**UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS**

Camilo Akimushkin Valencia

Propriedades de redes aplicadas à atribuição de autoria

São Carlos

2017

Camilo Akimushkin Valencia

Propriedades de redes aplicadas à atribuição de autoria

Tese apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para obtenção do título de Doutor em Ciências.

Área de concentração: Física Básica

Orientador: Prof. Dr. Osvaldo Novais de Oliveira Jr

Versão corrigida

(Versão original disponível na Unidade que aloja o Programa)

São Carlos

2017

AUTORIZO A REPRODUÇÃO E DIVULGAÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

Ficha catalográfica revisada pelo Serviço de Biblioteca e Informação do IFSC,
com os dados fornecidos pelo(a) autor(a)

Akimushkin, Camilo Valencia

Propriedades de redes aplicadas à atribuição de autoria / Camilo Valencia Akimushkin; orientador Osvaldo Novais de Oliveira Jr - versão corrigida -- São Carlos, 2017.

132 p.

Tese (Doutorado - Programa de Pós-Graduação em Física Básica) -- Instituto de Física de São Carlos, Universidade de São Paulo, 2017.

1. Línguas naturais. 2. Redes complexas. 3. Séries temporais. 4. Reconhecimento de autoria. I. Oliveira Jr, Osvaldo Novais de, orient. II. Título.

À minha mãe.

AGRADECIMENTOS

À minha mãe, a pessoa que mais amo; à minha avó, a pessoa mais admirável com que eu tive a honra de ter compartilhado experiências; e à minha irmã, a pessoa que mais xingo e que ainda é a minha irmãzinha.

À Gisela Ibañez; à gigante galera colombiana em São Carlos que não teria como listar aqui.

Aos meus amigos e colegas com os que tive a fortuna de compartilhar ideias entre eles ao Diego Raphael Amâncio, ao Filipi Nascimento Silva, ao Thomas Peron, ao César Comin e ao Henrique Ferraz de Arruda.

Aos meus amigos brasileiros a maioria deles agora longe. Ao Amadeu.

Por último mas não menos importante, ao meu orientador, Osvaldo Novais de Oliveira Jr. Fico grato por tudo o feito ao longo deste doutorado.

“By and large, language is a tool for concealing the truth.”

George Carlin

*“Like everything metaphysical the harmony between thought
and reality is to be found in the grammar of the language.”*

Ludwig Wittgenstein

RESUMO

AKIMUSHKIN, C. V. **Propriedades de redes aplicadas à atribuição de autoria.** 2017. 132p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2017.

O reconhecimento de autoria é uma área de pesquisa efervescente, com muitas aplicações, incluindo detecção de plágio, análise de textos históricos, reconhecimento de mensagens terroristas ou falsificação de documentos. Modelos teóricos de redes complexas já são usados para o reconhecimento de autoria, mas alguns aspectos importantes têm sido ignorados. Neste trabalho, exploramos a dinâmica de redes de co-ocorrência e a relação com as palavras que representam os nós e descobrimos que ambas são claras assinaturas de autoria. Com otimização dos descritores da topologia das redes e de algoritmos de aprendizado de máquina, foi possível obter taxas de acerto maiores que 85%, sendo atingida uma taxa de 98.75% em um caso específico, para coleções de 80 livros, cada uma compilada de 8 autores de língua inglesa com 10 livros por autor. Esta tese demonstra que existem ainda aspectos inexplorados das redes de co-ocorrência de textos, o que deve permitir avanços ainda maiores no futuro próximo.

Palavras-chave: Línguas naturais. Redes complexas. Séries temporais. Reconhecimento de autoria.

ABSTRACT

AKIMUSHKIN, C. V. **Network features for authorship attribution**. 2017. 132p. Tese (Doutorado em Ciências) - Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2017.

Authorship attribution is an active research area with many applications, including detection of plagiarism, analysis of historical texts, terrorist message identification or document falsification. Theoretical models of complex networks are already used for authorship attribution, but some issues have been ignored. In this thesis, we explore the dynamics of co-occurrence networks and the role of words, and found that they are both clear signatures of authorship. Using optimized descriptors for the network topology and machine learning algorithms, it has been possible to achieve accuracy rates above 85%, with a rate of 98.75% being reached in a particular case, for collections of 80 books produced by 8 English-speaking writers with 10 books per author. It is also shown that there are still many unexplored aspects of co-occurrence networks of texts, which seems promising for near future developments.

Keywords: Spoken languages. Complex networks. Time series. Authorship attribution.

LISTA DE ILUSTRAÇÕES

Figura 1 – Nós de uma rede coloridos segundo diferentes centralidades: a, intermediação; b, proximidade; c, autovetor; d, grau; e, harmônico; f, Katz. . .	35
Figura 2 – Lei de Zipf para varias línguas.	39
Figura 3 – Lei de Zipf estendida.	40
Figura 4 – Probabilidade das palavras.	40
Figura 5 – Rede de co-ocorrência.	40
Figura 6 – Conectividade	42
Figura 7 – Menor caminho	42
Figura 8 – Centralidade de intermediação	42
Figura 9 – Centralidade de carga	42
Figura 10 – Coeficiente de clusterização	42
Figura 11 – Intermitência	42
Figura 12 – Retrato de fase do mapa logístico com $r = 3.4$	45
Figura 13 – Retrato de fase de uma série estocástica.	45
Figura 14 – Autocorrelação considerada baixa. As linhas pontilhadas marcam os níveis $ACF = \pm 0.05$	46
Figura 15 – Variedade bidimensional embebida num espaço tridimensional.	56
Figura 16 – Distâncias entre dois pontos.	57
Figura 17 – Trajetória geodésica	57
Figura 18 – Distância (azul) e trajetória (vermelho) geodésicas.	57
Figura 19 – Exemplo de árvore de decisão.	61
Figura 20 – Exemplo de construção da árvore de decisão.	62
Figura 21 – Rede de função de base radial.	66
Figura 22 – Espaço de atributos	67
Figura 23 – Espaço de funções de base radial	67
Figura 24 – Projeção bidimensional das cidades da Europa usando MDS.	69
Figura 25 – Construção das séries temporais.	72
Figura 26 – Retrato de fase da clusterização	73
Figura 27 – Retrato de fase do número de nós	73
Figura 28 – Autocorrelação da centralidade de carga para o livro War and Peace. .	74
Figura 29 – Transformada de Fourier para o livro War and Peace.	75
Figura 30 – Séries temporais para as métricas do livro Moby Dick do Herman Melville.	75
Figura 31 – Histogramas da conectividade por autor.	76

Figura 32 – Atributos usando a seleção baseada em variância. Iniciais do eixo horizontal: C, clusterização; D, diâmetro; R, radio; Cq, numero de cliques; L, centralidade de carga; T, transitividade; B, centralidade de intermediação; S, Menor caminho; K, conetividade; I, intermitência; N, número de nós; e E, número de eixos.	78
Figura 33 – Taxas de acerto usando a seleção baseada em variância.	78
Figura 34 – Taxas de acerto usando a seleção baseada em acerto.	79
Figura 35 – Melhores atributos usando a seleção baseada em acerto. Iniciais do eixo horizontal: C, clusterização; D, diâmetro; R, radio; Cq, numero de cliques; L, centralidade de carga; T, transitividade; B, centralidade de intermediação; S, Menor caminho; K, conetividade; I, intermitência; N, número de nós; e E, número de eixos.	80
Figura 36 – Resultados da extração de atributos.	81
Figura 37 – Projecao Isomap bidimensional.	82
Figura 38 – Distribuição da centralidade de intermediação e palavras principais do livro The Memoirs of Sherlock Holmes.	86
Figura 39 – Distribuição da centralidade de intermediação e palavras principais do livro The Return of Sherlock Holmes.	87
Figura 40 – Matriz de distâncias usando centralidade de intermediação.	89
Figura 41 – Mapeamento de uma matriz de distâncias num espaço bidimensional usando MDS.	89
Figura 42 – Exemplo ilustrativo do método do cotovelo.	91
Figura 43 – Taxas de acerto com o método introduzido e com TF-IDF.	92
Figura 44 – Taxas de acerto aplicando Isomap.	121

LISTA DE TABELAS

Tabela 1 – Regras de associação para os diferentes valores da classe “Panorama”. .	61
Tabela 2 – Tabela de dados com atributos categóricos.	64
Tabela 3 – Parâmetros do RBFN.	67
Tabela 4 – Resumo das taxas de acerto obtidas com o conjunto original de 48 atributos e aplicando seleção de atributos e extração de atributos. Para PCA e Isomap foi usado $n_{componentes} = 10$ e para Isomap $n_{vizinhos} = 10$. O máximo erro absoluto foi de 5.7% obtido com KNN.	77
Tabela 5 – Palavras principais dos livros de Arthur Conan Doyle.	88
Tabela 6 – Produto interno dos livros de Arthur Conan Doyle.	88
Tabela 7 – Taxas de acerto (em percentagem) obtidas com o método proposto e com TF-IDF.	91
Tabela 8 – Averages of p_{values} obtained with various tests.	115
Tabela 9 – Número de séries ajustadas com um modelo ARIMA(p,d,q) para diferentes valores do parâmetro p	117
Tabela 10 – Número de séries ajustadas com um modelo ARIMA(p,d,q) para diferentes valores do parâmetro d	118
Tabela 11 – Número de séries ajustadas com um modelo ARIMA(p,d,q) para diferentes valores do parâmetro q	118
Tabela 12 – Número de séries ajustadas com um modelo ARIMA(p,d,q) para diferentes valores da soma $p + d + q$	119
Tabela 13 – Séries ajustadas com um modelo ARIMA(p,d,q) que possuem o maior valor da soma $p + d + q$	119
Tabela 14 – Taxas de acerto obtidas com KNN aplicando Isomap para vários valores dos parâmetros $n_{componentes}$ e $n_{vizinhos}$	122
Tabela 15 – Palavras retiradas dos textos (primeira parte).	131
Tabela 16 – Palavras retiradas dos textos (segunda parte).	132
Tabela 17 – Símbolos retirados dos textos.	132

LISTA DE ABREVIATURAS E SIGLAS

AR	Modelo auto-regressivo
MA	Modelo de média móvel
ARIMA	Modelo auto-regressivo integrado de média móvel
PCA	Principal Component Analysis
Isomap	Isometric feature mapping
0R	Zero Rule
1R	One Rule
KNN	k-Nearest Neighbors
RBFN	Radial Basis Function Networks
MDS	Multi-Dimensional Scaling
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
NB	Naive Bayes
C	Clusterização
D	Diâmetro
R	Radio
C	Número de cliques
L	Centralidade de carga
T	Transitividade
B	Centralidade de intermediação
S	Menor caminho
K	Conetividade
I	Intermitência
N	Número de nós
E	Número de eixos

BoW	Bag of Words
TF-IDF	Term Frequency-Inverse Document Frequency
e.g.	exempli gratia
i.e.	id est

LISTA DE SÍMBOLOS

α	Letra grega minúscula alfa
β	Letra grega minúscula beta
Δ	Letra grega delta
ε	Letra grega minúscula épsilon
λ	Letra grega minúscula lambda
ν	Letra grega minúscula nu
μ	Letra grega minúscula mu
ω	Letra grega minúscula omega
∞	Símbolo do infinito
$\in$	Pertence

SUMÁRIO

1	APRESENTAÇÃO	25
1.1	Motivação	25
1.2	Objetivos	26
2	LINGUAGEM	27
2.1	Áreas da Linguística	27
2.1.1	Fonologia	27
2.1.2	Morfologia	27
2.1.3	Lexicografia	27
2.1.4	Semântica	28
2.1.5	Sintaxe	28
2.2	Análise automática de textos	28
2.3	Pré-processamento	29
2.3.1	Stemming	30
2.3.2	Lematização	30
2.3.2.1	Part of speech tagging	30
2.3.3	Remoção de elementos	31
3	REDES COMPLEXAS	33
3.1	Complexidade	33
3.2	Métricas de rede	34
3.2.1	Peso	35
3.2.2	Conectividade	35
3.2.3	Menor caminho	36
3.2.4	Centralidade de intermediação	36
3.2.5	Centralidade de carga	37
3.2.6	Intermitência	37
3.2.7	Coeficiente de clusterização	37
3.2.8	Transitividade	37
3.2.9	Número de cliques	38
3.2.10	Diâmetro	38
3.2.11	Raio	38
3.3	Redes de Co-ocorrência	38
3.4	Métricas para uma rede de co-ocorrência	41
4	SÉRIES TEMPORAIS	43

4.1	Não-linearidade e caos	43
4.2	Estacionariedade	45
4.2.1	Auto-correlação	45
4.2.1.1	Teorema de Wiener-Khinchin	46
4.2.2	Testes de valor-p	47
4.2.3	Modelo ARIMA	49
4.2.3.1	Modelo auto-regressivo AR(p)	49
4.2.3.2	Modelo de média móvel MA(q)	50
4.2.3.3	Modelo integrado I(d)	50
4.2.3.4	Modelo auto-regressivo integrado de média móvel ARIMA(p,d,q)	51
5	ANÁLISE DE DADOS	53
5.1	Redução de dimensionalidade	53
5.1.1	Seleção de atributos	53
5.1.1.1	Critério baseado em variância	53
5.1.1.2	Critério baseado no acerto	54
5.1.2	Extração de atributos	54
5.1.2.1	Análise de componentes principais (PCA)	54
5.1.2.2	Isometric feature mapping (Isomap)	56
5.2	Classificação	58
5.2.1	Zero regras (Zero Rule)	59
5.2.2	Uma regra (One Rule)	60
5.2.3	J48 e C4.5	61
5.2.4	K-primeiros vizinhos (k-Nearest Neighbors)	63
5.2.5	Bayesiano ingênuo (Naive Bayes)	64
5.2.6	Redes de função de base radial (Radial Basis Function Networks)	65
5.3	Escalamento multi-dimensional (Multi-Dimensional Scaling)	68
6	DINÂMICA DE REDES	71
6.1	Construção de séries temporais	71
6.2	Análise de séries temporais	72
6.2.1	Não-linearidade e caos	72
6.2.2	Auto-correlação	73
6.2.3	Testes de valor-p	73
6.2.4	Modelo ARIMA	74
6.3	Medidas propostas	75
6.4	Resultados	76
6.4.1	Seleção de atributos	77
6.4.1.1	Baseada em variância	77
6.4.1.2	Baseada em acerto	78

6.4.2	Extração de atributos	80
6.4.2.1	PCA	80
6.4.2.2	Isomap	81
7	O PAPEL DAS PALAVRAS	85
7.1	Representação híbrida dos textos	85
7.2	Medidas de similaridade	86
7.3	União de matrizes de distâncias	88
7.4	Resultados	90
8	CONCLUSÕES	93
8.1	Conclusões gerais	93
8.2	Dinâmica de redes	93
8.3	O rol das palavras	94
8.4	Perspectivas para trabalhos futuros	94
8.5	Produção	95
	REFERÊNCIAS	97
	APÊNDICES	103
	APÊNDICE A – CORPORA DE TEXTOS	105
A.1	Primeiro corpus	105
A.2	Segundo corpus	108
A.3	Terceiro corpus	111
	APÊNDICE B – RESULTADOS DOS TESTES	115
	APÊNDICE C – AJUSTES COM O MODELO ARIMA	117
	APÊNDICE D – TAXAS DE ACERTO USANDO ISOMAP	121
	APÊNDICE E – OBTENÇÃO DAS RELAÇÕES PARA MDS	123
	ANEXOS	125
	ANEXO A – LISTA DE PART-OF-SPEECH TAGS USADOS NO NLTK	127
A.1	Part-of-Speech tags	127
	ANEXO B – ELEMENTOS REMOVIDOS	131

1 APRESENTAÇÃO

A linguagem falada é possivelmente a maior realização da humanidade e uma das suas maiores vantagens evolutivas. Hoje é usada para transmitir ideias, pensamentos, sentimentos, e às vezes, para distorcer as mesmas coisas. Isto é possível graças à sua grande riqueza, o que faz da linguagem um dos sistemas mais complexos para estudo na atualidade. Várias características permitem à linguagem ser de tão grande utilidade e facilidade de uso; de fato, praticamente todo ser humano aprende a se comunicar numa tenra idade.

