz	Portugal	Alves E Cia	1925	1997	Portal Domínio Público
Euclides da Cunha	Brasil	Os Sertões	1902	1984	Portal Domínio Público
Franklin Távora	Brasil	O Cabeleira	1876	1973	Portal Domínio Público
Franklin Távora	Brasil	O Matuto	1878	1902	Portal Domínio Público
Franklin Távora	Brasil	O Sacrifício	1879	1969	Portal Domínio Público
Germano Hasslocher	Brasil	A espelunca: Romance de atualidade	1889	1889	BLPL
João Carlos de Medeiros Pardal Mallet	Brasil	Hóspede	1887	2008	BLPL
João Carlos de Medeiros Pardal Mallet	Brasil	Hóspede	1887	1887	BLPL
João Carlos de Medeiros Pardal Mallet	Brasil	O lar	1888	2008	BLPL
João Carlos de Medeiros Pardal Mallet	Brasil	O esqueleto	1890	1944	BLPL
Joaquim Felício dos Santos	Brasil	Acaiaca	1894	1866	BLPL
Joaquim Manuel de Macedo	Brasil	O Culto do Dever	1865	1865	BLPL
Joaquim Manuel de Macedo	Brasil	A luneta mágica	1869	1990	Portal Domínio Público
Joaquim Manuel de Macedo	Brasil	As vítimas algozes	1869	1869	BLPL
Joaquim Manuel de Macedo	Brasil	O Rio do Quarto	1869	1869	BLPL
Joaquim Manuel de Macedo	Brasil	As Mulheres de Mantilha	1870	1870	BLPL
Joaquim Manuel de Macedo	Brasil	A Namoradeira	1870	1870	BLPL
Joaquim Manuel de Macedo	Brasil	Os quatro pontos cardeais	1872	1872	BLPL
Joaquim Manuel de Macedo	Brasil	A misteriosa	1872	1872	BLPL
José de Alencar	Brasil	O Guarani	1857	1996	Portal Domínio Público
José de Alencar	Brasil	Lucíola	1862	1988	Portal Domínio Público
José de Alencar	Brasil	Iracema	1865	1991	Portal Domínio Público
José de Alencar	Brasil	O Gaúcho	1870	1998	Portal Domínio Público
José de Alencar	Brasil	A pata da gazela	1870	1998	Portal Domínio Público
José de Alencar	Brasil	Sonhos d oro	1872	1998	Portal Domínio Público
Júlia Lopes de Almeida	Brasil	A falência	1901	1901	BLPL
Júlia Lopes de Almeida	Brasil	A Intrusa	1905	1905	BLPL
Júlia Lopes de Almeida	Brasil	Cruel amor	1908	1928	BLPL
Júlia Lopes de Almeida	Brasil	A Silveirinha	1912	1914	BLPL
Júlio Dinis	Portugal	As Pupilas do Senhor Reitor	1867	1987	Portal Domínio Público
Júlio Ribeiro	Brasil	A Carne	1888	1999	Portal Domínio Público
Lindolfo Rocha	Brasil	Maria Dusá	1910	1980	Portal Domínio Público
Machado de Assis	Brasil	Ressurreição	1872	1994	Portal Domínio Público
Machado de Assis	Brasil	A mão e a luva	1874	1874	BLPL
Machado de Assis	Brasil	Helena	1876	1876	BLPL
Machado de Assis	Brasil	Memórias Póstumas de Brás Cubas	1881	1994	Portal Domínio Público
Machado de Assis	Brasil	Memórias Póstumas de Brás Cubas	1881	1881	BLPL
Machado de Assis	Brasil	Quincas Borba	1891	1994	Portal Domínio Público
Machado de Assis	Brasil	Quincas Borba	1891	1891	BLPL
Machado de Assis	Brasil	Memorial de Aires	1908	1985	Portal Domínio Público
Machado de Assis	Brasil	Casa Velha	1944	1986	Portal Domínio Público

Autor	País de origem	Título	1ª publicação	Ano de edição	Fonte do Texto
Manuel Antonio de Almeida	Brasil	Memórias de um sargento de milícias	1852	1996	Portal Domínio Público
Manuel de Brito Camacho	Portugal	Ao de Leve	1903	1913	Gutenberg
Manuel de Oliveira Paiva	Brasil	Dona Guidinha do Poço	1981	1981	Portal Domínio Público
Mário de Sá-Carneiro	Portugal	A Confissão de Lúcio	1914	1991	Portal Domínio Público
Pinheiro Chagas	Portugal	Os Guerrilheiros da Morte	1872	1872	BLPL
Pinheiro Chagas	Portugal	O terramoto de Lisboa	1874	1874	BLPL
Raul Brandão	Portugal	Os Pobres	1906	2001	Portal Domínio Público
Raul Brandão	Portugal	Humus	1917	1921	Baixe Livros
Raul Brandão	Portugal	A Morte do Palhaço	1926	2016	Baixe Livros
Raul Pompeia	Brasil	As Jóias da Coroa	1883	1997	Portal Domínio Público
Raul Pompeia	Brasil	O Ateneu	1888	1996	Portal Domínio Público
Rilvan Batista de Santana	Brasil	O empresário	2007	na	Portal Domínio Público
Rilvan Batista de Santana	Brasil	Maria Madalena	2008	na	Portal Domínio Público
Rilvan Batista de Santana	Brasil	O DNA de Emanuel	2008	na	Portal Domínio Público
Salomão Rovedo	Brasil	Gardenia	2006	2006	Portal Domínio Público
Teixeira e Sousa	Brasil	O filho do pescador	1843	1843	BLPL
Teixeira e Sousa	Brasil	Gonzaga ou a conjuração do Tiradentes	1848	1848	BLPL
Visconde de Taunay	Brasil	A mocidade de Trajano	1871	1871	BLPL
Visconde de Taunay	Brasil	Inocência	1872	1991	Portal Domínio Público
Visconde de Taunay	Brasil	No declínio	1898	1898	BLPL
Visconti Coaracy	Brasil	Amor que mata	1873	1873	BLPL

Apêndice B

Lista de *stopwords*

A lista de *stopwords* abaixo contém artigos, pronomes, preposições e numerais.