Curiosamente, a linguagem não foi conscientemente pensada para possuir estas características. Pelo contrário, as linguagens atuais são obras conjuntas de todas as pessoas que as usaram, exceto por tentativas como a do esperanto. As teorias atuais sobre evolução da linguagem a modelam como produto de um sistema multiagente, onde o único objetivo de cada agente é se comunicar exitosamente. (1,2) Nesta linha de pesquisa também há estudos que apontam que várias línguas possam estar conectadas e possivelmente ter um ancestral comum. (3,4) O estudo da linguagem do ponto de vista estatístico começou no século XX. A descoberta da lei de potência na distribuição de frequência de aparição das palavras num texto é atribuída a Zipf, mas foi J. B. Estoup o primeiro a perceber este fenômeno em 1916. (5–7) Similarmente, em 1963 Benoit Mandelbrot apresentou a descoberta da lei de potência no tamanho das palavras. (7,8)

Um dos casos mais famosos dos aspectos práticos de análise computacional de textos é o dos Artigos Federalistas (9), que foi uma série de manifestos escritos anonimamente num jornal dos Estados Unidos entre 1787 e 1788 em favor da ratificação da nova constituição. Descobriu-se depois que os artigos tinham sido escritos por três autores. Dos 85 ensaios, 5 foram escritos por um autor, outro escreveu 14, outro escreveu 51 e três foram escritos conjuntamente por dois autores. Os 12 restantes, chamados “artigos sob disputa”, foram reivindicados por dois autores. A escola moderna dá unanimemente a autoria de todos os artigos sob disputa a um autor definido. Notavelmente, em 1964 Mosteller e Wallace apresentaram um método automático que chegava à mesma conclusão. (10, 11) Eles analisaram as distribuições de 30 palavras funcionais i.e. conjunções, preposições e artigos. Outro caso famoso é o de Don Foster, que afirmou ter encontrado um poema perdido de William Shakespeare, afirmação que posteriormente foi refutada por análise física do texto. (9)

1.1 Motivação

O uso de redes complexas para modelar a linguagem foi inspirado pela lei de Zipf e a aparição de modelo de Barabási e Albert (6, 12) que, quando usado como uma rede de

co-ocorrência, explicaria a lei perfeitamente. Porém, em (13) mostrou-se que a lei de Zipf teria duas leis de potência, contradizendo o modelo de Barabási. Este dilema foi resolvido por (14) que apresentaram um modelo de rede complexa que reproduzia a dupla lei de potência e mostrava como descrever as redes de co-ocorrência de palavras. Estes modelos tinham um parâmetro livre e portanto precisavam reproduzir os dados experimentais sem muitas possibilidades de ajuste. No entanto, surgiu a descoberta que as medidas de rede dependiam de diferentes fatores, como época, língua, estilo ou autor. (15–19) Ou seja, deve haver vários possíveis parâmetros ajustáveis, mas generalizações dos modelos de redes permitem explicar esse comportamento. Há, porém, uma importante questão a ser resolvida: mesmo ajustando os parâmetros existiria uma interdependência entre as métricas de rede, o que não foi observado explicitamente. (18,20)

As redes de co-ocorrência têm mostrado grande poder de descrição da linguagem, especificamente em problemas como o reconhecimento de autoria, mas ainda é possível estender este conhecimento. Esta é a motivação do presente trabalho: entender a complexidade da linguagem a partir das redes de co-ocorrência. (7,21,22)

1.2 Objetivos

Estudar diferentes aspectos das redes de co-ocorrência e sua viabilidade como atributos para o reconhecimento da autoria de textos literários.

- Definir e estudar as propriedades dinâmicas das redes de co-ocorrência de palavras a partir de textos.
- Definir e estudar medidas de similaridade entre textos que levem em conta a identidade das palavras que conformam a rede.

O documento compõe-se de três partes: os cinco primeiros capítulos dedicam-se à base teórica; os capítulos 6 e 7 mostram os resultados obtidos; e o capítulo 8 apresenta as conclusões do trabalho.

2 LINGUAGEM

O objeto do estudo deste trabalho são os textos literários, portanto a continuação apresentamos alguns conceitos e definições necessárias, assim como algumas técnicas que foram usadas para preparar os textos. A grande vantagem que têm algumas abordagens incluindo a nossa é que as técnicas que tratam as particularidades e sutilezas da linguagem já foram implementadas em detalhe e disponibilizadas por especialistas na matéria (i.e., linguistas, lexicólogos, etc.) portanto somente precisamos implementar as ferramentas adequadas.

2.1 Áreas da Linguística

O estudo da linguagem abrange várias escalas, desde sons e sílabas até frases inteiras e ainda escalas maiores. Listamos a continuação algumas das áreas mais conhecidas desta ciência. (23)

2.1.1 Fonologia

A fonologia trata com os sons e a sua relação com a linguagem: como eles podem ser classificados e associados com os símbolos escritos ou caracteres; a unidade básica estudada pela fonologia é chamada de fonema. Uma área relacionada é a fonética que estuda os aspectos físicos da emissão de sons.

2.1.2 Morfologia

As regras que governam a construção de palavras são estudadas pela morfologia, tanto na forma grafêmica (i.e., escrita) quanto na fonêmica (i.e., falada). Os blocos que compõem uma palavra são conhecidos como morfemas e estão definidos como a mínima unidade na linguagem que pode ser associada a um significado ou alternativamente como a mínima unidade de análise gramática. Um exemplo simples é o morfema -s que, quando colocado no final de uma palavra, usualmente denota o plural desta.

2.1.3 Lexicografia

De especial interesse para áreas de análise computacional tais como o processamento de línguas naturais e recuperação de informação, a lexicografia é a técnica utilizada para compilar um léxico ou dicionário. Dependendo do léxico, outras características como inflexões, relações semânticas e hierarquias também podem ser tratadas.

2.1.4 Semântica

A semântica é o estudo do significado. Ela estuda o conteúdo semântico (a mensagem) dos diferentes elementos da linguagem como palavras, frases ou expressão corporal. Por exemplo as palavras funcionais como “de” têm pouco conteúdo semântico já que não expressam muito individualmente.

2.1.5 Sintaxe

A sintaxe estuda as regras da organização das palavras numa frase de acordo ao tipo de cada uma. Por exemplo, no português ao igual que na grande maioria das línguas uma frase declarativa segue a sequência Sujeito-Verbo-Objeto. (4)

2.2 Análise automática de textos

À parte da representação em forma de redes, que é o tema central deste manuscrito, têm existido várias propostas para a análise automática dos textos. A análise automática tem virado uma ferramenta mais necessária a cada dia especialmente com o grande volume de texto gerado atualmente e a sua disponibilização na Internet e por isto centenas de abordagens têm sido propostas. (24–26) Podemos separar estas técnicas em três grandes grupos:

1. *Análise de elementos lexicais.* como ficou evidente na seção anterior, a linguagem possui características lexicais a muitos níveis, estudam-se entre outros: palavras, n -gramas, palavras funcionais, classes de palavras, expressões conectoras de discursos, gírias, contrações, dialetos, erros de ortografia, nomes próprios, características semânticas (polissemia), n -gramas de caracteres, sufixos frequentes, pontuação, etc.
2. *Formato do texto.* comprimento de linhas, palavras e frases, formatação (espaços em branco), capitalização, caracteres não alfanuméricos, começos e fins de textos.
3. *Caraterísticas sintáticas.* função sintática de n -gramas, tipos de frase, perplexidade e complexidade morfológica.

O segundo grupo é usado principalmente para análise de editoração ou análise de texto não-literário (e.g. código fonte) (9), enquanto o terceiro acostuma precisar de conhecimentos específicos em linguística. Vamos dar uns exemplos do primeiro grupo usados para o reconhecimento de autoria. São estudados elementos lexicais de baixo nível como os caracteres (27, 28), os fonemas (29), e os morfemas. (30, 31) Existem outros elementos que podem ser estudados como os n -gramas: grupos pequenos de elementos que co-ocorrem em um texto (32, 33), ou os papéis sintáticos das palavras nas frases. (34, 35)

Mas o arquétipo destes métodos é chamado Bag of Words (BoW). (36) O BoW representa um texto como o conjunto das palavras que este contém e a cada palavra é associada a frequência com que ela aparece. A similaridade entre dois textos é calculada como o produto interno dos vetores onde cada dimensão é uma palavra e a componente do vetor nessa dimensão é a frequência. Dado que usualmente um texto de maior tamanho possui um vocabulário maior existe a distorção de considerar que os maiores textos são mais similares a quaisquer outros. Para corrigir isto foi criada a variante chamada Frequência de Termo-Frequência Inversa de Documento, do Inglês *Term Frequency-Inverse Document Frequency* (TF-IDF). (37, 38) No TF-IDF a frequência do termo (TF) é dividida pela frequência inversa nos documentos da coleção considerada (IDF) que é calculada como sendo o logaritmo do quociente entre o número total de documentos e o número de documentos que contém o termo. Desta forma, palavras que são em geral frequentes têm a sua frequência reduzida em maior grau. Há variações do BoW e do TF-IDF, em particular a chamada Bag of Features é amplamente usada, por exemplo para reconhecimento de imagens (39) onde particularidades das imagens ocupam o papel das palavras.

2.3 Pré-processamento

Particularidades da língua ou dialeto, o estilo da escrita do autor, a época, ou até a edição de um texto criaram a necessidade de formatar ou pré-processar um texto antes da sua análise. Mas o motivo mais importante para realizar um pré-processamento é a existência das inflexões de uma palavra que estão presentes na maioria das línguas. Uma palavra pode ser modificada ou inflexionada para expressar diferentes categorias gramaticais tais como tempo, caso, voz, aspecto, número, gênero, e humor que podem ou não existir dependendo da língua e da categoria da palavra.

Existem diversos motivos para pré-processar um texto, por exemplo na tarefa de busca automática de uma palavra os documentos que possuam formas inflexionadas podem ser relevantes. Mesma situação acontece nas tarefas de sumarização e extração de palavras. Para o presente trabalho existem duas motivações principais: procuramos relações semânticas de curto e longo alcance, e o objeto de trabalho (as redes de co-ocorrência) são inerentemente esparsas e portanto é desejável condensar a informação antes de tratá-la. Existem também motivos para não aplicar certos métodos de pré-processamento, por exemplo alguns autores propõem métodos de reconhecimento de autoria baseados em características tais como o uso de palavras funcionais como mencionado acima e portanto a remoção destas palavras é inteiramente contraproducente para eles. Vamos mencionar umas poucas tarefas de pré-processamento relevantes ao presente trabalho.

2.3.1 Stemming

Cada palavra tem associada uma raiz ou *stem*, da qual foi derivada, usualmente adicionando um sufixo à semente. O stemming é aplicado sobre as palavras num processo chamado de conflagação (*conflation*) que é amplamente usado por algoritmos de busca com o fim de capturar as palavras que também foram derivadas da mesma semente. Os algoritmos de stemming tentam inferir a semente usando uma série de regras quando aplicáveis à palavra. A semente não é a unidade mínima com significado de uma palavra, e uma semente pode na sua vez provir de outra semente. Se bem o stemming pode se considerar uma abordagem de força bruta, muitas vezes é igual ou mais eficiente para juntar palavras relacionadas. (38)

2.3.2 Lematização

Diferentes formas de uma palavra tem a mesma semente e uma palavra pode estar associada a mais de uma semente. Por exemplo, a palavra em Inglês *book* pode se referir a um substantivo (livro) ou um verbo (reservar, registrar, marcar, etc.) o que se conhece como polissemia. Neste caso o stemming tradicional pode retornar a semente errada o que pode prejudicar a caracterização semântica de um texto. A lematização é um processo que alguns consideram uma classe particular de stemming na qual se procura determinar a semente mais adequada usando uma abordagem chamada tagging.

2.3.2.1 Part of speech tagging

É necessário analisar o contexto no qual se encontra uma palavra para escolher corretamente a semente. A cada palavra se associa um elemento chamado tag que define sua classe. A forma mais comum de fazer isto é mediante o chamado *part of speech tagging* ou tagging por parte do discurso. Uma parte do discurso é uma categoria de palavras com propriedades gramaticais similares, por exemplo substantivos, verbos, adjetivos, advérbios, pronomes, preposições, conjunções, interjeições, numerais, artigos, ou determinantes. No entanto, à medida que a complexidade de uma frase aumenta podem aparecer tags mais específicas, no anexo A listam-se os tags usados pelo pacote NLTK. Existem várias propostas sobre a escolha do tipo de tag a ser associado a uma palavra (40):

- *N*-gramas. (41)
- Modelos ocultos de Markov.
- Modelos de linguagem simbólicos derivados usando aprendizado de máquina.
- Regras globais manuais.
- Regras locais manuais.

- Híbridos.

Neste trabalho usamos part of speech tagging baseado na maximização da entropia.
(42)

2.3.3 Remoção de elementos

As palavras funcionais como os artigos, preposições ou conjunções realizam uma importante tarefa servindo de suporte na construção das frases, e por isso é fácil encontrá-las em qualquer lugar, no entanto elas tem pouco conteúdo semântico. Estes dois fatores fazem imprescindível a sua remoção dos textos. Por exemplo, os primeiros lugares na lei de Zipf estão dominados por palavras funcionais. Similarmente os símbolos aparecem com muita frequência sem aportar maior conteúdo semântico. No anexo B listam-se as palavras e símbolos que foram removidos dos textos na etapa de pré-processamento.

3 REDES COMPLEXAS

Uma rede ou grafo é um conjunto N de nós ligados por um conjunto E de arestas com uma distribuição dada. Os nós podem representar objetos, e as arestas relações entre os objetos. Leonard Euler começou o estudo de redes resolvendo um problema conhecido como as pontes de Königsberg onde se perguntava se sete pontes dessa cidade poderiam ser percorridas num caminho fechado passando por cada ponte só uma vez. Euler construiu uma rede e percebeu que, se bem este era um problema da geometria, era possível fazer transformações sem afetar a rede, resolvendo assim o problema e criando um novo ramo da matemática chamado topologia.

Muitos pesquisadores continuaram a estudar as propriedades topológicas de redes pequenas, mas uma segunda revolução surgiu com o trabalho de Paul Erdős e Alfréd Rényi (43), que estudaram não uma rede mas um conjunto de redes criadas a partir de uma mesma regra com uma parte estocástica conhecido como ensemble. No modelo de Erdős-Rényi as arestas são sorteadas entre pares de vértices escolhidos aleatoriamente. Desde então uma grande parte dos estudos sobre redes dedica-se às suas propriedades estatísticas, tanto em modelos como em sistemas reais. Um exemplo é a rede de mundo pequeno de Duncan J. Watts e Steven Strogatz. (44) O efeito de mundo pequeno (conhecido pelos famosos seis graus de separação) mostrou como umas poucas conexões de longo alcance podem diminuir consideravelmente a distância entre os nós.

3.1 Complexidade

Uma rede que merece especial atenção é a rede complexa de Albert-László Barabási e Réka Albert. (12) A complexidade é um dos paradigmas mais importantes da física das últimas décadas e consiste na emergência de fenômenos macroscópicos a partir de interações de curto alcance. (45,46) Em muitos sistemas isto se manifesta na aparição de distribuições livres de escala e em particular leis de potência. No modelo de Barabasi-Albert, a rede é construída adicionando um nó por vez conectado a um número fixo de novas arestas. O outro extremo destas arestas é escolhido segundo um critério de ligação preferencial: cada nó existente tem uma probabilidade de ser escolhido proporcional ao seu número atual de arestas. Este fenômeno é conhecido como *rich gets richer* (o rico fica mais rico), e é comum a vários sistemas complexos. Do mecanismo de crescimento e ligação preferencial originam-se distribuições em forma de leis de potência i.e. aquelas onde o número de elementos (nós) com um valor dado de uma métrica (usualmente a conectividade) é proporcional ao valor da métrica elevado a uma potência negativa. Algumas das redes do mundo real que se acredita seguir o modelo de Barabási-Albert são:

- Interação proteica na célula, por exemplo entre as proteínas da *saccharomyces cerevisiae*. (47)
- Regulação de expressão de genes.
- Rede de neurônios do cérebro.
- Rede funcional do cérebro.
- Voos comerciais.
- Elementos físicos conectados à Internet.
- World Wide Web (rede de sites).
- Colaborações científicas.
- Amizades.
- Co-stardom: rede de colaborações de atores.
- Correlações entre preços de ações do mercado.
- Relações sintáticas entre palavras.

Existem variações do modelo e modelos similares para reproduzir distribuições de longa cauda, isto é, aquelas que têm distribuições que, se não necessariamente têm forma de leis de potência, abrangem uma longa faixa de valores.

3.2 Métricas de rede

Muitas métricas de rede têm sido e continuam sendo propostas na literatura. (48) As métricas podem ser globais (um número para toda a rede), ou locais, tanto para cada nó como para pares de nós (e.g., menor caminho entre dois nós) ou mesmo para grupos de nós. Cada métrica define uma propriedade diferente da rede como se ilustra na figura 1. Na figura mostram-se os nós de uma rede coloridos segundo o valor de uma medida para seis medidas diferentes locais aos nós. A figura evidencia que o valor de uma medida de rede não necessariamente tem relação com os valores das outras medidas e que a importância de um nó na rede depende da medida em consideração.

Já temos definidas duas medidas globais de rede: N e E , mas a representação de rede mais importante e geral é a matriz de adjacência A . (46) Esta matriz é de tamanho $N \times N$ e o elemento A_{ij} é igual a um se existe uma aresta do i -ésimo ao j -ésimo nó e igual a zero em caso contrário. Do ponto de vista da topologia, a matriz de adjacência define completamente a rede, no entanto é de difícil interpretação e mais importante, torna-se computacionalmente ineficiente para ser usada com redes medianas ou grandes. Além de N e E , neste trabalho foram usadas as seguintes medidas de rede:

Figura 1 – Nós de uma rede coloridos segundo diferentes centralidades: a, intermediação; b, proximidade; c, autovetor; d, grau; e, harmônico; f, Katz.

Fonte: Adaptada de COMPRESSIVE... (49)

3.2.1 Peso

O peso é uma métrica que pode ser atribuída em certos casos especiais. Quando uma rede é construída pode acontecer que se crie uma nova aresta entre dois nós entre os quais já há uma aresta. Neste caso há várias alternativas: criar mais uma aresta entre os nós, descartar a nova aresta, recalculá-la (caso seja uma aresta aleatória) ou dar um peso a cada aresta, sendo este inicialmente igual à unidade e aumentando uma unidade com cada novo enlace. Neste trabalho optamos pela última opção. O peso também é usado nas redes que têm um certo fluxo (como redes de transporte ou de servidores de Internet) para medir a congestão de uma linha ou aresta.

3.2.2 Conectividade

A conectividade ou grau é uma métrica local de cada nó igual ao seu número de vizinhos (arestas). Para redes não-direcionadas esta pode ser expressa em termos da matriz de adjacência por:

$$k_i = [A^2]_{ii}. \quad (3.1)$$

Isto é evidente do fato que neste caso a matriz A é simétrica e o termo à direita

em 3.1 é a norma de um vetor composto por zeros e um número de uns igual ao número de vizinhos do i -ésimo nó. A conectividade é muito usada por sua fácil interpretação e relação com o conceito de complexidade. Outra vantagem é a rapidez de cálculo.

3.2.3 Menor caminho

O (comprimento do) menor caminho é uma métrica local que depende de dois nós e expressa o número mínimo de arestas que deve ser percorrido para chegar de um nó a outro. o número $l_{ij}^{(c)}$ de caminhos de comprimento c entre o i -ésimo e o j -ésimo nó vem dado por:

$$l_{ij}^{(c)} = [A^c]_{ij}. \quad (3.2)$$

Considere o caso dos caminhos de comprimento dois: se $A_{ik} = A_{kj} = 1$ existe uma aresta entre os nós i e k , e entre os nós k e j , o que contribui uma unidade ao valor de $l_{ij}^{(2)*}$. Analogamente, $[A^2]_{kj}$ contém o número de caminhos de comprimento dois entre k e j que vão contribuir ao número de caminhos de comprimento três entre i e j caso $A_{ik} = 1$ demonstrando a relação 3.3 para $c = 3$ e para comprimentos maiores por indução. Definimos portanto o (comprimento do) menor caminho como:

$$c_{ij}^{(m)} = \min\{c \in \mathbb{N} \mid l_{ij}^{(c)} \neq 0\}. \quad (3.3)$$

Ele é importante nas redes de mundo pequeno onde geralmente é muito menor que o número de arestas E . No entanto, é uma das métricas computacionalmente mais exigentes sendo a sua complexidade de $O(E^2)$ caso todos os menores caminhos de um nó sejam calculados usando o algoritmo de Dijkstra, e $O(E^3)$ caso todos os menores caminhos sejam calculados com o algoritmo de Floyd–Warshall.

3.2.4 Centralidade de intermediação

A centralidade de intermediação é uma métrica local a um nó e uma medida da proporção de menores caminhos que passam pelo nó, definida como:

$$B_i = \sum_{s \neq i \neq t} \frac{c_{st}^{(m)i}}{c_{st}^{(m)}}, \quad (3.4)$$

onde o super-índice i no numerador denota um menor caminho que passa pelo nó em consideração. Uma vez que a expressão 3.4 precisa do cálculo de todos os menores caminhos, a complexidade computacional é $O(E^3)$.

* Em particular, os elementos da diagonal de A correspondem ao número de possíveis maneiras de ir e voltar ao mesmo nó em dois passos i.e. a conectividade 3.1.

3.2.5 Centralidade de carga

A centralidade de carga L_i é definida analogamente à centralidade de intermediação considerando-se também o peso nas arestas. Neste caso se substituem os comprimentos l da expressão 3.4 pela soma dos pesos nas arestas que compõem o caminho. Por exemplo, se o caminho está composto por três arestas, dois com peso um e um com peso dois o correspondente caminho é $c_{ij}^{(m)} = 3$ enquanto o caminho com peso é $\hat{l}_{ij}^{(m)} = 4$. Esta métrica busca dar maior importância aos nós relacionados a arestas com alto peso.

3.2.6 Intermitência

A intermitência não foi criada dentro do contexto de redes, mas nos sistemas onde existe uma série de elementos empilhados em certa ordem que podem se repetir, como é o caso dos textos. Neste caso ela mede quanto as sucessivas aparições do elemento são irregulares. A intermitência define-se como

$$I_i = \frac{\text{var}(\Delta)}{\bar{\Delta}}, \quad (3.5)$$

onde Δ é a distância (em número de elementos) entre duas aparições sucessivas. Caso o elemento apareça só uma vez Δ é o tamanho da série e a intermitência é zero. No contexto das redes a intermitência está associada com a variação da largura dos ciclos (caminhos fechados) sobre um nó.

3.2.7 Coeficiente de clusterização

O coeficiente de clusterização é uma métrica local de cada nó que mostra a coesão entre os seus vizinhos próximos, sendo definido por:

$$c_i = \frac{e_i}{k_i(k_i - 1)}, \quad (3.6)$$

onde k_i é a conectividade do nó e e_i é o número de conexões entre os seus vizinhos. O denominador é o número máximo de possíveis conexões e portanto o coeficiente de clusterização encontra-se na faixa entre zero e um. O coeficiente de clusterização foi usado por Watts e Strogatz para mostrar que é possível ter redes com alta clusterização e baixo menor caminho (como as redes sociais). No entanto, a maioria das redes medianas e grandes tem uma clusterização baixa, excetuando as redes regulares ou grades.

3.2.8 Transitividade

A transitividade é uma métrica global da rede que mede quanto a conexão entre nós é transitiva, isto é, se o nó i tem conexão com o nó j e o j com o k , qual é a possibilidade

de ter uma conexão entre o nó i e o k . Ela é relevante principalmente no estudo das redes sociais e está dada pela expressão:

$$T = 3 \frac{\text{Triângulos}}{\text{Triadas}}, \quad (3.7)$$

onde um triângulo são três nós completamente conectados e uma tríade são três nós com no mínimo duas arestas entre eles. O fator três normaliza a transitividade entre zero e um.

3.2.9 Número de cliques

O número de cliques é uma métrica global. Um clique é um subgrafo completo, isto é, cujos nós estão todos mutuamente conectados. Um clique tem uma ordem que é o número de nós do subgrafo. Neste trabalho consideram-se cliques de todas as ordens.

3.2.10 Diâmetro

O diâmetro de uma rede é uma métrica global definida por:

$$D = \max\{c_{ij}^{(m)}\}, \quad (3.8)$$

isto é, a maior das menores distâncias. Em outras palavras, a menor distância entre os dois nós mais afastados da rede.

3.2.11 Raio

O raio de uma rede é uma métrica global definida como o mínimo sobre todos os nós da máxima distância do nó a qualquer outro nó da rede. Note-se que a partir de uma métrica local a um ou vários nós, sempre é possível obter uma métrica global promediando sobre todos os nós.

3.3 Redes de Co-ocorrência

A representação de textos nesta tese é feita através de redes de co-ocorrência. Nesta seção, discutimos brevemente a relação entre linguagem, redes e complexidade. A relação entre linguagem e complexidade é de longa data e precede a abordagem estatística de redes apresentada no começo do capítulo. Em 1935, George Kingsley Zipf apresentou um estudo sobre a frequência de aparição das palavras em um texto. Zipf contou a frequência f de aparição das palavras de um texto e organizou as palavras de acordo com esta frequência atribuindo um ranking r de um à palavra mais frequente, dois à seguinte e assim sucessivamente. Ele descobriu que a palavra mais frequente aparecia o dobro de

vezes que a segunda e a relação se repetia entre uma palavra e a seguinte rankeada, ou segundo a relação,

$$f \sim 1/r, \quad (3.9)$$

isto é, uma lei de potência com expoente $1^\dagger$. A lei de Zipf tem recebido críticas e correções já que não é satisfeita em todos os casos ou pelo menos não exatamente. A figura 2 mostra a lei de Zipf para várias línguas, evidenciando que existem diferentes expoentes assim como algumas irregularidades. Em 2001 Ramon Ferrer i Cancho e Ricard V. Solé descobriram um aspecto desconhecido da lei de Zipf. (13) Como se pode ver na figura 3, há dois e não um regime de potência na distribuição. Ferrer i Cancho e Solé apontaram que originalmente o Zipf não considerou as palavras com maior ranking (lado direito da figura 3) e somente concentrou-se nas palavras mais frequentes (lado esquerdo da figura 3), em cujo caso o expoente encontrado foi quase o mesmo ($\alpha = 1.01$).

Figura 2 – Lei de Zipf para varias línguas.

Fonte: BIEMANN; QUASTHOFF. (50)

Se existir uma faixa que satisfaz uma relação do tipo $f \sim r^{-\alpha}$ (segundo Zipf $\alpha = 1$) a distribuição correspondente das frequências das palavras é dada por:

$$P(n) \sim n^{-\beta}, \quad \beta = \frac{1}{\alpha} + 1, \quad (3.10)$$

que é apresentada na figura 4. Note-se que neste caso a diferença entre os expoentes é menor. A explicação deste comportamento foi dada por Sergey N. Dorogovtsev e José Fernando Ferreira Mendes no mesmo ano, utilizando redes de co-ocorrência. (14)

[†] É uma prática comum considerar só a parte positiva do expoente, portanto neste caso teríamos $\alpha = 1$

Figura 3 – Lei de Zipf estendida.

Fonte: FERRER-I-CANCHO; SOLÉ. (13)

Figura 4 – Probabilidade das palavras.

Fonte: FERRER-I-CANCHO; SOLÉ. (13)

Uma rede de co-ocorrência se constrói a partir de um texto considerando a posição relativa das palavras. Nestas redes, as palavras, ou melhor, as raízes das palavras depois do pré-processamento comentado na seção 2.3, fazem o papel de nós. Se duas palavras encontram-se consecutivamente no texto existirá uma aresta saindo do nó da primeira palavra para o nó da segunda. Considere por exemplo o trecho inicial da novela “*A Tale of Two Cities*” de Charles Dickens: “*It was the best of times; It was the worst of times; It was the age of wisdom; It was the age of foolishness...*”. Pré-processando o texto e construindo a rede de co-ocorrência correspondente teríamos a rede da figura 5.

Figura 5 – Rede de co-ocorrência.

Fonte: Elaborada pelo autor.