stopwords = {a, à, algo, alguém, algum, alguma, algumas, alguns, ante, ao, aonde, aos, após, aquela, àquela, aquelas, àquelas, aquele, àquele, aqueles, àqueles, àquilo, as, às, assim, até, avos, bilhão, bilhões, bilionésimo, cada, caso, catorze, cem, centésimo, cêntuplo, certa, certas, certo, certos, cinco, cinquenta, co, com, comigo, como, conforme, conosco, conquanto, consigo, consoante, contigo, contra, contudo, convosco, cuja, cujas, cujo, cujos, da, dacolá, daí, dalém, dalgo, dalguém, dalgum, dalguma, dalgumas, dalguns, dalgures, dalhures, dali, dantes, daquela, daquelas, daquele, daqueles, daquém, daqui, daquilo, das, de, décimo, décuplo, dela, delas, dele, deles, dentre, desde, dessa, dessas, desse, desses, desta, destas, deste, destes, dez, dezenove, dezesseis, dezessete, dezoito, disso, disto, do, dobro, dois, donde, dos, doutra, doutras, doutrem, doutro, doutrora, doutros, doze, ducentésimo, dum, duma, dumas, duns, duodécuplo, duplo, duzentos, e, ela, elas, ele, eles, em, embora, então, entre, entretanto, essa, essas, esse, esses, esta, estas, este, estes, eu, exceto, fora, la, las, lhe, lhes, lo, logo, los, mas, me, mediante, meio, meu, meus, mil, milésimo, milhão, milhões, milionésimo, mim, minha, minhas, muita, muitas, muito, muitos, na, nada, naquela, naquelas, naquele, naqueles, naquilo, nas, nela, nelas, nele, neles, nem, nenhum, nenhuma, nenhumas, nenhuns, nessa, nessas, nesse, nesses, nesta, nestas, neste, nestes, ninguém, nisso, nisto, no, nonagésimo, nongentésimo, noningentésimo, nono, nônuplo, nos, nós, nossa, nossas, nosso, nossos, nove, novecentos, noventa, o, octingentésimo, octogésimo, óctuplo, oitavo, oitenta, oito, oitocentos, onde, onze, os, outra, outras, outrem, outro, outros, para, pela, pelas, pelo, pelos, perante, pois, por, porém, porquanto, porque, portanto, pouca, poucas, pouco, poucos, pra, pras, primeiro, pro, pros, prum, pruma, prumas, pruns, quadragésimo, quadringentésimo, quádruplo, quais, quaisquer, qual, qualquer, quando, quanta, quantas, quanto, quantos, quarenta, quarto, quatro, quatrocentos, que, quem, quingentésimo, quinhentos, quinquagésimo, quinto, quíntuplo, quinze, salvo, se, segundo, seis, seiscentos, sem, senão, septingentésimo, septuagésimo, sessenta, sete, setecentos, setenta, sétimo, sétuplo, seu, seus, sexagésimo, sexcentésimo, sexto, sêxtuplo, si, sob, sobre, sua, suas, tanta, tantas, tanto, tantos, te, terceiro, terço, teu, teus, ti, tirante, toda, todas, todavia, todo, todos, trás, trecentésimo, três, treze, trezentos, trigésimo, trinta, tríplice, triplo, tu, tua, tuas, tudo, um, uma, umas, undécuplo, uns, vária, várias, vário, vários, vigésimo, vinte, visto, você, vocês, vos, vós, vossa, vossas, vosso, vossos}

Apêndice C

Algoritmo: Extrair Corpus

Linhas	Códigos
1	
2	# Algoritmo: Extrair Corpus
3	#
4	# Este algoritmo extrai subconjuntos de textos
5	# baseados nos parâmetros:
6	# - Escolas literárias;
7	# - País de origem do autor;
8	# - Anos das edição;
9	# - Total de palavras;
10	#
11	# Observação:
12	# O corpus principal deve conter arquivos textos em formato UTF-8,
13	# A primeira linha de cada texto precisa conter os meta-dados
14	# das obras do seguinte formato:
15	# [nome do autor]; [título da obra]; [ano da 1a publicação]; [ano da edição]; [total de palavras]; [país de origem do autor];
16	#
17	# Exemplo: Franklin Távora; O CABELEIRA;1876;1973;5088;romantismo;Brasil;
18	#
19	# Atenção: O algoritmo não apaga os arquivos que existem previamente na pasta de destino.
20	
21	#extrair.corpus <- function(){
22	
23	#Limpar variáveis da memória
24	rm(list=ls(all=TRUE))
25	
26	library('beepr')
27	
28	#transforma um vetor de palavras em itens numerados;
29	msg <- function(vetor){
30	resp <- ''
31	for(i in 1:length(vetor)){
32	resp <- paste(resp,i,' - ',vetor[i],'; ', sep='')
33	}
34	resp <- paste(resp,i+1,' - todas;', sep='')
35	return(resp)
36	}
37	
38	
39	cat('\n\n##(início)\n\n')
40	cat('\n')
41	cat('**** Início da Execução ****')
42	
43	cat("\n\nInforme o diretorio onde estão os arquivos\n")
44	pasta.origem <- choose.dir(default = "C:\\Testes\\corpus\\", caption = "Selecione o diretório:")
45	dir.corpus <- dir(pasta.origem)
46	
47	autor <- vector()
48	titulo <- vector()
49	prim.edi <- vector()
50	ano.edi <- vector()
51	tot.pal <- vector()
52	escola <- vector()
53	pais <- vector()
54	
55	tab.meta.dados <- data.frame()
56	tmp <- 0
57	for(i in 1:length(dir.corpus)){
58	arquivo <- paste(pasta.origem, dir.corpus[i],sep='\\')
59	txt <- readLines(arquivo, encoding = "UTF-8")
60	meta.dados <- strsplit(tolower(txt[1]),';')
61	
62	autor <- meta.dados[[1]][1]
63	titulo <- meta.dados[[1]][2]