Nestas redes o peso nas arestas de entrada é igual ao peso nas arestas de saída,

o que é válido para todos os nós exceto o da primeira e o da última palavra. Observe-se que o peso total nas arestas de entrada (ou de saída) é igual ao número de vezes que a palavra apareceu no texto, ou em outras palavras, à sua frequência. Portanto, se todas as arestas têm peso um, podemos relacionar a distribuição das conectividades à lei de Zipf. Dorogovtsev e Mendes criaram um modelo para explicar a dupla lei de potência da linguagem argumentando que a co-ocorrência de palavras é governada pelo mecanismo de ligação preferencial. O modelo consegue reproduzir a dupla lei de potência e explicar os dados experimentais sem parâmetros ajustáveis com uma modificação muito simples ao mecanismo de Barabási-Albert: a cada momento que um novo nó se liga preferencialmente a outro na rede, ct novas conexões também se criam dentro da rede, onde c é uma constante e t é o tamanho atual da rede ou “tempo”.

3.4 Métricas para uma rede de co-ocorrência

As distribuições das métricas de rede locais de uma rede de co-ocorrência típica são apresentadas nesta seção para dar uma ideia do tipo de perfil a esperar. No caso do menor caminho, que é uma métrica local dependente de dois nós, calculamos a média sobre um dos índices de $c_{ij}^{(m)}$. Algumas métricas são diferentes dependendo de se a rede é ou não-direcionada. Por exemplo a conectividade de um nó numa rede de co-ocorrência direcionada é a metade do caso de uma não-direcionada, com exceção dos nós correspondentes à primeira e à última palavra do texto. Devido a que as redes de co-ocorrência são esparsas escolhemos converter a rede numa rede não-direcionada quando extrairmos certas métricas. As métricas calculadas usando uma rede não direcionada são: a conectividade, a centralidade de intermediação, a centralidade de carga e o menor caminho. Para gerar as distribuições usamos a rede de co-ocorrência do maior livro da coleção: “War and Peace” do Leo Tolstoi. O número de bins está indicado em cada figura e as distribuições estão representadas por pontos para maior clareza[‡]. Algumas características gerais podem ser inferidas: a conectividade e as centralidades de intermediação e de carga seguem leis de potência decrescentes; a intermitência segue uma lei de potência crescente com uma pequena explosão nos valores menores; o menor caminho, mesmo abrangendo somente uma faixa de uma década dos valores tem duas regiões com leis de potência, uma seguindo uma lei crescente e a outra seguindo uma lei decrescente com um máximo bem definido em $l = 3$; o coeficiente de clusterização possui uma distribuição em forma de sino centrada em valores baixos da clusterização e que possui muito ruído.

[‡] Os bins são cada uma das partes em que está dividido um histograma.

Figura 6 – Conetividade (200 bins).

Fonte: Elaborada pelo autor.

Figura 7 – Menor caminho (300 bins).

Fonte: Elaborada pelo autor.

Figura 8 – Centralidade de intermediação (200 bins).

Fonte: Elaborada pelo autor.

Figura 9 – Centralidade de carga (400 bins).

Fonte: Elaborada pelo autor.

Figura 10 – Coeficiente de clusterização (800 bins).

Fonte: Elaborada pelo autor.

Figura 11 – Intermitência (1000 bins).

Fonte: Elaborada pelo autor.

4 SÉRIES TEMPORAIS

Um dos objetivos do trabalho é estudar a dinâmica das redes de co-ocorrência. Existem duas formas de estudar a dinâmica no contexto das redes. Numa linha de pesquisa estuda-se a possível dinâmica que ocorre no interior de uma rede. Por exemplo, é de interesse em neurociência entender o padrão de disparos dos neurônios, que pode ser descrito usando uma dinâmica contínua como no modelo de Hodgkins e Huxley (51) ou com uma dinâmica discreta usando autômatos celulares como no modelo de Kinouchi e Copelli. (52) Outras aplicações da dinâmica dentro da rede incluem a propagação de doenças, o fluxo de informação na WWW e rumores em redes sociais, e o estudo do fenômeno de sincronização usando o modelo de Kuramoto. (53) Infelizmente, não parece haver um processo que naturalmente possa ser modelado como uma dinâmica sobre uma rede de co-ocorrência aparte da própria estrutura. A outra grande linha de pesquisa consiste em estudar a dinâmica da topologia da rede, como mencionado no capítulo anterior. (12,14) Neste caso, a ideia da dinâmica da rede de co-ocorrência surge naturalmente levando em conta que esta descreve um texto que tem uma direção bem definida: a da leitura. Vamos associar a direção na qual transcorre a leitura de um livro ao conceito de tempo. Diferentemente das abordagens anteriores, para estudar a dinâmica usamos a teoria das séries temporais; isto permite analisar cada parte do texto ao invés de somente observar a estrutura final. Outra vantagem do uso de séries temporais em relação à atribuição de autoria é poder comparar facilmente textos de tamanhos diferentes.

As séries temporais são definidas como sistemas com uma dinâmica em tempo discreto, também são conhecidas como mapas. (54) Seu estudo, e principalmente predição, é de grande interesse em áreas como finanças e meteorologia. O passo temporal pode ser variável, como no caso das mudanças no preço de uma ação ou dos disparos de um neurônio. Por construção as séries temporais de nosso interesse são de passo temporal fixo. A variável de interesse pode também ser discreta ou contínua; aqui consideramos variáveis contínuas ainda que algumas métricas de rede possam gerar séries de variável discreta.

4.1 Não-linearidade e caos

A dinâmica de um sistema em geral pode ser descrita em termos de dois tipos de processos, uma parte determinística e uma parte estocástica. A parte determinística se refere ao que pode ser descrito de forma exata por meio de leis e certas condições iniciais, enquanto a parte estocástica pode predizer, no máximo, a probabilidade de ocorrência. O determinismo tem sido um ponto de debate desde a época de Descartes e Newton, pois em princípio seria possível conhecer a evolução de qualquer sistema a partir das condições iniciais. O estudo dos sistemas com dinâmicas não lineares trouxe nova

perspectiva com a teoria do caos. Mostrou-se que mesmo com regras conhecidas muito simples, a evolução de um sistema pode ser extremamente sensível às condições iniciais e criar padrões complicados, no que se denomina regime caótico. (54) Ao contrário, a indeterminação é inerente à dinâmica estocástica, mesmo que toda a informação possível esteja disponível.

É importante portanto descobrir se as nossas séries temporais são geradas por uma dinâmica determinística não-linear, onde o valor de um elemento da série é dado pela história da série ou se, ao contrário, os valores são independentes. Nesta seção apresentaremos um método simples para detectar dinâmicas não-lineares, chamado retrato de fase. Para séries temporais, um retrato de fase mostra um conjunto de pontos onde o valor da abscissa de cada ponto consiste no valor da série num determinado instante e o valor da ordenada é o valor da série no instante seguinte. (55) Existe também uma definição de retrato de fase para sistemas com dinâmica contínua. (54) É importante destacar que os retratos de fase não descartam a possibilidade de ter uma dinâmica determinística, mas servem para detectar elementos típicos deste tipo de situação: os atratores e os ciclos limite. Um exemplo simples de sistema com atratores é o mapa logístico cuja dinâmica é descrita pela relação de recorrência:

$$x_{t+1} = rx_t(1 - x_t). \quad (4.1)$$

Dependendo do valor do parâmetro r , há diferentes regimes. Por exemplo, para $1 \leq r \leq 3$ a série converge ao valor $(r - 1)/r$ para qualquer condição inicial; para $3 \leq r \leq 1 + \sqrt{6}$ a série oscila entre dois valores (os atratores) gerando dois pontos simetricamente posicionados em relação à diagonal no retrato de fase; e para valores maiores o comportamento geral da série varia entre o regime caótico e regimes com um número de atratores que dobra à medida que aumenta r .

Na figura 12 é mostrado o retrato de fase do mapa logístico com o parâmetro $r = 3.4$ que corresponde ao regime com dois atratores com valores 0.45196 e 0.84215, e que podem ser claramente identificados pelo traço dos pontos ocupados enquanto o sistema convergia. Dado que o valor da coordenada vertical depende da horizontal segundo a simples relação 4.1, os pontos estão confinados a uma parábola invertida. No regime caótico os pontos estão distribuídos sobre a parábola. Na figura 13 há um retrato de fase obtido de uma série temporal no qual não se observam atratores ou pontos limite. Além disso, os pontos não estão espalhados uniformemente, mas encontram-se preferencialmente numa região, que é um comportamento típico de uma dinâmica estocástica. O contraste com a distribuição dos pontos da figura 12 é ainda mais notório considerando-se que para o mapa logístico se usaram 1000 iterações enquanto esta série somente possui 252 elementos.

Figura 12 – Retrato de fase do mapa logístico com $r = 3.4$.

Fonte: Elaborada pelo autor.

Figura 13 – Retrato de fase de uma série estocástica.

Fonte: Elaborada pelo autor.

4.2 Estacionariedade

Um dos conceitos mais importantes na análise das séries temporais é a estacionariedade e curiosamente não existe consenso na sua definição. Existem duas definições denominadas de forma fraca e forma forte. Na forma fraca um processo é estacionário se o valor esperado da média é constante no tempo. Porém, segundo alguns autores o valor da variância também deve ser constante, sendo que esta segunda condição está relacionada a uma correlação baixa (ver seção 4.2.1). Por outro lado, a forma forte de estacionariedade impõe que todos os valores esperados sejam constantes no tempo; na prática a forma forte é raramente usada e dependendo da aplicação se tem um compromisso do grau de estacionariedade requerida, mais ainda quando se trabalha com séries finitas. Nesta seção apresentaremos as técnicas para avaliar a existência de estacionariedade.

4.2.1 Auto-correlação

A correlação entre duas séries temporais é um número entre 1 e -1 que mede o grau em que as séries se comportam do mesmo modo. Uma alta correlação pode então significar que existe algum mecanismo que relacione as séries. Ressalte-se que uma crítica comum à utilização desta medida é que a correlação não implica causalidade e pode ser um caso fortuito. Existem várias definições de correlação dependendo do contexto e dos objetos de análise (e.g. séries, funções, operadores, etc.). Para séries temporais costuma-se

usar o coeficiente de correlação de Pearson:

$$Cor(x, y) = \frac{\sum_{i=1}^T (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^T (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^T (y_i - \bar{y})^2}}. \quad (4.2)$$

Podemos definir a autocorrelação $ACF(\tau)$ como a correlação dos valores de uma série com os valores dela deslocada um número τ de posições (por definição $ACF(0) = 1$):

$$ACF(\tau) = Cor(x^*(t), x(t + \tau)) \quad (4.3)$$

onde o asterisco indica o complexo conjugado da função. A autocorrelação é interpretada como a relação entre os valores presentes e passados de uma série. Uma alta autocorrelação (positiva ou negativa) significa que os valores não são independentes. Uma série com baixa autocorrelação é apresentada na figura 14.

Figura 14 – Autocorrelação considerada baixa. As linhas pontilhadas marcam os níveis $ACF = \pm 0.05$.

Fonte: Elaborada pelo autor.

As linhas pontilhadas marcam os níveis $ACF = \pm 0.05$. Observam-se alguns valores que ultrapassam estes limiares, mas como já comentado, ocorrem casos nos quais isto acontece por mera coincidência. Outra situação ocorre para valores baixos de τ onde a densidade de valores altos é maior. Isto acontece com a maioria das séries já que é necessário um deslocamento mínimo, não muito grande, para des-correlacionar a série.

4.2.1.1 Teorema de Wiener-Khinchin

Existe ainda uma relação direta entre a autocorrelação de uma série e sua expansão em séries de Fourier, conhecida como o teorema de Wiener-Khinchin. O teorema é facilmente

dedutível da seguinte forma. Partimos das definições de autocorrelação (equação 4.3) e da expansão em séries de Fourier:

$$x(t) = \sum_{\mu=1}^T x_{\mu} e^{-2\pi i \mu t}, \quad (4.4)$$

com conjugada complexa,

$$x^*(t) = \sum_{\mu=1}^T x_{\mu}^* e^{2\pi i \mu t}. \quad (4.5)$$

Substituindo 4.5 e 4.4 na definição de autocorrelação temos:

$$\begin{aligned} ACF(\tau) &= A^{-1} \sum_{t=1}^T \sum_{\nu=1}^T x_{\nu}^* e^{2\pi i \nu t} \sum_{\nu'=1}^T x_{\nu'} e^{-2\pi i \nu' (t+\tau)} \\ &= A^{-1} \sum_{t=1}^T \sum_{\nu=1}^T \sum_{\nu'=1}^T x_{\nu}^* x_{\nu'} e^{2\pi i (\nu - \nu') t} e^{-2\pi i \nu' \tau} \\ &= A^{-1} \sum_{\nu=1}^T \sum_{\nu'=1}^T x_{\nu}^* x_{\nu'} \delta(\nu - \nu') e^{-2\pi i \nu' \tau} \\ &= A^{-1} \sum_{\nu=1}^T x_{\nu}^* x_{\nu} e^{-2\pi i \nu \tau} \\ &= \mathcal{F}_{\nu} [|x_{\nu}|^2] (\tau), \end{aligned} \quad (4.6)$$

sendo A uma constante de proporcionalidade e ν, ν' os parâmetros na representação de frequências. A equação 4.6 mostra que a autocorrelação está dada pela série de Fourier do quadrado da série original. (56) Portanto, é de se esperar que uma autocorrelação baixa seja refletida numa transformada de Fourier também de magnitude baixa e em geral constante.

4.2.2 Testes de valor-p

Uma série estacionária tem propriedades interessantes. Uma série em geral pode ser decomposta em três partes: uma parte aleatória, uma tendência e uma sazonalidade. A parte aleatória está presente mesmo em séries estacionárias e tem a ver com as flutuações ou ruído a pequena escala no tempo. A tendência (do inglês *trend*) é um movimento regular da série, que varia em escalas maiores, por exemplo numa série crescente, e tem a ver com a mudança da média. A sazonalidade (do inglês *seasonality*) é a componente periódica da variação da série em grande escala, e é útil para analisar séries que devem ter um período natural de variação. Por exemplo, o volume de vendas numa loja de produtos de Natal provavelmente tem periodicidade anual. Nesta seção apresentam-se testes sobre a existência destas componentes.

Os testes de valor-p (do inglês *p-value*) retornam um número entre zero e um que mede a probabilidade de o sujeito do teste ter sido gerado sob certa condição, chamada hipótese

nula. Os testes supõem uma certa densidade de probabilidade da hipótese nula e comparam o valor obtido para o caso particular com esta distribuição. Além da hipótese nula, um teste de valor-p contém uma hipótese alternativa; se o valor obtido for comparável aos valores típicos da distribuição de probabilidade, conserva-se a hipótese nula. No entanto, se o valor for muito diferente, a hipótese nula é rejeitada em favor da hipótese alternativa. Para quantificar isto escolhe-se um nível de significância. Este é o limiar de probabilidade entre aceitar e rejeitar a hipótese nula: se o valor obtido for menor que o nível de significância, a hipótese nula é rejeitada e então aceita-se a hipótese alternativa. Se o valor for maior, conserva-se a hipótese nula. Um valor típico do nível de significância é 0.05, que se traduz em que a hipótese nula não pode ser rejeitada com 95% de confiança.

Antes de introduzir os testes é preciso definir os conceitos usados. Definimos dois casos particulares de séries: o passeio aleatório:

$$x(t) = x(t-1) + \varepsilon(t), \quad (4.7)$$

e o processo determinístico com tendência:

$$x(t) = \beta t + \varepsilon(t), \quad (4.8)$$

onde os $\varepsilon(t)$ são variáveis aleatórias independentes e identicamente distribuídas (com a mesma distribuição de probabilidade). No primeiro caso se diz que existe uma tendência estocástica já que cada valor da série afetará os valores seguintes, sempre com a mesma intensidade. No segundo, existe uma tendência determinística já que a série crescerá ou decairá com o tempo dependendo do sinal de β . A tendência deve ser tratada de forma diferente nos dois casos. No primeiro, pode ser eliminada por diferenciação como será mostrado na próxima seção; e no segundo se utiliza um procedimento chamado *detrending*. Generalizando o processo 4.7 como

$$x(t) = \alpha x(t-1) + \varepsilon(t), \quad (4.9)$$

observamos três possíveis cenários: se $\alpha < 1$ a influência de um valor da série sobre os valores seguintes vai diminuir com o tempo; se $\alpha = 1$ como no passeio aleatório a influência vai se manter constante; e se $\alpha > 1$ os valores mais afastados vão ter uma influência cada vez maior nos valores seguintes, o que se conhece como choque. Portanto, o valor $\alpha = 1$ marca uma mudança de regime, num comportamento que pode ser descrito com o polinômio característico da série. No caso simples da equação 4.9 o polinômio característico é:

$$1 - \alpha z = 0, \quad (4.10)$$

onde z é um número complexo. Para $\alpha = 1$ a solução de 4.10 encontra-se sobre o círculo no plano complexo de raio um e se diz que a série possui uma raiz unitária. O primeiro teste que será usado é o de Kwiatkowski-Phillips-Schmidt-Shin (KPSS). (57) Neste caso a hipótese nula é a estacionariedade com a possibilidade de uma tendência determinística e a hipótese alternativa é a existência de uma raiz unitária. No teste de Phillips-Perron a hipótese nula é que a série é integrada de primeira ordem (o que será explicado na próxima seção) (58), e nos testes aumentados de Dickey-Fuller e de McKinnon a hipótese nula é a existência de uma raiz unitária. Há, porém, duas versões do teste de McKinnon cada uma assumindo uma distribuição de probabilidade diferente da hipótese nula. (59)

4.2.3 Modelo ARIMA

Existem modelos para descrever séries reais que muitas vezes conseguem reproduzir corretamente as séries com uns poucos parâmetros ajustáveis. Estes modelos podem ser generalizados facilmente e permitem uma interpretação do processo que gera a série, assim como resultados analíticos. Nesta seção apresentamos os modelos mais simples nos quais o valor presente da série depende de valores passados em forma de polinômios de primeira ordem.

4.2.3.1 Modelo auto-regressivo AR(p)

O estado atual $x(t)$ de uma série temporal pode depender dos estados anteriores. A forma mais simples de dependência é a linear. Considere uma série onde o estado atual depende dos p estados imediatamente anteriores segundo a relação:

$$x(t) = a_1x(t-1) + a_2x(t-2) + \cdots + a_px(t-p) + \varepsilon(t), \quad t > p, \quad (4.11)$$

onde $\varepsilon(t)$ é um ruído branco com média zero e variância σ^2 . Este tipo de comportamento é chamado de modelo auto-regressivo de ordem p , AR(p). (60) O processo estatístico com tendência da equação 4.9 é o processo AR(1), neste caso particular temos

$$\begin{aligned} x(0) &= \epsilon(0) \\ x(1) &= a_1x(0) + \epsilon(1) = a_1\epsilon(0) + \epsilon(1) \\ x(2) &= a_1x(1) + \epsilon(2) = a_1^2\epsilon(0) + a_1\epsilon(1) + \epsilon(2), \end{aligned} \quad (4.12)$$

e por indução,

$$x(t) = \sum_{i=0}^t a_1^i \varepsilon(t-i). \quad (4.13)$$

Da equação 4.13, a média do processo AR(1) é zero enquanto a variância é $Var(x(t)) = \sigma^2/(1 - a_1^2)$. Fica então quantificado o comportamento mencionado na

seção anterior: se a_1 (i.e. α em 4.9) for menor que a unidade as flutuações são baixas e o sistema não se afasta consideravelmente da média, mas se $a_1 = 1$ (passeio aleatório) a variância é infinita. Um resultado importante do modelo AR(p) é que o seu polinômio característico é dado pela expressão:

$$1 - a_1 z - a_2 z^2 - \cdots - a_p z^p = 0. \quad (4.14)$$

4.2.3.2 Modelo de média móvel MA(q)

O valor atual pode também depender dos q valores anteriores da variável aleatória ϵ da forma

$$x(t) = \epsilon(t) + b_1 \epsilon(t-1) + b_2 \epsilon(t-2) + \cdots + b_p \epsilon(t-p), \quad t > q, \quad (4.15)$$

que é chamado de modelo de média móvel MA(q). Embora a dependência dos valores anteriores da série não seja explícita, os valores estão relacionados. Por exemplo, no caso do modelo MA(1),

$$y(t) = \epsilon(t) + b_1 \epsilon(t-1), \quad (4.16)$$

e assumindo $\epsilon(t) = 0$, temos a relação de recorrência:

$$\begin{aligned} x(1) &= \epsilon(1) \\ x(2) &= b_1 \epsilon(1) + \epsilon(2) = b_1 x(1) + \epsilon(2) \\ x(3) &= b_1 \epsilon(2) + \epsilon(3) = b_1 (-b_1 x(1) + x(2)) + \epsilon(3) \\ &\vdots \\ x(t) &- b_1 x(t-1) + b_1^2 x(t-2) - b_1^3 x(t-3) \cdots = \epsilon(t). \end{aligned} \quad (4.17)$$

A relação de recorrência anterior mostra que o valor atual da série depende implicitamente dos valores anteriores. De fato, a relação 4.17 pode ser interpretada como o modelo AR(∞). (60) O modelo MA(q) é inversível se pode ser convertido no modelo AR(∞), isto se satisfaz se todas as soluções da equação

$$1 + b_1 z + b_2 z^2 + \cdots + b_q z^q = 0 \quad (4.18)$$

estão fora do círculo de raio um ($|z| > 1$). Em particular MA(1) é inversível se $|b_1| < 1$.

4.2.3.3 Modelo integrado I(d)

Uma importante ferramenta para obter estacionariedade a partir de séries não-estacionárias é a derivada, que para tempos discretos é uma equação de diferenças.

Considere por exemplo o passeio aleatório (equação 4.7), o qual sabemos não ser estacionário. No passeio aleatório as contribuições do ruído passado estão contidas no termo $x(t-1)$ e não diminuem com o tempo. Porém, se definirmos

$$y(t) = x(t) - x(t-1) = \varepsilon(t), \quad (4.19)$$

observamos que de fato não existe mais dependência com valores passados. Definimos dois operadores que serão úteis posteriormente, o operador de atraso $L^p y(t) = y(t-p)$, e o operador derivada $\Delta_i^d = (1 - L^i)^d$.

4.2.3.4 Modelo auto-regressivo integrado de média móvel ARIMA(p,d,q)

Podemos juntar os modelos anteriores para obter o modelo mais geral ARIMA(p,d,q). Uma das vantagens do modelo ARIMA é que em geral requer menos parâmetros para descrever uma série que os modelos anteriores. Primeiro definimos o modelo ARMA(p,q) (61) como aquele que satisfaz

$$\begin{aligned} x(t) = & a_1 x(t-1) + a_2 x(t-2) + \dots + a_p x(t-p) + \varepsilon(t) \\ & + b_1 \varepsilon(t-1) + b_2 \varepsilon(t-2) + \dots + b_p \varepsilon(t-p). \end{aligned} \quad (4.20)$$

Usando o operador de atraso a equação 4.20 pode ser escrita de forma mais compacta:

$$\left(1 - \sum_{i=1}^p a_i L^i\right) x(t) = \left(1 + \sum_{i=1}^q b_i L^i\right) \varepsilon(t). \quad (4.21)$$

Finalmente um processo ARIMA se obtém introduzindo o operador derivada dos modelos integrados:

$$\left(1 - \sum_{i=1}^p a_i L^i\right) (1 - L)^d x(t) = \left(1 + \sum_{i=1}^q b_i L^i\right) \varepsilon(t). \quad (4.22)$$

O uso conjunto dos três parâmetros permite uma melhor e mais simples descrição de várias séries temporais.

5 ANÁLISE DE DADOS

Neste trabalho, textos literários são modelados usando redes de co-ocorrência e propõem-se medidas a partir das propriedades das redes. Supomos que estas medidas capturam a essência do estilo de cada autor e portanto podem servir para o reconhecimento da autoria. O método científico baseia-se na criação de hipóteses e na sua posterior validação sob as condições da hipótese. Neste capítulo apresentamos os métodos de avaliação usados neste trabalho junto com outros algoritmos que foram de utilidade para extrair informação dos dados de forma mais eficiente. Precisamos inicialmente definir conceitos das áreas de análise de dados e aprendizado de máquina. Trabalhamos com coleções de textos, uma coleção típica consta de 8 autores com 10 textos cada, perfazendo um total de 80 textos, e cada texto é representado como um número fixo de medidas de rede. Na terminologia do aprendizado de máquina cada texto da coleção é uma instância, cada autor é uma classe e cada uma das medidas de rede usadas para representar um texto é um atributo. Por último, a tabela onde cada fila corresponde a uma instância e cada coluna a um atributo é chamada tabela de dados e se imaginarmos cada fila desta tabela como um vetor, este vetor existiria num espaço chamado espaço de atributos.

5.1 Redução de dimensionalidade

A redução de dimensionalidade visa a diminuir o número total de atributos, seja removendo alguns deles como no caso da seleção de atributos ou criando novos atributos a partir dos existentes, como no caso da extração de atributos. A redução de dimensionalidade é desejável quando o número de atributos é muito grande, o que acontece por exemplo nos métodos tradicionais para classificar um texto onde cada palavra é um atributo; para remover informação redundante ou irrelevante; ou para encontrar relações significativas entre os atributos.

5.1.1 Seleção de atributos

Na seleção de atributos alguns destes são retirados com base em determinados critérios. Apresentamos dois destes.

5.1.1.1 Critério baseado em variância

Considere um atributo, i.e., uma coluna da tabela de dados. A variância é a soma dos quadrados das diferenças entre os valores do atributo para cada instância e o valor médio sob essa coluna. Suponha agora que o valor do atributo é o mesmo para todas as instâncias e portanto a variância é zero; é claro que este atributo não é de nenhuma utilidade para distinguir entre uma instância e outra e pode ser removido sem risco. Da

mesma forma podemos esperar que atributos com baixa variância contribuam pouco na caracterização das instâncias. Portanto, o critério baseado em variância impõe um limiar mínimo de variância, sendo que atributos com variâncias menores sejam removidos.

5.1.1.2 Critério baseado no acerto

Em princípio a forma segura de achar os atributos mais úteis para a predição da classe de uma instância é gerar todas as combinações possíveis entre os atributos disponíveis e escolher as combinações que obtiverem as maiores taxas de acerto. No entanto, há um problema de complexidade com este procedimento. Caso tivéssemos n atributos poderíamos representar cada combinação como um número binário com n dígitos: em cada posição 0 representaria a ausência e 1 a presença do atributo, e uma fila de n números um seria o número total de combinações (em binário). O valor em representação decimal deste número é $2^n - 1$, que cresce extremamente rápido com n , o que faz este procedimento exaustivo ser ineficiente.

Existe uma alternativa que, mesmo precisando considerar um grande número de combinações, é viável. Começa-se com todos os atributos disponíveis e se remove um por vez precisando assim calcular n combinações. Destas combinações se escolhem aquelas que obtiveram a maior taxa de acerto e para cada uma destas combinações se remove um dos $n - 1$ atributos existentes e se repete o processo até não ter atributos restantes. Como a cada iteração podemos ter mais de uma combinação com a maior taxa de acerto não podemos calcular o número exato de combinações a serem consideradas, porém supondo que somente exista uma combinação a cada iteração obtemos um limite inferior igual a $\sum_{i=0}^n i = n(n + 1)/2$, que é um número muito menor do que o total de combinações, mesmo para valores pequenos de n .

5.1.2 Extração de atributos

A extração de atributos pode ser feita com uma série de métodos nos quais novos atributos são gerados a partir dos originais. Desta forma é possível obter informação nova, e muitas vezes melhor, a partir dos dados originais. Isso pode ser usado para visualização ou, num sentido geral, análise de dados. Existem muitos algoritmos considerados como sendo de extração de atributos; no entanto, consideraremos só dois deles que resultaram de particular importância para este trabalho.

5.1.2.1 Análise de componentes principais (PCA)

A análise de componentes principais, do inglês *Principal Component Analysis* (PCA), talvez seja o mais famoso algoritmo de análise de dados, cuja principal aplicação é na redução de dimensionalidade. PCA foi inventado em 1901 por Karl Pearson e tem recebido vários nomes na literatura dependendo da área de aplicação como: transformada

de Hotelling, transformada de Karhunen-Loève, Decomposição Ortogonal Própria, Decomposição em Valores Singulares, etc. (62,63) É importante ressaltar que o PCA não foi projetado como algoritmo de classificação, apesar de que às vezes seus resultados permitam uma separação dos dados.

A filosofia do PCA baseia-se na análise da correlação (ou covariância) das diferentes filas da tabela de dados X . No capítulo 4, especificamente na equação 4.2 define-se a correlação entre duas séries temporais. Como tanto uma série como um vetor são listas ordenadas de números, podemos usar a mesma equação com vetores. A correlação entre as filas i e j da tabela X tem relação com quanto a i -ésima e a j -ésima instância seguem o mesmo padrão e por isto o PCA procura uma combinação (linear) dos atributos na qual a correlação entre diferentes instâncias seja mínima. A covariância é igual à correlação a menos de uma constante; a matriz de covariância pode ser obtida a partir da tabela de dados pela expressão:

$$A = X^T X, \quad (5.1)$$

onde o elemento A_{ij} é igual à correlação das filas i e j de X . A matriz A é quadrada e simétrica, portanto ela é sempre diagonalizável e os seus autovetores são mutuamente ortogonais. Sejam v_i e λ_i o i -ésimo autovetor e autovalor respectivamente. Seja $V = [v_1 | v_2 | \dots | v_n]$ a matriz cuja i -ésima coluna é igual ao i -ésimo autovetor. Do produto AV obtemos:

$$AV = [\lambda_1 v_1 | \lambda_2 v_2 | \dots | \lambda_n v_n] = VD, \quad (5.2)$$

onde D é uma matriz diagonal com elementos,

$$D = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \vdots & 0 \\ \vdots & \dots & \ddots & 0 \\ 0 & 0 & 0 & \lambda_n \end{pmatrix}. \quad (5.3)$$

Usando o fato de que V é ortogonal, podemos obter sua matriz inversa e escrever a expressão 5.2 como $V^{-1}AV = D$. Considerando que $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$ concluímos que os autovetores da matriz de covariância A são as componentes principais da tabela X que apontam nas direções de máxima variância dos dados em ordem decrescente da variância. Podemos então escolher as direções dos primeiros k autovetores sabendo que nas direções ortogonais a estas os dados têm menos variância, assim conseguindo uma redução da dimensionalidade.