Linhas	Códigos
64	prim.edi <- meta.dados[[1]][3]
65	ano.edi <- meta.dados[[1]][4]
66	tot.pal <- as.numeric(meta.dados[[1]][5])
67	escola <- meta.dados[[1]][6]
68	pais <- meta.dados[[1]][7]
69	arquivo <- dir.corpus[i]
70	
71	linha <- cbind.data.frame(autor,titulo, prim.edi, ano.edi,escola,pais, arquivo)
72	if(tmp == 0){
73	tmp <- 1
74	tab.meta.dados <- linha
75	}else{
76	tab.meta.dados <- rbind.data.frame(tab.meta.dados, linha)
77	}
78	
79	}
80	
81	colnames(tab.meta.dados) <- c('autor','titulo','prim.edi','ano.edi','escola','pais', 'arquivo')
82	#View(tab.meta.dados)
83	
84	cat("Informe o diretorio de destino\n\n")
85	pasta <- choose.dir(default = "C:\\Testes\\destino\\", caption = "Selecione o diretório:")
86	dir.destino <- dir(pasta)
87	
88	#Filtro: Escola
89	escola <- unique(tab.meta.dados$escola)
90	cat('Selecione uma escola literária: \n')
91	cat(msg(escola))
92	resp <- as.numeric(readline(prompt='Escola: '))
93	
94	if(resp<0 | resp>(length(escola)+1)){
95	stop('Opção inválida!')
96	}
97	
98	if((length(escola)+1)==resp){
99	escola.filtro <- escola
100	} else {
101	escola.filtro <- escola[resp]
102	}
103	
104	#Filtro: País
105	pais <- unique(tab.meta.dados$pais)
106	mensagem <- paste0('Selecione o país de origem (',msg(pais),'): ')
107	resp <- as.numeric(readline(prompt=mensagem))
108	
109	if(resp<0 | resp>(length(pais)+1)){
110	stop('Opção inválida!')
111	}
112	
113	if((length(pais)+1)==resp){
114	pais.filtro <- pais
115	} else {
116	pais.filtro <- pais[resp]
117	}
118	
119	#Filtro: Ano de Edição
120	ano.edi <- unique(tab.meta.dados$ano.edi)
121	x <- grepl("^[0-9]{1,}$", ano.edi) #verificar se são números
122	ano.edi <- as.numeric(ano.edi[x])
123	cat('\tInforme a data de início para o ano de edição ou\n\ttecle Enter para ignorar o ano: ')
124	resp <- as.numeric(readline(prompt='Ano de início: '))
125	
126	if(is.na(resp)){

Linhas	Códigos
127	ano.ini <- as.numeric(min(ano.edi))
128	} else {
129	ano.ini <- resp
130	}
131	
132	cat('\tInforme a data de término para o ano de edição ou\n\ttecle Enter para ignorar o ano: ')
133	resp <- as.numeric(readline(prompt='Ano de Término: '))
134	
135	if(is.na(resp)){
136	ano.fim <- as.numeric(max(ano.edi))
137	} else {
138	ano.fim <- resp
139	}
140	
141	#Filtro: Total de palavras
142	tot.pal <- unique(tab.meta.dados$tot.pal)
143	cat('\tInforme o total de palavras ou\n\ttecle Enter para ignorá-lo: ')
144	resp <- as.numeric(readline(prompt='Total de palavras : '))
145	
146	if(is.na(resp)){
147	tot.pal <- as.numeric(max(tot.pal))
148	} else {
149	tot.pal <- resp
150	}
151	
152	id <- which(tab.meta.dados$escola%in%escola.filtro &
153	tab.meta.dados$pais%in%pais.filtro &
154	tab.meta.dados$ano.edi >= ano.ini &
155	tab.meta.dados$ano.edi <= ano.fim)
156	
157	if(identical(id,integer(0))){
158	stop('Não há corpus para os parâmetros informados.')
159	}
160	tab.corpus <- tab.meta.dados[id,]
161	View(tab.corpus)
162	
163	for(i in 1:nrow(tab.corpus)){
164	
165	arquivo <- paste(pasta.origem, tab.corpus$arquivo[i],sep='\\')
166	txt <- readLines(arquivo, encoding = "UTF-8")
167	
168	total.de.palavras <- 0
169	parar <- 0
170	texto <- ''
171	
172	for(j in 2:length(txt)){
173	anterior <- ' '
174	linha <- txt[j]
175	linha <- gsub('-',' ',linha)
176	if(!identical(linha,'')){
177	
178	for(k in 1:nchar(linha)){
179	letra <- substr(linha,k,k)
180	if(letra==' '){
181	if(anterior!=letra){
182	total.de.palavras <- total.de.palavras + 1
183	if(tot.pal <= total.de.palavras){
184	x <- k
185	break
186	}
187	}
188	}
189	anterior <- letra

Linhas	Códigos
190	}
191	
192	if(anterior!=' '){
193	total.de.palavras <- total.de.palavras + 1
194	}
195	#
196	if(tot.pal <= total.de.palavras){
197	texto <- c(texto, substr(linha,1,x))
198	texto <- c(texto, '.\n')
199	parar <- 1
200	break
201	}else{
202	texto <- c(texto, linha)
203	}
204	}
205	if(parar>0){
206	break
207	}
208	}
209	Encoding(texto) <- "UTF-8"
210	writeLines(texto, paste(pasta,'\\',tab.corpus$arquivo[i],sep=''),useBytes = TRUE)
211	}
212	
213	cat('\nInício do processamento: ', date(),'\n')
214	
215	cat('\nTérmino do processamento: ', date(),'\n\n')
216	cat('**** Fim da Execução ****')
217	cat('\n\n##(fim)\n\n')
218	
219	beep(sound = 5)
220	beep(sound = 1)
221	
222	rm(list=ls(all=TRUE))
223	
224	#}

Algoritmo: Etapa 01 - Transformar textos em conjuntos de itens

Linhas	Códigos
1	
2	#Limpar variáveis da memória
3	rm(list=ls(all=TRUE))
4	
5	#bibliotecas
6	library(utf8)
7	library('beepr')
8	
9	#Caracteres usados para identificar o fim de uma frase.
10	pontuacao_de_parada <- c(".","?","!")
11	
12	#Stopwords
13	stopwords <- c('a', 'à', 'algo', 'alguém', 'algum', 'alguma', 'algumas', 'alguns', 'ante', 'ao', 'aonde', 'aos',
14	'após', 'aquela', 'àquela', 'aquelas', 'àquelas', 'aquele', 'àquele', 'aqueles', 'àqueles', 'àquilo',
15	'as', 'às', 'assim', 'até', 'avos', 'bilhão', 'bilhões', 'bilionésimo', 'cada', 'caso', 'catorze', 'cem',
16	'centésimo', 'cêntuplo', 'certa', 'certas', 'certo', 'certos', 'cinco', 'cinquenta', 'co', 'com', 'comigo',
17	'como', 'conforme', 'conosco', 'conquanto', 'consigo', 'consoante', 'contigo', 'contra', 'contudo', 'convosco',
18	'cuja', 'cujas', 'cujo', 'cujos', 'da', 'dacolá', 'daí', 'dalém', 'dalgo', 'dalguém', 'dalgum', 'dalguma',
19	'dalgumas', 'dalguns', 'dalgures', 'dalhures', 'dali', 'dantes', 'daquela', 'daquelas', 'daquele', 'daqueles',
20	'daquém', 'daqui', 'daquilo', 'das', 'de', 'décimo', 'décuplo', 'dela', 'delas', 'dele', 'deles', 'dentre',
21	'desde', 'dessa', 'dessas', 'desse', 'desses', 'desta', 'destas', 'deste', 'destes', 'dez', 'dezenove',
22	'dezesseis', 'dezessete', 'dezoito', 'disso', 'disto', 'do', 'dobro', 'dois', 'donde', 'dos', 'doutra',
23	'doutras', 'doutrem', 'doutro', 'doutrora', 'doutros', 'doze', 'ducentésimo', 'dum', 'duma', 'dumas', 'duns',
24	'duodécuplo', 'duplo', 'duzentos', 'e', 'ela', 'elas', 'ele', 'eles', 'em', 'embora', 'então', 'entre',
25	'entretanto', 'essa', 'essas', 'esse', 'esses', 'esta', 'estas', 'este', 'estes', 'eu', 'exceto', 'fora', 'la',
26	'las', 'lhe', 'lhes', 'lo', 'logo', 'los', 'mas', 'me', 'mediante', 'meio', 'meu', 'meus', 'mil', 'milésimo',
27	'milhão', 'milhões', 'milionésimo', 'mim', 'minha', 'minhas', 'muita', 'muitas', 'muito', 'muitos', 'na',
28	'nada', 'naquela', 'naquelas', 'naquele', 'naqueles', 'naquilo', 'nas', 'nela', 'nelas', 'nele', 'neles',
29	'nem', 'nenhum', 'nenhuma', 'nenhumas', 'nenhuns', 'nessa', 'nessas', 'nesse', 'nesses', 'nesta', 'nestas',
30	'neste', 'nestes', 'ninguém', 'nisso', 'nisto', 'no', 'nonagésimo', 'nongentésimo', 'noningentésimo', 'nono',
31	'nônuplo', 'nos', 'nós', 'nossa', 'nossas', 'nosso', 'nossos', 'nove', 'novecentos', 'noventa', 'o',
32	'octingentésimo', 'octogésimo', 'óctuplo', 'oitavo', 'oitenta', 'oito', 'oitocentos', 'onde', 'onze', 'os',
33	'outra', 'outras', 'outrem', 'outro', 'outros', 'para', 'pela', 'pelas', 'pelo', 'pelos', 'perante', 'pois',
34	'por', 'porém', 'porquanto', 'porque', 'portanto', 'pouca', 'poucas', 'pouco', 'poucos', 'pra', 'pras',
35	'primeiro', 'pro', 'pros', 'prum', 'pruma', 'prumas', 'pruns', 'quadragésimo', 'quadringentésimo',
36	'quádruplo', 'quais', 'quaisquer', 'qual', 'qualquer', 'quando', 'quanta', 'quantas', 'quanto', 'quantos',
37	'quarenta', 'quarto', 'quatro', 'quatrocentos', 'que', 'quem', 'quingentésimo', 'quinhentos', 'quinquagésimo',
38	'quinto', 'quíntuplo', 'quinze', 'salvo', 'se', 'segundo', 'seis', 'seiscentos', 'sem', 'senão',
39	'septingentésimo', 'septuagésimo', 'sessenta', 'sete', 'setecentos', 'setenta', 'sétimo', 'sétuplo', 'seu',
40	'seus', 'sexagésimo', 'sexcentésimo', 'sexto', 'sêxtuplo', 'si', 'sob', 'sobre', 'sua', 'suas',
41	'tanta', 'tantas', 'tanto', 'tantos', 'te', 'terceiro', 'terço', 'teu', 'teus', 'ti', 'tirante', 'toda',
42	'todas', 'todavia', 'todo', 'todos', 'trás', 'trecentésimo', 'três', 'treze', 'trezentos',
43	'trigésimo', 'trinta', 'tríplice', 'triplo', 'tu', 'tua', 'tuas', 'tudo', 'um', 'uma', 'umas', 'undécuplo',
44	'uns', 'vária', 'várias', 'vário', 'vários', 'vigésimo', 'vinte', 'visto', 'você', 'vocês', 'vos', 'vós', 'vossa',
45	'vossas', 'vosso', 'vossos')
46	
47	######################## FUNÇÕES ########################
48	
49	### Função: Ler cada arquivo .txt na pasta de origem
50	identificar_textos <- function(pasta, diretorio){
51	
52	tot_txt <- length(diretorio)
53	enderecos <- vector()
54	identificacao <- vector()
55	
56	for(h in 1:tot_txt){
57	enderecos <- c(enderecos,paste(pasta,"\\",diretorio[h],sep=""))
58	arq_nome <- tolower(diretorio[h])
59	identificacao <- c(identificacao, arq_nome)
60	}
61	
62	id <- 1:tot_txt
63	tab_corpus <- data.frame(id, enderecos, identificacao)
64	return(tab_corpus)
65	}

Linhas	Códigos
66	
67	
68	### Função: Transformar um texto em uma lista de sentenças
69	# Etapa: Excluir caracteres especias, números, etc.
70	
71	identificar_sentencas <- function(tab_textos){
72	
73	tab_txt <- tab_textos
74	lista <- list()
75	
76	for(ii in 1:nrow(tab_txt)){
77	txt <- as.character(tab_txt$enderecos[ii])
78	arq <- readLines(txt, encoding = "UTF-8")
79	arq <- tolower(arq)
80	
81	idx <- 1
82	tam <- length(arq)
83	letra <- ''
84	sentenca <- ''
85	lista_de_sentencas <- vector()
86	
87	for(i in 1:tam){
88	txt <- arq[i]
89	
90	for(j in 1:nchar(txt)){
91	letra <- substr(txt, j, j)
92	#Remover caracteres especiais
93	#intToUtf8(8211) or utf8ToInt(":") = ["--"""]
94	if(letra!=''){
95	if(utf8ToInt(letra)%in%c(45, 58, 8211, 8212, 8216, 8217, 8220,8221, 65279)){
96	letra <- " "
97	}
98	if(letra %in% c(0,1,2,3,4,5,6,7,8,9)){
99	letra <- " "
100	}
101	}
102	
103	resp <- letra %in% c("\"","\'","[","]","(",")","#","$","%", "&", "*", "+")
104	
105	if(resp==FALSE){
106	sentenca <- paste(sentenca,letra,sep="")
107	}
108	
109	if (letra %in% pontuacao_de_parada){
110	if(substr(sentenca,1,1)==' '){
111	sentenca <- substr(sentenca,2,nchar(sentenca))
112	}
113	if(nchar(sentenca)>1){
114	lista_de_sentencas[[idx]] <- sentenca
115	idx <- idx + 1
116	}
117	sentenca <- ''
118	}
119	}
120	
121	if(substr(sentenca, nchar(sentenca), nchar(sentenca))!=" "){
122	sentenca <- paste(sentenca," ",sep="")
123	}
124	}
125	
126	lista[[ii]] <- list(lista_de_sentencas)
127	
128	}
129	return(lista)
130	}