5.1.2.2 Isometric feature mapping (Isomap)

O PCA retorna direções no espaço de atributos e portanto é um algoritmo de redução de dimensionalidade linear. Às vezes o uso de algoritmos lineares não contribui na separação dos dados i.e., não existe uma rotação tal que permita observar uma separação das instâncias, como é o caso dos dados da figura 15.

Figura 15 – Variedade bidimensional embebida num espaço tridimensional.

Fonte: COMPRESSIVE... (64)

No entanto, parece que existe uma certa organização dos pontos, e de fato, a superfície curva acima é uma distorção de um quadrado plano onde os pontos de cada cor estão distribuídos ordenadamente sobre linhas paralelas. Os algoritmos de redução de dimensionalidade não-linear conseguem lidar com este tipo de situação. Como exemplo, o Isomap, apresentado em 2000 por Joshua B. Tenenbaum, Vin de Silva, e John C. Langford (65), conseguiu tratar com sucesso a superfície da figura 15. O Isomap serve para organizar coleções de imagens, uma missão extremamente difícil, e é baseado na ideia de identificar e desenrolar uma hiper-superfície ou variedade* de uma certa dimensionalidade embebida num espaço de dimensionalidade maior. Esta variedade é usada posteriormente para alimentar outros algoritmos de análise de dados.

A razão pela qual os pontos no espaço original são problemáticos se usados diretamente nos algoritmos é ilustrada na figura 16. A distância euclidiana (comumente usada

* Do inglês *manifold*. Aliás, os termos “redução de dimensionalidade não-linear” e “manifold learning” são usados indistintamente por alguns autores.

nos algoritmos de classificação), e representada pela linha pontilhada, indica que os pontos ressaltados estão relativamente próximos. No entanto, o grau de similaridade entre eles deveria estar dado pela distância geodésica sobre a variedade representada pela linha sólida que mostra uma grande dissimilaridade, o que faz pensar que os pontos provavelmente pertencem a classes distintas.

Figura 16 – Distâncias entre dois pontos.

Fonte: TENENBAUM et al. (65)

Figura 17 – Trajetória geodésica

Fonte: TENENBAUM et al. (65)

Para poder usar as distâncias geodésicas dentro da variedade, Tenenbaum et al. introduziram uma distância que fosse igual à euclidiana localmente e similar à distância geodésica em escalas maiores. Para um determinado ponto existem pontos próximos, chamados vizinhos, e pontos afastados. A distância entre dois pontos vizinhos será igual à distância euclidiana no espaço de alta dimensão. Já a distância entre dois pontos afastados será igual ao comprimento do menor caminho que teria que ser percorrido para ir do primeiro ao segundo ponto, passando de cada ponto a um dos seus vizinhos, como mostra a linha vermelha da figura 17.

Figura 18 – Distância (azul) e trajetória (vermelho) geodésicas.

Fonte: TENENBAUM et al. (64)

Existem duas versões do algoritmo, o ε -Isomap e o K -Isomap, que diferem na definição de um parâmetro: ε é o raio de uma hiper-esfera centrada num ponto tal que todos os pontos interiores são considerados seus vizinhos e K é um numero fixo de pontos mais próximos considerados vizinhos. Na figura 18 pode-se observar que a distância assim definida é muito parecida à verdadeira distância geodésica ao interior da variedade. Desta forma tem-se uma estimativa das distâncias geodésicas entre todos os pares de pontos; no entanto, ainda falta reduzir a dimensionalidade. Para isto usa-se um algoritmo chamado MDS que será discutido na seção 5.3. Em resumo, MDS converte uma matriz de distâncias numa configuração de pontos num espaço. Formalizando a construção anterior, o Isomap consta de três passos (65):

1. Criar o grafo de vizinhança usando as distâncias euclidianas $d_x(i, j)$ se i e j forem vizinhos.
2. Inicializar a matriz de distâncias d_G sendo $d_G(i, j) = d_x(i, j)$ se i e j forem vizinhos e $d_G(i, j) = \infty$ caso contrário, e iterar para todos os pontos $k = 1, 2, \dots, N$ substituindo $d_G(i, j)$ por $\min\{d_G(i, j), d_G(i, k) + d_G(k, j)\}$.
3. Aplicar o algoritmo MDS na matriz D_G para obter uma representação dos pontos num espaço de dimensão reduzida.

5.2 Classificação

Os algoritmos anteriores serviram para melhorar a qualidade dos dados a serem usados. Nesta seção discutimos a avaliação dos dados para a tarefa solicitada. O objetivo geral do aprendizado de máquina é que um algoritmo use os dados disponíveis e consiga tomar uma decisão sobre um dado desconhecido. O aprendizado de máquina se divide em aprendizado supervisionado, não-supervisionado, e semi-supervisionado. No primeiro caso, também conhecido como classificação, existe um conjunto de dados rotulados chamados de dados de treinamento para ensinar o algoritmo a tomar uma decisão sobre dados sem rótulo, chamados dados de teste. No segundo caso, também conhecido como agrupamento, existe somente um conjunto de dados sem rótulos e o algoritmo precisa encontrar a estrutura subjacente que seria equivalente ao rotulamento. Por último, o aprendizado semi-supervisionado é similar ao não-supervisionado, mas algumas instâncias estão rotuladas, o que pode ajudar a adjudicar um rótulo às instâncias próximas.

Neste trabalho usamos somente o aprendizado supervisionado, já que a sua avaliação é muito mais clara como mostraremos a seguir. Na classificação somente existe uma resposta correta para cada instância: ela tem um rótulo que é desconhecido e o algoritmo deve conseguir prever esse rótulo. Os resultados da classificação podem ser apresentados de forma concisa por meio da matriz de confusão. As filas da matriz de confusão correspondem a cada uma das classes da coleção, portanto a soma dos elementos de cada fila deve ser

igual ao número original de instâncias pertencentes à classe. As colunas da matriz de confusão correspondem à cada classe prevista com o algoritmo das instâncias das filas; assim, se o único elemento não nulo da coluna encontra-se na diagonal da matriz significa que nenhum elemento de outra classe foi erroneamente classificado como sendo da classe em consideração. Em geral, numa classificação perfeita a matriz não tem elementos não-nulos fora da diagonal. A avaliação da classificação dos elementos de uma classe é feita usando dois índices: a precisão e a cobertura. Primeiro vamos renomear os elementos da matriz de confusão: Os verdadeiros positivos VP da classe da fila i são as instâncias que se encontram na posição ii (na diagonal), os falsos positivos FP encontram-se na i -ésima coluna e fora da célula diagonal, e os falsos negativos FN na i -ésima fila e fora da célula diagonal. A precisão da classe A é então:

$$P_A = \frac{VP_A}{VP_A + FP_A}, \quad (5.4)$$

e pode ser interpretada como a tendência da classe de ser diferenciada das outras. A cobertura[†] é definida como:

$$R_A = \frac{VP_A}{VP_A + FN_A}, \quad (5.5)$$

e tem uma interpretação intuitiva como sendo a taxa de acerto da classe. Para ter índices globais de avaliação da classificação da coleção obtemos a média sobre todas as classes. Duas médias são geralmente usadas. As médias micro e macro são respectivamente:

$$micro_M(\{A_i\}) = \frac{1}{k_i} \sum_i k_i M_{A_i}, \quad macro_M(\{A_i\}) = \frac{1}{n} \sum_i M_{A_i}, \quad (5.6)$$

onde M é uma medida genérica (precisão ou cobertura no nosso caso), k é o número de instâncias, e n é o número de classes. Em particular, a média micro da cobertura é a taxa de acerto global, i.e., o número total de instâncias corretamente classificadas sobre o número total de instâncias da coleção. Observe que para coleções onde o número de instâncias de cada classe é o mesmo as duas médias são equivalentes. Apresentamos agora os algoritmos de classificação utilizados no trabalho.

5.2.1 Zero regras (Zero Rule)

A classificação mais simples é a classificação por chance na qual os valores dos atributos são ignorados, daí o nome zero regras (0R). No entanto, existe uma forma eficiente de garantir a melhor taxa de acerto possível sem considerar os atributos. Suponha uma coleção composta por 99 instâncias da classe A e uma instância da classe B . Se perguntarmos a qual classe pertence uma instância desconhecida existe uma boa chance

[†] A notação deve-se ao termo original em inglês, *Recall*.

que a resposta seja A . Usando a mesma lógica, o algoritmo 0R atribui a classe com o maior número de instâncias a todas as instâncias de teste. O pior cenário para este algoritmo ocorre quando todas as classes têm o mesmo número de instâncias, neste caso a probabilidade de acerto é $p = 1/n$, onde n é o número de classes da coleção e todas as instâncias serão atribuídas a uma classe escolhida aleatoriamente. Para a classe escolhida a precisão será $P = p = 1/n$ e a cobertura será $R = 1$; para as outras classes a precisão não está definida e a cobertura será $R = 0/n$; é fácil verificar que a probabilidade de acerto (a cobertura global) é efetivamente $p = (1 + 0(n - 1))/n = 1/n$.

5.2.2 Uma regra (One Rule)

O algoritmo seguinte na ordem de sofisticação é o da uma regra (1R). Neste caso se consideram os atributos, porém, no final a maior parte da informação é desconsiderada. 1R analisa um atributo por vez e associa uma regra a cada valor do atributo e cada classe e escolhe aquele atributo cuja regra gere a maior taxa de acerto entre todos os atributos. A taxa de acerto do atributo escolhido é a taxa de acerto do algoritmo. O pseudocódigo do algoritmo no caso de atributos categóricos (com um conjunto finito de possíveis valores) é:

INICIO

PARA Atributo EM Atributos

PARA Valor EM ValoresAtributo

CONTE O NÚMERO DE VEZES QUE APARECE CADA CLASSE;

ENCONTRE A CLASSE MAIS FREQUENTE;

Regra <- Valor-Classe;

CALCULAR ERRO TOTAL DAS REGRAS DO ATRIBUTO;

FIM PARA;

FIM PARA;

RETORNA Atributo COM O MENOR ERRO;

Um exemplo clássico (66) do uso do algoritmo é prever a classe entre “jogar golfe” e “não jogar golfe” dados os atributos “Panorama”, “Temperatura”, “Umidade”, e “Vento”. Assim, por exemplo para o atributo “Panorama” temos os valores “Ensolarado”, “Nublado”, e “Chuvoso” e para cada valor contamos o número de vezes que aparece com cada classe nas instâncias de teste associando o atributo com a classe onde esteja presente o maior número de vezes. Na tabela 1 é mostrada a contagem do número de aparições de cada valor para cada classe e em negrito se marcam os valores mais frequentes que definem as regras de associação: se “Panorama” é “Ensolarado” então “Jogar golfe”.

Este atributo é justamente o que tem o menor erro (igual a quatro), e portanto é escolhido como o resultado do algoritmo 1R com uma taxa de acerto de $p = (3 + 4 +$

de instâncias I que cheguem até o nó será escolhido. Para um atributo A , o ganho de informação $G(I, A)$ é a diferença entre a entropia de Shannon $S(I)$ entre as instâncias de entrada e a organização das instâncias na saída:

$$G(I, A) = S(I) - \sum_{\nu \in \text{Valores}(A)} \frac{|I_\nu|}{|I|} S(I_\nu). \quad (5.7)$$

Vamos ilustrar como usar o ganho de informação para construir uma árvore com um exemplo simples. Novamente usaremos atributos categóricos, A_1 e A_2 , os quais somente podem assumir os valores “Verdadeiro” e “Falso” e no nó haste teremos 32 instâncias, 16 pertencentes à classe “+” e 16 à classe “-” como se mostra na figura 20.

Figura 20 – Exemplo de construção da árvore de decisão.

Fonte: Elaborada pelo autor.

Observamos na parte esquerda da figura que nas 32 instâncias existem 20 para as quais o valor do atributo A_1 é “Verdadeiro” e 12 para as quais A_1 é “Falso”. Para este caso onde somente existem duas classes possíveis, a entropia é:

$$S(I) = -p_+ \text{Log}_2 p_+ - p_- \text{Log}_2 p_-, \quad (5.8)$$

onde p_+ e p_- são as proporções das instâncias (e.g. $p_+ = |I_+| / |I|$). O ganho de informação em cada caso é calculado a seguir. Para o nó haste, nos dois casos

$$S = -\frac{16}{32} \text{Log}_2 \frac{16}{32} - \frac{16}{32} \text{Log}_2 \frac{16}{32} = 1, \quad (5.9)$$

que é a máxima entropia possível. Para o atributo A_1 ,

$$\begin{aligned} S_V &= -\frac{12}{20} \log_2 \frac{12}{20} - \frac{8}{20} \log_2 \frac{8}{20} = 0.971, \\ S_F &= -\frac{4}{12} \log_2 \frac{4}{12} - \frac{8}{12} \log_2 \frac{8}{12} = 0.918, \\ G &= 1 - \frac{20}{32} 0.971 - \frac{12}{32} 0.918 = 0.0488. \end{aligned} \tag{5.10}$$

Para o atributo A_2 ,

$$\begin{aligned} S_V &= -\frac{4}{20} \log_2 \frac{4}{20} - \frac{16}{20} \log_2 \frac{16}{20} = 0.722, \\ S_F &= -\frac{12}{12} \log_2 \frac{12}{12} = 0, \\ G &= 1 - \frac{20}{32} 0.722 = 0.549. \end{aligned} \tag{5.11}$$

Vemos que a melhor escolha é o atributo A_2 . De fato, o ganho de informação é muito alto, o que era esperado considerando-se que no galho correspondente ao valor “Falso” somente há instâncias da classe “+” e portanto a entropia é nula, ou equivalentemente a compressão da informação é perfeita.

5.2.4 K-primeiros vizinhos (k-Nearest Neighbors)

K-primeiros vizinhos ou *k-Nearest Neighbors* (kNN) é um algoritmo extremamente simples. Porém, às vezes pode ter uma eficácia comparável à de outros algoritmos mais complexos. Ele pertence à família dos algoritmos chamados preguiçosos. Os algoritmos preguiçosos são aqueles que não constroem nenhum modelo de classificação (como é o caso do kNN) ou fazem poucos cálculos antes de serem solicitados para classificar uma instância de teste. Estes algoritmos demoram pouco na fase de treinamento e muito na fase de teste. Curiosamente, o resultado destes algoritmos pode ser melhor que o dos assim chamados algoritmos ansiosos, especialmente na presença de ruído. Isto porque os algoritmos ansiosos costumam criar um modelo sobre o espaço de atributos que suaviza as particularidades deste, ignorando as irregularidades a pequena escala. O kNN também pode ser usado dentro de outros algoritmos (e.g. Radial basis function networks) com entidades diferentes às instâncias.