Linhas	Códigos
131	
132	cat("\n\nInforme o diretorio onde estao os arquivos\n\n")
133	pasta <- choose.dir(caption = 'Selecione o diretório:')
134	diretorio <- dir(pasta)
135	
136	stpwrd <- '0'
137	stpwrd <- tolower(readline(prompt="Stopwords? 1)Remover; 2)Manter; 3)So stopwords. Opcao: "))
138	
139	cat('\n')
140	cat('**** Início da Execução ****')
141	cat('\n\nInício do processamento: ', date(),'\n\n')
142	
143	#1. Tabular endereços dos arquivos
144	tab_textos <- identificar_textos(pasta, diretorio)
145	
146	#2. Listar sentenças por autoria
147	lis_sentencas <- identificar_sentencas(tab_textos)
148	
149	##
150	# FASE 02: Converte as frases em vetores de palavras #
151	##
152	
153	lis_palavras <- list()
154	pontos <- c('\\.','\\!','\\?','\\,','\\;') #remover estas pontuacoes
155	
156	tam.i <- length(lis_sentencas)
157	for(i in 1:tam.i){
158	cat(paste0(round(i/tam.i*30,digits = 1), '% processado. \n', '\b\r'))
159	frases <- unlist(lis_sentencas[[i]])
160	a <- list()
161	tam.j <- length(frases)
162	
163	for(j in 1:tam.j){
164	frase <- frases[j]
165	#remover pontuação
166	for(k in 1:length(pontos)){
167	frase <- gsub(pontos[k],",frase)
168	}
169	vetor <- strsplit(frase, ' ')
170	vetor <- unlist(vetor)
171	
172	a[[j]] <- vetor
173	}
174	lis_palavras[[i]] <- a
175	}
176	
177	##
178	# FASE 03: Tabular resultados #
179	##
180	
181	lis_itens <- list()
182	'%notin%' <- Negate('%in%')
183	
184	tam.i <- length(lis_palavras)
185	for(i in 1:tam.i){
186	l <- list()
187	tam.j <- length(lis_palavras[[i]])
188	for(j in 1:tam.j){
189	v <- unlist(lis_palavras[[i]][j])
190	v <- unique(v)
191	v <- v[order(v)]
192	id <- which(v == '')
193	if(!(identical(integer(0),id))){
194	v <- v[-id]
195	}

Linhas	Códigos
196	#remover stopword
197	if(stpwrd=='1'){
198	id <- which(v %notin% stopwords)
199	if(!(identical(integer(0),id))){
200	v <- v[id]
201	} else {
202	v <- ''
203	}
204	}
205	#trabalhar só com stopword
206	if(stpwrd=='3'){
207	id <- which(v %in% stopwords)
208	if(!(identical(integer(0),id))){
209	v <- v[id]
210	} else {
211	v <- ''
212	}
213	}
214	l[[j]] <- v
215	}
216	lis_itens[[i]] <- l
217	}
218	
219	#(1)tabular o resultado com tabela verdade (TRUE or FALSE)
220	vet.palavras <- vector()
221	tot.livros <- length(lis_itens)
222	tot.linhas <- 0
223	id.autor <- vector()
224	for(i in 1:tot.livros){
225	tam.j <- length(lis_itens[[i]])
226	for(j in 1:tam.j){
227	v <- unlist(lis_itens[[i]][j])
228	vet.palavras <- c(vet.palavras,v)
229	tot.linhas <- tot.linhas + 1
230	id.autor <- c(id.autor,i)
231	}
232	}
233	
234	vet.palavras <- unique(vet.palavras)
235	
236	
237	
238	m <- matrix(FALSE, nrow = tot.linhas, ncol = length(vet.palavras))
239	nomes <- vet.palavras
240	colnames(m)<-nomes
241	
242	linha <- 0
243	for(i in 1:tot.livros){
244	tam.j <- length(lis_itens[[i]])
245	a <- lis_itens[[i]]
246	for(j in 1:tam.j){
247	linha <- linha + 1
248	v <- unlist(a[j])
249	for(k in 1:length(v)){
250	id <- which(nomes%in%v[k])
251	m[linha,id] <- TRUE
252	}
253	}
254	}
255	#(1)
256	
257	#Right outer:
258	df1 <- as.data.frame(id.autor)
259	colnames(df1) <- 'id'

Linhas	Códigos
260	df2 <- merge(x = df1, y = tab_textos, by = "id", all.y = TRUE)
261	df3 <-as.data.frame(m)
262	
263	df4 <- cbind(df2,df3)
264	
265	arquivos <- tolower(diretorio)
266	for(i in 1:length(arquivos)){
267	id <- which(df4$identificacao==arquivos[i])
268	arq <- df3[id,]
269	nome <- arquivos[i]
270	if(stpwrd=='1'){
271	nome <- gsub('.txt', 'ss - itens.csv', nome)
272	} else {
273	if(stpwrd=='2'){
274	nome <- gsub('.txt', 'cs - itens.csv', nome)
275	}else{
276	nome <- gsub('.txt', 'st - itens.csv', nome)
277	}
278	}
279	write.csv2(arq,nome,row.names = FALSE)
280	}
281	
282	cat(paste0(100, '% processado.\n', "\b\r"))
283	
284	cat('\n')
285	cat('\nTérmino do processamento: ', date(),'\n\n')
286	cat('Os arquivos de itens estão em:\n', getwd(),'.\n\n')
287	cat('**** Fim da Execução ****')
288	cat('\n\n##(fim)\n\n')
289	
290	beep(sound = 5)
291	beep(sound = 1)
292	
293	#Limpar variáveis da memória
294	rm(list=ls(all=TRUE))

Algoritmo: Etapa 02 - Transformar conjuntos de itens em regras de associação

Linhas	Códigos
1	#
2	# Converte uma tabela de verdade em regras de associação
3	# - Os títulos das colunas representam os itens (e.g., 'laranja', 'maça', etc)
4	# - Os registros com TRUE ou FALSE representam a 'presença do item' com TRUE e ausência com FALSE
5	#
6	
7	library('arules')
8	library('beepr')
9	
10	cat('\n##(Início)\n\n')
11	
12	resp <- 0
13	resp <- readline(prompt="Infome um valor maior que 0 e menor que 1 para o cálculo a confiança: ")
14	
15	resp <- as.numeric(resp)
16	
17	if(resp >= 0 & resp <=1){
18	
19	cat("\n\nInforme o diretorio onde estao os arquivos\n\n")
20	pasta <- choose.dir(default = "C:\\00_Testes", caption = "Selecione o diretório:")
21	diretorio <- dir(pasta)
22	
23	for(i in 1:length(diretorio)){
24	
25	endereco <- paste(pasta,'\\',diretorio[i],sep='')
26	
27	base_dados <- read.table(endereco, header = TRUE, sep = ';')
28	
29	base_dados_transacional <- as(base_dados,'transactions')
30	
31	regras <- apriori(data = base_dados_transacional,
32	parameter = list(minlen=2, maxlen=3, supp=0.02, conf=resp))
33	
34	
35	if(nrow(regras@quality)==0){
36	
37	cat('\nNão há regras para confiança ',resp,' em ', diretorio[i],'.')
38	
39	} else {
40	
41	tab.regras <- DATAFRAME(regras)
42	nome <- paste('- ra - conf_',resp,'.csv', sep = '')
43	nome.arq <- grep('- itens.csv',basename(endereco))
44	
45	if(identical(nome.arq,integer(0))){
46	nome.arq <- gsub('.csv', nome, nome.arq)
47	} else {
48	nome.arq <- gsub('- itens.csv', nome, basename(endereco))
49	}
50	
51	write.csv2(tab.regras, nome.arq)
52	cat('\nArquivo:', nome.arq)
53	rm(tab.regras,base_dados_transacional)
54	}
55	}
56	} else {
57	cat('\n\n\tValor inválido!')
58	}
59	
60	beep(sound = 5)
61	beep(sound = 1)
62	
63	rm(nome, endereco, base_dados, regras, nome.arq)
64	rm(resp, i, pasta, diretorio)
65	
66	cat('\n##(Fim)\n\n')

Algoritmo: Etapa 03 - Transformar regras de associação em redes

```
#Limpar variáveis da memória
rm(list=ls(all=TRUE))

#
# Converte uma tabela com regras de associação em uma rede
#
library(igraph)
library('beepr')

cat('\n#################################################################(Início)\n')
cat(date(),'\n')

cat("\n\nInforme o diretorio onde estao os arquivos\n\n")
pasta <- choose.dir(default = "C:\\00_Testes", caption = "Selecione o diretório:")
diretorio <- dir(pasta)

tam <- length(diretorio)

for(i in 1:tam){

 cat('\r        ')
 cat('\r',i,'/',tam)

 endereco <- paste(pasta,'\\',diretorio[i],sep='')

 #Tabela com as regras
 regras <- read.table(endereco, header = TRUE, sep = ';',dec = ',')

 ini <- c(1.0, 0.8, 0.6, 0.4, 0.2, 0.0)
 fim <- c(2.0, 1.0, 0.8, 0.6, 0.4, 0.2)

 for(ii in 1:6){

  id <- which(regras$confidence < fim[ii] & regras$confidence >= ini[ii] )

  if( !identical(id,integer(0)) ){

   tab.regras <- regras[id,]
   #View(tab.regras)

   # #Tabela de adjacencias entre itens
   # tab.adj <- data.frame()
   #
   # for(i in 1:nrow(tab.regras)){
   #   a <- as.vector(tab.regras$LHS[i])
   #   b <- as.vector(tab.regras$RHS[i])
   #   if(a != '{}' && b != '{}'){
   #     a <- unlist(strsplit(a, "[\\{\\}\\,]"))
   #     b <- unlist(strsplit(b, "[\\{\\}\\,]"))
   #     for(j in 1:length(a)){
   #       if(a[j]!=''){
   #         for(k in 1:length(b)){
   #           if(b[k]!=''){
   #             c <- c(a[j],b[k])
   #             tab.adj <- rbind(tab.adj,c)
   #           }
   #         }
   #       }
   #     }
   #   }
   # }
   # }
   #
```

Linhas	Códigos
64	#Tabela de adjacencias entre itens
65	tab.adj <- data.frame()
66	v01 <- vector()
67	v02 <- vector()
68	cont <- 0
69	
70	tab.regras$LHS <- gsub('[\\{\\}]', '', tab.regras$LHS)
71	tab.regras$RHS <- gsub('[\\{\\}]', '', tab.regras$RHS)
72	
73	lis.lhs <- strsplit(tab.regras$LHS, ",")
74	lis.rhs <- strsplit(tab.regras$RHS, ",")
75	
76	for(t in 1:nrow(tab.regras)){
77	
78	if(lis.lhs[t] != '' && lis.rhs[t] != ''){
79	a <- unlist(lis.lhs[t])
80	b <- unlist(lis.rhs[t])
81	for(j in 1:length(a)){
82	for(k in 1:length(b)){
83	cont <- cont + 1
84	v01[cont] <- a[j]
85	v02[cont] <- b[k]
86	}
87	}
88	}
89	}
90	
91	tab.adj <- cbind.data.frame(v01,v02)
92	
93	if(nrow(tab.adj) > 0) {
94	#sumarizar o peso dos arcos
95	colnames(tab.adj) <- c('v1', 'v2')
96	tab.adj <- aggregate(tab.adj$v1, by=list(tab.adj$v1 , tab.adj$v2), FUN = length)
97	colnames(tab.adj) <- c('v1', 'v2', 'peso')
98	# View(tab.adj)
99	
100	g <- graph.data.frame(tab.adj, directed = TRUE)
101	
102	E(g)$weight <- tab.adj$peso
103	E(g)$label <- E(g)$weight
104	
105	#g <- set.edge.attribute(g, "arrow.size", value = 0.1)
106	#E(g)$curved <- 0.3
107	# plot(g)
108	
109	#Extrair um arquivo .net
110	V(g)$id <- V(g)$name
111	
112	nome.arq <- gsub('.csv', ' - rede.net', basename(endereco))
113	nome.arq <- paste(ii,nome.arq,sep = '-')
114	
115	#converter nome do arquivo para o encoding nativo
116	x <- nome.arq
117	y <- enc2native(x)
118	
119	write.graph(g, file = y, format = 'pajek')
120	cat('\n\n A rede: ', nome.arq)
121	}
122	}
123	}
124	}
125	
126	beep(sound = 5)
127	beep(sound = 1)
128	cat('\n',date(),'\n')
129	cat('##(Fim)\n\n')