Como mencionamos, o kNN não faz nenhum cálculo na fase de treinamento e simplesmente guarda os dados (i.e. as posições das instâncias de treinamento no espaço de atributos). Na fase de teste, as distâncias desde a posição da nova instância até todas as instâncias de treinamento são calculadas; escolhem-se então as k instâncias mais próximas e se faz um processo de “votação”. À nova instância será atribuída a classe majoritária entre

os seus k vizinhos. Desta forma, o algoritmo atribui classes de acordo com as instâncias efetivamente mais próximas e não segundo algum tipo de função. O maior problema do algoritmo é a quantidade de distâncias que precisa calcular desnecessariamente quando o número de instâncias é muito grande; para atacar este problema têm se desenvolvido técnicas como a busca do primeiro vizinho (*nearest neighbor search*). (69)

5.2.5 Bayesiano ingênuo (Naive Bayes)

A tabela de dados pode ser interpretada como uma tabela de probabilidades condicionais. Lembrando que as probabilidades condicionais de dois eventos independentes[‡] A e B pode ser obtida dos termos da equação redundante $P(A \cap B) = P(B \cap A)$, podemos desenvolver cada lado em relação a cada evento e obter o teorema de Bayes:

$$P(A|B)P(B) = P(B|A)P(A). \quad (5.12)$$

A interpretação probabilística da tabela de dados é mais intuitiva usando dados de categorias. Suponha que tanto os possíveis valores V de cada atributo A como as possíveis classes C são eventos. A probabilidade condicional de ocorrer um valor dado que “ocorreu” uma classe é dada pela proporção das vezes que o valor apareceu nas instâncias pertencentes à classe. Considere por exemplo a seguinte tabela de atributos:

Tabela 2 – Tabela de dados com atributos categóricos.

	A₁	A₂	A₃
C₁	+	−	+
	−	−	+
	+	+	+
C₂	−	+	−
	+	+	+
	−	−	+

Fonte: Elaborada pelo autor.

Teríamos as seguintes probabilidades condicionais: $P(A_1 = +|C_1) = 2/3$, $P(A_2 = -|C_2) = 1/3$, $P(A_3 = -|C_1) = 0$. O problema da classificação de uma nova instância pode ser descrito como “achar a classe dado que os valores dos atributos são aqueles da nova instância”. Isto é, achar os $P(C|A_1, A_2, A_3)$. Por enquanto consideramos a aparição do conjunto de valores dos atributos como um único evento e usamos o teorema de Bayes 5.12 para escrever,

$$P(C_i|\{A_j\}) = \frac{P(\{A_j\}|C_i)P(C_i)}{P(\{A_j\})}. \quad (5.13)$$

[‡] A independência de eventos será usada mais uma vez posteriormente.

onde $P(C_i)$ é a proporção de instâncias pertencentes à classe C_i da coleção. Usamos a hipótese de independência para separar o termo acima:

$$P(C_i|\{A_j\}) = \frac{\prod_j P(A_j|C_i)P(C_i)}{Z}, \quad (5.14)$$

onde substituímos o denominador por uma constante genérica de normalização sabendo que a soma das probabilidades de todas as C_i é igual à unidade. No exemplo, se quisermos classificar uma nova instância que possui atributos $(+, -, +)$ calculamos as probabilidades condicionais: $P(A_1 = +|C_1) = 2/3$, $P(A_2 = -|C_1) = 2/3$, $P(A_3 = +|C_1) = 1$, e $P(A_1 = +|C_2) = 1/3$, $P(A_2 = -|C_2) = 1/3$, $P(A_3 = +|C_2) = 2/3$, e desconsiderando $P(C_i)$ já que o número de instâncias das duas classes é o mesmo, aplicamos a equação 5.13 para C_1 ,

$$P(C_1|+, -, +) = Z^{-1} \cdot \frac{2}{3} \cdot \frac{2}{3} \cdot 1 = \frac{4}{9Z}, \quad (5.15)$$

e para C_2 ,

$$P(C_2|+, -, +) = Z^{-1} \cdot \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{27Z}. \quad (5.16)$$

Desde já podemos ver que a nova instância tem maior chance de ser rotulada como sendo da classe C_1 . Especificamente, usando a condição de normalização encontramos $Z = 14/27$, $P(C_1|+, -, +) = 6/7$, e $P(C_2|+, -, +) = 1/7$. Mesmo com atributos dependentes, a hipótese ingênua de independência pode gerar bons resultados.

5.2.6 Redes de função de base radial (Radial Basis Function Networks)

As redes de função de base radial (*Radial Basis Function Networks* (RBFN)) fazem parte de um conjunto de algoritmos baseados em redes neurais, usados em várias aplicações, incluindo o aprendizado de máquina. Elas formam o coração do paradigma de *deep learning* que tem tido grande sucesso e é muito promissor. Estas redes são compostas por camadas e o resultado da análise de uma camada converte-se nos dados de entrada para a camada seguinte[§]. Assim, o conceito de deep learning consiste na capacidade de incluir várias camadas intermediárias para melhorar a análise.

A rede de função de base radial consiste somente de três camadas, uma camada de entrada composta por um número de neurônios (nós) igual ao número de atributos n (i.e., à dimensionalidade do espaço) os quais estão completamente conectados aos neurônios da camada intermediária, como mostra a figura 21.

Em cada neurônio da camada intermediária encontra-se uma função de base radial num espaço de dimensionalidade n . Em termos gerais, uma função de base radial é

[§] Porém, existem variantes que incluem retroalimentação.

Figura 21 – Rede de função de base radial.

Fonte: Adaptada de RADIAL... (70)

monotônica esfericamente simétrica, que associa um peso dependendo do valor do raio. Uma escolha típica para a função de base radial é a função gaussiana,

$$h(x) = \exp\left(-\frac{(x - c)^2}{r^2}\right), \quad (5.17)$$

onde c é a localização do centro da função de base radial e r é uma medida do tamanho da função. Assim, para cada instância introduzida será calculado o valor da função h em cada um dos neurônios da camada intermediária. O valor f na saída é a predição da rede que neste caso será a classe da instância de entrada. O enlace entre cada neurônio da camada intermediária e a saída terá um peso, mostrando a influência de cada centro na predição:

$$f(x) = \sum_{j=1}^m w_j h_j(x). \quad (5.18)$$

No caso das redes do deep learning existe mais de um neurônio nesta terceira camada e os resultados destes neurônios podem por sua vez ser juntados, repetindo-se o processo várias vezes. Vamos ilustrar o funcionamento de uma rede RBFN com um exemplo simples tomado de (71): suponha que haja quatro instâncias de treinamento como ilustra a figura 22, onde a cor de cada instância corresponde à classe. Claramente esta

configuração é problemática para ser tratada com algoritmos lineares de classificação já que não existe uma linha reta que possa separar as classes por regiões.

Figura 22 – Espaço de atributos.

Fonte: Adaptada de RADIAL... (71)

Figura 23 – Espaço de funções de base radial.

Fonte: Adaptada de RADIAL... (71)

Se associarmos o valor 0 à classe dos pontos brancos e o valor 1 à classe dos pontos vermelhos podemos expressar a tabela de dados nas terceira e quarta colunas da tabela 3, onde a primeira coluna é o índice da instância e a segunda é a classe. Se criarmos dois centros, um em[¶] (0,0) e outro em (1,1) e aplicarmos a equação 5.17 com as distâncias dos pontos aos centros e fixando $r = 1$ teremos os valores listados nas duas últimas colunas.

Tabela 3 – Parâmetros do RBFN.

p	C	x_1	x_2	h_1	h_2
1	0	0	0	1.0000	0.1353
2	1	0	1	0.3678	0.3678
3	1	1	0	0.3678	0.3678
4	0	1	1	0.1353	1.0000

Fonte: Elaborada pelo autor.

Com estes valores das funções radiais aplicamos a equação 5.18 para obter os pesos w . Neste caso temos,

$$f(x) = w_1 h_1(x) + w_2 h_2(x) + \theta, \quad (5.19)$$

[¶] Na prática as posições dos centros são escolhidas aleatoriamente.

onde θ é uma constante. Usando os valores da tabela 3 temos o sistema de equações lineares,

$$\begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1.0000 & 0.1353 & -1 \\ 0.3678 & 0.3678 & -1 \\ 0.3678 & 0.3678 & -1 \\ 0.1353 & 1.0000 & -1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \theta \end{pmatrix}, \quad (5.20)$$

que tem três equações independentes (e uma dependente), três variáveis, cuja solução é $w_1 = w_2 = -2.5018$ e $\theta = -2.8404$. Usando a equação 5.19 com estes valores podemos estimar a classe de uma instância de teste: o valor y obtido será comparado com 0 e 1 para estimar a probabilidade de a instância pertencer a cada classe.

5.3 Escalamento multi-dimensional (Multi-Dimensional Scaling)

Na seção 5.1.2.2 comentamos que o Isomap utiliza o escalamento multi-dimensional ou MDS. Certamente este é um dos algoritmos mais usados apesar de não ser tão conhecido, pois não é um algoritmo de classificação. Pelo contrário, o resultado é uma distribuição das instâncias num espaço de atributos. O algoritmo deriva posições a partir de distâncias. Por exemplo, ingressando uma matriz das distâncias entre algumas cidades europeias é possível obter uma projeção bidimensional que assemelha-se às localizações geográficas, como mostra a figura 24. De fato, o mapeamento para duas ou três dimensões é amplamente usado em visualização de dados.

Uma matriz de distâncias D , definida como a matriz simétrica cuja célula d_{ij} contém a distância entre os elementos i - e j -ésimo, é dada ao algoritmo, que por sua vez retorna uma distribuição das posições x_i dos elementos num espaço de dimensão definida pelo usuário. A distribuição dos elementos no espaço é escolhida para que as distâncias entre os elementos sejam mantidas da melhor forma possível. Isto é, que a função de perda, ou estresse, seja minimizada. Existem muitas escolhas para o estresse, sendo uma das mais populares:

$$Stress_D(x_1, x_2, \dots, x_n) = \left(\sum_{i \neq j=1, \dots, n} (d_{ij} - |x_i - x_j|)^2 \right)^{1/2}. \quad (5.21)$$

Em geral, a função de estresse é uma medida da diferença entre as distâncias introduzidas e as distâncias das posições no espaço de atributos retornadas. O MDS clássico procura encontrar a matriz de posições $X = [x_1, x_2, \dots, x_n]$ tal que:

$$d_{ij} = |x_i - x_j|. \quad (5.22)$$

Figura 24 – Projeção bidimensional das cidades da Europa usando MDS.

Fonte: NOEL... (72)

Podemos impor a condição de as posições estarem centradas, i.e.,

$$\sum_{i=1}^n x_{ik} = 0, \quad (5.23)$$

para todas as coordenadas k . Usando a matriz Gram $B = X'X$ de tamanho $n \times n$ podemos escrever a relação 5.22 como,

$$d_{ij}^2 = b_{ii}^2 + b_{jj}^2 - 2b_{ij}. \quad (5.24)$$

Somando sobre os índices, usando a condição 5.23, e definindo o traço $T = \sum_{i=1}^n b_{ii}$ obtemos as relações (ver apêndice E):

$$\sum_{i=1}^n d_{ij}^2 = T + nb_{jj}, \quad \sum_{j=1}^n d_{ij}^2 = T + nb_{ii}, \quad \sum_{i,j=1}^n d_{ij}^2 = 2nT. \quad (5.25)$$

Definindo as matrizes H e S tais que $h_{ij} = (\delta_{ij} - 1/n)$ e $s_{ij} = d_{ij}^2$, usando as relações 5.25 e substituindo em 5.24 para obter b_{ij} , verificamos que:

$$B = -\frac{1}{2}HSH, \quad (5.26)$$

satisfaz 5.22 (ver apêndice E). As coordenadas podem ser obtidas a partir de B se considerarmos a decomposição nos autovetores V e na matriz diagonal de autovalores Λ

de $B = V\Lambda V'$:

$$X = \Lambda^{1/2}V'. \quad (5.27)$$

Recapitulando, a condição imposta 5.22 é satisfeita definindo as matrizes S , H e B segundo a expressão 5.26, e os valores das coordenadas são obtidos com a decomposição em autovalores e autovetores de B . Os autovalores em Λ estão na ordem decrescente e portanto podemos fixar a dimensionalidade $m \leq n$ escolhendo os primeiros m autovalores e autovetores. Por último, uma técnica para conhecer o valor mais adequado de m consiste em voltar à equação 5.21 e observar a dependência do estresse com a dimensionalidade. Em geral, o estresse é reduzido monotonicamente à medida que aumenta o número de dimensões, salvo exceções como o exemplo da figura 24 onde existe um mínimo na dimensionalidade original dos dados. Se o estresse diminui monotonicamente o número de dimensões é escolhido tal que a variação no estresse deixe de ser significativa se adicionarmos mais uma dimensão, o que também é conhecido como o método do cotovelo. (73)

6 DINÂMICA DE REDES

Com os conceitos e métodos apresentados nos capítulos anteriores, temos os elementos necessários para desenvolver a abordagem de análise de textos a partir de redes de co-ocorrência. Neste capítulo usaremos a dinâmica das redes para tal propósito. A ideia desta abordagem consiste em estudar as propriedades dinâmicas da estrutura de um livro ao longo da leitura deste. O objetivo é verificar se é possível caracterizar um texto a partir da dinâmica de redes, em particular para a tarefa de reconhecimento de autoria. A abordagem baseia-se na evidência de que medidas estáticas (i.e. fixas) de redes têm sido bem sucedidas para reconhecer autoria (17–19), assim como para outras tarefas de classificação (15, 16). Como comentado no capítulo 4, existem abordagens teóricas para o crescimento ou evolução de redes. No entanto, até o momento não há trabalhos que apliquem esta abordagem à classificação de textos. A inconveniência destas abordagens é a de que textos, e portanto redes, são de diferentes tamanhos. A análise por séries temporais contorna esse problema, além de enfatizar a dinâmica e não os mecanismos de evolução.

6.1 Construção de séries temporais

Na nossa abordagem, um livro é representado como uma série temporal de redes como mostrado na figura 25. Primeiro o texto é preprocessado como um todo, e em seguida procede-se a fatiar o texto em partes contendo o mesmo número de palavras, com a condição de contar a repetição de palavras, isto é, uma palavra que aparecer duas vezes é contada duas vezes. Esta forma de dividir o texto cria redes com o mesmo peso total nas arestas. As redes são construídas como descrito na seção 3.3 e posteriormente são calculadas as métricas apresentadas na seção 3.2. Como calculamos tanto métricas locais (e.g. conectividade) quanto métricas globais (e.g. diâmetro), são obtidas métricas globais das locais calculando-se as médias sobre toda a rede. Assim, cada rede é representada por um conjunto de doze números, ou seja, cada livro é representado por doze séries temporais. A vantagem desta abordagem é que, para os tamanhos típicos das redes usadas, é possível trabalhar com uma estatística confiável ao invés de usar outras características das distribuições.