Algoritmo: Etapa 04 - Extrair métricas das redes

Código em R do algoritmo Extrair propriedade das redes

Linhas	Códigos
1	#
2	# Algoritmo: Extrair métricas das redes dirigidas e ponderadas
3	#
4	# - Entrada: Diretório onde estão as redes.
5	#
6	# - Arquivos: Os arquivos precisam identificar o nome da
7	# classe/objeto com [nome] - [detalhe].net
8	# (e.g. Elefante - 01.net; Elefante - 02.net;)
9	#
10	
11	#Limpar variáveis da memória
12	rm(list=ls(all=TRUE))
13	
14	tot.metr <- 16 #métricas
15	
16	library(igraph)
17	library('centiserve')
18	library('beepr')
19	##########################
20	# Normalizar
21	norm_min_max <- function(x){
22	min = 0
23	max = 1
24	valor <- (x - min(x)) / (max(x)-min(x)) * (max - min) + min
25	return(valor)
26	}
27	
28	##########################
29	# Função converter net para igraph
30	de.NET.para.IGRAPH <- function(txt){
31	
32	vertice.label <- vector()
33	linha <- ''
34	tam <- nrow(txt)
35	
36	n <- 0
37	ini <- ifelse(substr(txt[1,1],1,1)=='*',2,1)
38	for(i in ini:tam){
39	n <- n + 1
40	linha <- txt[i,1]
41	if(substr(linha,1,1)=='*'){
42	i <- i + 1
43	break
44	} else {
45	vertice.label[n] <- gsub("[0-9]", "", linha)
46	}
47	}
48	
49	tam <- length(vertice.label)
50	m <- matrix(0, nrow = tam, ncol = tam)
51	colnames(m) <- vertice.label
52	rownames(m) <- vertice.label
53	
54	tam <- nrow(txt)
55	vetor <- vector()
56	x <- y <- z <- ''
57	
58	for(k in i:tam){
59	
60	linha <- txt[k,1]
61	vetor <- unlist(strsplit(linha,' '))
62	
63	x <- as.numeric(vetor[1])

Linhas	Códigos
64	y <- as.numeric(vetor[2])
65	z <- as.numeric(vetor[3])
66	
67	m[x,y] <- m[x,y] + z
68	
69	}
70	g <- graph_from_adjacency_matrix(m, mode='directed',weighted = TRUE)
71	return (g)
72	}
73	#########################
74	
75	
76	#Função: Extrair atributo da rede
77	fn.atrib.rede <- function(opcao, g){
78	atributo <- 0
79	
80	#Centralidade de intermedição
81	if(opcao==1){ atributo <- betweenness(g, directed = TRUE) }
82	#Grau Ponderado
83	if(opcao==2){ atributo <- strength(g, mode='all') }
84	#Grau Ponderado (Entrada)
85	if(opcao==3){ atributo <- strength(g, mode='in') }
86	#Grau Ponderado (Saída)
87	if(opcao==4){ atributo <- strength(g, mode='out') }
88	#Grau
89	if(opcao==5){ atributo <- degree(g, mode='all') }
90	#Grau (Entrada)
91	if(opcao==6){ atributo <- degree(g, mode='in') }
92	#Grau (Saída)
93	if(opcao==7){ atributo <- degree(g, mode='out') }
94	#Excentricidade
95	if(opcao==8){ atributo <- eccentricity(g, mode='all') }
96	#Excentricidade (Entrada)
97	if(opcao==9){ atributo <- eccentricity(g, mode='in') }
98	#Excentricidade (Saída)
99	if(opcao==10){ atributo <- eccentricity(g, mode='out') }
100	#Grau Ponderado + grau
101	if(opcao==11){ atributo <- strength(g, mode='all') + degree(g, mode='all')}
102	#Grau Ponderado + grau (Entrada)
103	if(opcao==12){ atributo <- strength(g, mode='in') + degree(g, mode='in')}
104	#Grau Ponderado + grau (Saída)
105	if(opcao==13){ atributo <- strength(g, mode='out') + degree(g, mode='out')}
106	#Laplace
107	if(opcao==14){ atributo <- laplacian(g, mode='all') }
108	#Laplace (Entrada)
109	if(opcao==15){ atributo <- laplacian(g, mode='in') }
110	#Laplace (Saída)
111	if(opcao==16){ atributo <- laplacian(g, mode='out') }
112	
113	return(atributo)
114	}
115	
116	#Função: Extrair nome da classe
117	# Extrai da classe a partir da primeira palavra no nome do arquivo
118	# inicia na posição 1 e termina ante do símbolo '-'
119	fn.nome <- function(n){
120	nome <- n
121	n <- substr(n,3,nchar(n))
122	for(i in 1:nchar(n)){
123	if(substr(n,i,i)=='-'){
124	nm <- substr(n,1,i-1)
125	break
126	}

Linhas	Códigos
127	}
128	if(substr(nm,nchar(nm),nchar(nm))==' '){
129	nm <- substr(n,1,nchar(nm)-1)
130	}
131	return(nm)
132	}
133	
134	
135	cat('\n##(Início)')
136	cat('\n',date(),'\n')
137	
138	cat("\n\n\tInforme o diretorio onde estao os arquivos.\n")
139	pasta <- choose.dir(caption = "Selecione o diretório:")
140	diretorio <- dir(pasta)
141	
142	for(nn in 1:6){
143	
144	cat('\r ')
145	cat('\r',nn,'/',6)
146	
147	lis.g <- list()
148	arquivos <- vector()
149	vertices.vt <- vector()
150	
151	ctrl <- ''
152	cont <- 0
153	
154	for(i in 1:length(diretorio)){
155	
156	letra <- as.numeric(substr(diretorio[i],1,1))
157	
158	if(letra==nn){
159	ctrl <- '*'
160	endereco <- paste(pasta,'\\',diretorio[i],sep='')
161	txt <- read.delim(endereco, encoding = 'ANSI')
162	g <- de.NET.para.IGRAPH(txt)
163	#g <- read.graph(endereco, format = 'pajek')
164	cont <- cont + 1
165	lis.g[[cont]] <- g
166	arquivos[cont] <- diretorio[i]
167	vertices.vt <- c(vertices.vt,V(g)$name)
168	}
169	}
170	
171	if(ctrl == '*') {
172	objetos <- vector()
173	for(i in 1:length(arquivos)){
174	objetos <- c(objetos, fn.nome(arquivos[i]))
175	}
176	
177	nome.arq <- 'indefinido.csv'
178	
179	resp <- 1
180	for(ii in 1:tot.metr){
181	
182	vertices.vt <- unique(vertices.vt)
183	id <- order(vertices.vt)
184	vertices.vt <- vertices.vt[id]
185	
186	tam <- length(lis.g)
187	tab <- as.data.frame(matrix(0,nrow = tam, ncol=length(vertices.vt)))
188	colnames(tab) <- vertices.vt
189	

Linhas	Códigos
190	for(i in 1:tam){
191	a <- vector()
192	b <- vector()
193	g <- lis.g[[i]]
194	
195	atributo <- fn.atrib.rede(resp, g)
196	
197	nome.atrib <- names(atributo)
198	ids <- order(nome.atrib)
199	nome.atrib <- nome.atrib[ids]
200	
201	id <- which(vertices.vt%in%nome.atrib)
202	
203	tab[i,id] <- atributo[ids]
204	
205	}
206	
207	tab.norm <- tab
208	for(i in 1:nrow(tab.norm)){
209	tab.norm[i,] <- norm_min_max(tab[i,])
210	}
211	
212	tab.ext <- cbind(objetos,tab.norm)
213	# View(tab.ext)
214	
215	if(resp == 1) { nome.arq <- paste(nn,'Intermediacao_norm.csv',sep = '-') }
216	if(resp == 2) { nome.arq <- paste(nn,'Grau_ponderado_norm.csv',sep = '-') }
217	if(resp == 3) { nome.arq <- paste(nn,'Grau_ponderado_entrada_norm.csv',sep = '-') }
218	if(resp == 4) { nome.arq <- paste(nn,'Grau_ponderado_saida_norm.csv',sep = '-') }
219	if(resp == 5) { nome.arq <- paste(nn,'Grau_norm.csv',sep = '-') }
220	if(resp == 6) { nome.arq <- paste(nn,'Grau_entrada_norm.csv',sep = '-') }
221	if(resp == 7) { nome.arq <- paste(nn,'Grau_saida_norm.csv',sep = '-') }
222	if(resp == 8) { nome.arq <- paste(nn,'Excentricidade_norm.csv',sep = '-') }
223	if(resp == 9) { nome.arq <- paste(nn,'Excentricidade_entrada_norm.csv',sep = '-') }
224	if(resp ==10) { nome.arq <- paste(nn,'Excentricidade_saida_norm.csv',sep = '-') }
225	if(resp ==11) { nome.arq <- paste(nn,'Grau_mais_peso_norm.csv',sep = '-') }
226	if(resp ==12) { nome.arq <- paste(nn,'Grau_entrada_mais_peso_norm.csv',sep = '-') }
227	if(resp ==13) { nome.arq <- paste(nn,'Grau_saida_mais_peso_norm.csv',sep = '-') }
228	if(resp ==14) { nome.arq <- paste(nn,'Laplace_norm.csv',sep = '-') }
229	if(resp ==15) { nome.arq <- paste(nn,'Laplace_norm_entrada.csv',sep = '-') }
230	if(resp ==16) { nome.arq <- paste(nn,'Laplace_norm_saida.csv',sep = '-') }
231	
232	write.csv2(tab.ext,nome.arq)
233	
234	resp <- resp + 1
235	}
236	}
237	}
238	cat('\n\nOrigem: ', pasta)
239	cat('\n\n',date(),'\n')
240	cat('\n##(Fim)\n\n')
241	
242	
243	beep(sound = 5)
244	beep(sound = 1)
245	

Algoritmo: Processos de Predições

Linhas	Códigos
1	
2	
3	#
4	# Algoritmo: Predições por Naive Bayes, SVM e Floresta aleatória
5	#
6	# Entrada:
7	# - Tabela com objetos e atributos dos objetos:
8	# -- 1ª coluna: objetos;
9	# -- Demais colunas: atributos dos objetos;
10	#
11	
12	#Limpar variáveis da memória
13	rm(list=ls(all=TRUE))
14	
15	library('e1071')
16	library(randomForest)
17	
18	cat('\n###(Início)')
19	cat('\n\n')
20	resp <- ''
21	resp <- tolower(readline(prompt="\tConverter para os valores para 0 ou 1 (s/n): "))
22	
23	if(resp=='n'){
24	resp <- readline(prompt="\tNormalizar os dados (s/n): ")
25	if(resp=='s'){
26	resp <- 'norm'
27	}
28	}
29	
30	if(resp %in% c('s','n','norm')){
31	cat('\n\n\tInforme o arquivo .csv com objetos e atributos.\n\n')
32	endereco <- file.choose()
33	
34	#Tabela com as regras
35	tab.tmp <- read.table(endereco, header = TRUE, sep = ';', dec = ",")
36	colnames(tab.tmp)[1] <- 'Classes'
37	
38	c1 <- as.factor(tab.tmp$Classes)
39	cn <- tab.tmp[,2:ncol(tab.tmp)]
40	
41	if(tolower(resp)=='s'){
42	for(i in 1:ncol(cn)){
43	id <- which(cn[,i]>0)
44	if(!identical(id,integer(0))){
45	cn[id,i] <- 1
46	}
47	}
48	rm(id)
49	}
50	
51	if(resp == 'norm'){
52	norm_min_max <- function(x){
53	min = 0
54	max = 1
55	valor <- (x - min(x)) / (max(x)-min(x)) * (max - min) + min
56	#
57	id <- which(is.nan(valor))
58	if(!identical(id,integer(0))){
59	valor[id] <- 0
60	}
61	#
62	return(valor)
63	}

Linhas	Códigos
64	for(i in 1:ncol(cn)){
65	v <- as.vector(cn[,i])
66	if(sum(v)>0){
67	v <- norm_min_max(v)
68	cn[,i] <- v
69	}
70	}
71	rm(v)
72	}
73	
74	tab <- cbind(c1,cn)
75	objetos <- unique(tab[,1])
76	n <- nrow(tab)
77	
78	oficial <- vector()
79	previsao <- vector()
80	prev_nb <- vector()
81	prev_sv <- vector()
82	prev_rf <- vector()
83	
84	for(i in 1:n){
85	teste <- i
86	treino <- setdiff(1:n, teste)
87	
88	tab.treino <- tab[treino,]
89	tab.teste <- tab[teste,]
90	
91	#Naive Bayes
92	prev.nb <- naiveBayes(tab.treino[,-1], tab.treino[,1])
93	#SVM
94	prev.sv <- svm(c1 ~ . , data = tab.treino, scale = FALSE, kernel = "linear")
95	#Floresta aleatória
96	prev.rf <- randomForest(c1 ~ ., data = tab.treino, importance = FALSE, ntree=500)
97	
98	oficial <- c(oficial, as.character(tab$c1[i]))
99	prev_nb <- c(prev_nb, as.character(predict(prev.nb, tab.teste[,-1], type = 'class')))
100	prev_sv <- c(prev_sv, as.character(predict(prev.sv, tab.teste)))
101	prev_rf <- c(prev_rf, as.character(predict(prev.rf, tab.teste[,-1], type = 'class')))
102	
103	}
104	
105	resposta <- cbind(oficial,prev_nb, prev_sv, prev_rf)
106	View(resposta)
107	
108	write.csv2(resposta,'Predição.csv')
109	
110	} else {
111	cat('\n\n\t Resposta invalida!')
112	}
113	cat('##(Fim)\n\n')
114	
115	
116	
117	
118	
119	
120	
121	
122	
123	
124	
125	
126	

Método para analisar autoria de textos baseado em regras de associação e redes de palavras

Cleônidas Tavares de Souza Júnior

Salvador, Outubro, 2022.