O peso total, isto é, o tamanho dos sub-textos foi escolhido como sendo $W = 200$, já que os melhores resultados entre os valores testados foram obtidos para este caso. A explicação para este comportamento é o compromisso entre a estatística nas redes e o ruído nas séries temporais. Por um lado, para este tamanho encontrou-se que o número de nós era sempre maior que 100; portanto podemos confiar que a média global obtida é estatisticamente representativa para os perfis das distribuições obtidas. Por outro lado, quanto maiores forem as redes menor será o número de elementos da série temporal

correspondente, dificultando sua análise.

Figura 25 – Construção das séries temporais.

Fonte: Elaborada pelo autor.

6.2 Análise de séries temporais

Havendo construído as $12 \times 80 = 960$ séries temporais para a coleção de livros, procedemos à análise destas como mostrado no capítulo 4.

6.2.1 Não-linearidade e caos

O estudo preliminar da existência de dinâmica não-linear mostrou retratos de fase que não apresentavam atratores ou ciclos limite claramente definidos, e portanto esta hipótese não foi prosseguida. Por exemplo, as figuras 26 e 27 mostram os retratos de fase da clusterização e do número de nós para o livro “War and Peace” de Leo Tolstoi (escolhido por ser o maior da coleção) que apresentam grande dispersão, típica de todos os retratos de fase obtidos. Mesmo que existam medidas em geral mais dispersas do que outras*, a grade formada na segunda figura deve-se ao fato de o número de nós ser inteiro. A presença de atratores e ciclos limite indicaria uma dinâmica não-linear. No entanto, a conclusão contrária não é sempre verdadeira e esta é uma área muito avançada que poderia continuar sendo pesquisada.

* Na figura 13 mostrou-se o retrato de fase da centralidade de carga do livro The Sea-Wolf do Jack London.

Figura 26 – Retrato de fase da clusterização para o livro War and Peace.

Fonte: Elaborada pelo autor.

Figura 27 – Retrato de fase do número de nós para o livro War and Peace.

Fonte: Elaborada pelo autor.

6.2.2 Auto-correlação

Novamente considerando o livro War and Peace, que poderia ter correlações de longo alcance devido ao seu tamanho, as auto-correlações encontram-se na faixa de $\pm 5\%$. Este valor é considerado próprio de uma série descorrelacionada, como mostrado na figura 28 para a intermediação de carga. Nota-se também que o valor cai rapidamente do valor inicial, igual a um, para esta faixa, o que reflete pouca relação entre um elemento da série e os imediatamente anteriores.

A transformada de Fourier correspondente à autocorrelação anterior é mostrada na figura 29. Neste caso a função não está normalizada e devemos comparar as alturas relativas das componentes. A componente zero da transformada de Fourier está relacionada com o valor médio da função, que é muito maior do que as outras, mostrando que não existe componente periódica considerável. Além disto, as flutuações assemelham-se a um ruído branco, que é uma característica de sinais não-modulados.

6.2.3 Testes de valor-p

Os testes de valor-p estão sumarizados no apêndice B. Os valores listados são médias sobre todos os livros e não resultados para livros individuais, mas certas tendências são claras. O teste mais importante é provavelmente o de KPSS que avalia a hipótese nula de estacionariedade com tendência. Neste caso a hipótese nula não foi rejeitada para nenhuma das métricas e elas podem ser consideradas estacionárias. Para os outros testes que avaliam a existência de uma raiz nula há uma discrepância: segundo o teste de Phillips-Perron, todas as medidas, exceto a centralidade de carga e o número de arestas,

Figura 28 – Autocorrelação da centralidade de carga para o livro War and Peace.

Fonte: Elaborada pelo autor.

não passam no teste e portanto não têm raiz unitária. Por outro lado, nos testes restantes a hipótese nula de raiz unitária é mantida sem exceção.

6.2.4 Modelo ARIMA

O modelo ARIMA ajusta os dados segundo os três conjuntos de parâmetros da equação 4.22 p , d , e q . Lembrando que o primeiro é o número de elementos da série imediatamente anteriores a um elemento dado dos quais este depende. Para o mesmo elemento, q é o número de valores aleatórios ou ruído imediatamente anteriores que o influenciam. Estes valores são importantes, pois se referem à memória do sistema, mas o mais importante é o valor d . É possível obter séries estacionárias a partir de certas classes de séries não estacionárias por derivação, isto é, subtraindo-se dois elementos sucessivos da série. Neste processo a série diminui de tamanho uma unidade cada vez que é derivada, e o parâmetro d é o número de derivadas necessárias para obter uma série estacionária. Se uma série tem um dado valor do parâmetro d , diz-se que a mesma é integrável de ordem d .

As séries obtidas para os livros tiveram em geral valores baixos dos três parâmetros. Para cada um em separado, e para todos juntos, como apresentado no apêndice C. Encontrou-se que 27% das séries são integráveis de primeira ordem ($d = 1$), sendo as restantes estacionárias. Em relação aos outros parâmetros, o máximo valor foi de cinco nos dois casos. Poderia ainda acontecer de a soma dos valores para várias séries ser grande, mas isso não ocorreu, pois o máximo valor dos três parâmetros foi nove para uma série ARIMA(5, 0, 4).

Figura 29 – Transformada de Fourier para o livro War and Peace.

Fonte: Elaborada pelo autor.

6.3 Medidas propostas

Os resultados das seções anteriores mostram que em geral as séries podem ser consideradas estacionárias, o que é ilustrado pela figura 30 em que são mostradas as séries para as doze métricas do livro Moby Dick, de Herman Melville. Observe que o número de nós e especialmente de arestas está muito perto da janela escolhida de $W = 200$ e que a escala é logarítmica.

Figura 30 – Séries temporais para as métricas do livro Moby Dick do Herman Melville.

Fonte: Elaborada pelo autor.

Como já mencionado, a forma forte de estacionariedade exige que todas as estatísticas sejam constantes ao longo da série. Não utilizamos essa condição tão restritiva, mas pressupomos que os quatro primeiros momentos da distribuição dos valores da série sejam constantes. Com isto em mente, propomos as seguintes medidas para classificar os textos:

$$\mu_i = \left[\frac{1}{T-1} \sum_{j=1}^T (x_j - \mu_1)^i \right]^{1/i} \quad (6.1)$$

onde x_j são os valores da série, $2 \leq \mu_i \leq 4$ e μ_1 é a média. Se calcularmos a distribuição dos valores da série para uma métrica e para todos os autores obteremos a figura 31. Nota-se que as medidas propostas podem ser bons indicativos da assinatura de cada autor: a média e a variância são claramente diferentes e algumas distribuições possuem alta assimetria, como no caso de Herman Melville e de Jonathan Swift.

Figura 31 – Histogramas da conectividade por autor.

Fonte: Elaborada pelo autor.

Temos portanto $12 \times 4 = 48$ medidas (i.e. números) que definem cada texto e que serão usadas como atributos nos algoritmos de classificação.

6.4 Resultados

As medidas propostas foram avaliadas usando os algoritmos de classificação apresentados na seção 5.2 para o conjunto original representado por uma tabela de dados de 80 filas e 48 colunas e aplicando cada uma das técnicas de redução da dimensionalidade da seção 5.1 que geraram tabelas alternativas. Os algoritmos foram escolhidos por serem

representativos dos principais tipos de algoritmos usados na atualidade. Os resultados estão sumarizados na tabela 4 e serão discutidos ao longo desta seção.

Tabela 4 – Resumo das taxas de acerto obtidas com o conjunto original de 48 atributos e aplicando seleção de atributos e extração de atributos. Para PCA e Isomap foi usado $n_{componentes} = 10$ e para Isomap $n_{vizinhos} = 10$. O máximo erro absoluto foi de 5.7% obtido com KNN.

Atributos*	J48(%)	KNN(%)	NB(%)	RBFN(%)
Conjunto original	45	62.5	61.25	56.25
Seleção por variância	55	67.5	63.75	63.75
Seleção por acerto	75	78.75	77.5	75
$\{\mu_1\}$	45	43.75	46.25	40
$\{\mu_2, \mu_3, \mu_4\}$	38.75	63.75	60	57.5
PCA	40	46.25	48.75	42.5
Isomap	63.75	88.75	81.25	83.75

*Para os conjuntos de atributos com um parâmetro livre, escolheu-se o melhor resultado encontrado.

Fonte: Elaborada pelo autor.

6.4.1 Seleção de atributos

A seleção de atributos entendida como a remoção seletiva de alguns permite conhecer a importância relativa de um atributo na classificação. Apresentamos dois critérios de seleção de atributos e mais dois conjuntos alternativos.

6.4.1.1 Baseada em variância

Este critério define um limiar de variância mínima que deve satisfazer o atributo, i.e. os elementos da respectiva coluna. Percorremos os diferentes valores do limiar como se mostra na figura 32, onde os quatro atributos obtidos da mesma métrica estão agrupados e a parte preta significa que o atributo está presente na seleção. Por exemplo, na parte superior da figura o limiar é zero e todos os atributos estão presentes. À medida que observamos a figura de cima para baixo, começam a aparecer atributos com baixa variância, sendo o primeiro o quarto momento do número de cliques $\mu_4^{(cliques)}$. Curiosamente o último atributo a ser removido (o de maior variância) é $\mu_3^{(cliques)}$. Observe que quase todos os atributos são removidos numa estreita faixa de valores.

As taxas de acerto para os diferentes valores do limiar de variância são apresentadas na figura 33. No extremo esquerdo estão as taxas de acerto originais, i.e. sem redução de dimensionalidade, com taxa máxima de 62.5% usando KNN. À medida que se aumenta o limiar, há um comportamento irregular das taxas de acerto mas para todos os algoritmos

Figura 32 – Atributos usando a seleção baseada em variância. Iniciais do eixo horizontal: C, clusterização; D, diâmetro; R, radio; Cq, numero de cliques; L, centralidade de carga; T, transitividade; B, centralidade de intermediação; S, Menor caminho; K, conetividade; I, intermitência; N, número de nós; e E, número de eixos.

Fonte: Elaborada pelo autor.

de classificação, e encontramos uma melhora em algum limiar intermediário, atingindo máximos representados por esferas pretas na figura. Os valores dos limiares nos quais foram atingidos esses máximos estão sinalizados na figura 33. Por último, quando removermos todos os atributos atingimos a taxa de acerto do ZeroR igual a 12.5%.

Figura 33 – Taxas de acerto usando a seleção baseada em variância.

Fonte: Elaborada pelo autor.

6.4.1.2 Baseada em acerto

Uma situação diferente ocorre com as taxas de acerto para a seleção baseada em acerto ou aquelas nas quais um atributo é removido por vez visando a otimizar o valor da taxa de acerto, como mostrado na figura 34. Neste caso o crescimento é constante (monotônico salvo poucas exceções) à medida que removemos os atributos e as curvas em

geral apresentam comportamento similar entre si. Após atingir os valores máximos começa uma queda rápida para o valor do ZeroR. Duas particularidades merecem especial atenção: os números de atributos com as maiores taxas são muito menores que o número original de 48 atributos. Além disto, o valor destas taxas é comparável às taxas encontradas na literatura (com um máximo de 78.75% usando KNN) (18,19), mostrando que uma remoção estratégica de atributos melhora consideravelmente a classificação.

Figura 34 – Taxas de acerto usando a seleção baseada em acerto.

Fonte: Elaborada pelo autor.

Infelizmente não encontramos uma tendência a priori dos atributos que deveriam ser removidos a fim de obter uma taxa de acerto ótima como mostrado na figura 35, onde se apresentam as combinações de atributos que atingiram os valores máximos. Novamente, na figura os quadros pretos representam atributos presentes em cada combinação. No caso do KNN duas combinações atingiram o valor máximo. Não há uma similitude entre os atributos ótimos para todos os algoritmos, salvo algumas exceções. A mais notória é a ausência total da centralidade de carga e quase total do coeficiente de clusterização e do diâmetro. Ao contrário, o número de arestas e a transitividade são frequentemente usados.

A seleção de atributos mostrou que em geral existem subconjuntos dos atributos que têm taxas de acerto maiores que as do conjunto original. No entanto, é difícil saber quais atributos a serem escolhidos e se esta seleção é válida em geral, ainda mais porque depende fortemente do algoritmo de classificação. A seleção baseada em variância não mostrou ser muito efetiva, devido a seu comportamento irregular, o que pode ser em razão da similitude entre as variâncias dos atributos considerados. Mencione-se que a seleção baseada no acerto foi otimizada para obter boas taxas de acerto, mas a otimização não foi exaustiva e portanto pode haver outras combinações com melhores resultados.

Figura 35 – Melhores atributos usando a seleção baseada em acerto. Iniciais do eixo horizontal: C, clusterização; D, diâmetro; R, radio; Cq, numero de cliques; L, centralidade de carga; T, transitividade; B, centralidade de intermediação; S, Menor caminho; K, conetividade; I, intermitência; N, número de nós; e E, número de eixos.

Fonte: Elaborada pelo autor.

Dois últimos sub-conjuntos de atributos foram considerados. Como sabe-se que atributos que não consideram a dinâmica conseguem uma boa classificação (17–19), e dado que os primeiros momentos μ_1 representam a parte estática das medidas, comparamos o conjunto destes primeiros momentos com o conjunto complementar dos momentos de ordem superior. O conjunto dos momentos $\{\mu_2, \mu_3, \mu_4\}$ captura a variabilidade com relação à média e portanto a dinâmica das séries. Os resultados variam dependendo do algoritmo e não é possível estabelecer qual conjunto é superior para a classificação. No entanto, observou-se que o sub-conjunto $\{\mu_1\}$ nunca superou o conjunto original, demonstrando que a adição dos momentos de ordem superior melhora o desempenho.

6.4.2 Extração de atributos

Consideremos agora o caso em que a remoção de informação não é feita com atributos individuais, mas analisando a organização no espaço de atributos. Apresentamos os resultados de dois métodos: um método linear, o PCA; e um método não linear, o Isomap. Em ambos os casos a redução foi feita do espaço original 48-dimensional a um espaço de dez dimensões. Os resultados da extração de atributos com ambos os algoritmos estão na figura 36.

6.4.2.1 PCA

Dada a popularidade do PCA, os resultados podem parecer surpreendentes: para todos os algoritmos tanto a precisão quanto a taxa de acerto diminuem, em comparação com os valores da coleção original. O maior valor da taxa de acerto ocorre com NB de 48.75%, consideravelmente menor que os 61.25% obtidos com o mesmo algoritmo na coleção original. Como o PCA é uma rotação no espaço de atributos que otimiza a variabilidade

Figura 36 – Resultados da extração de atributos.

Fonte: Elaborada pelo autor.

entre as dimensões rotacionadas e eliminando as restantes, concluímos que não existe uma rotação tal que seja possível conservar a informação relevante nesse número de dimensões. Para outras dimensionalidades as taxas de acerto também são menores ou no máximo iguais às taxas da coleção original.

6.4.2.2 Isomap

Ilustramos o que pode ser esperado do Isomap com a figura 37 feita com o software de visualização Projection EXplorer (PEX) (74). Na figura foi feita a projeção para o espaço bidimensional e fica em evidência a separação dos textos dos autores, em particular do Arthur Conan Doyle e do Bernard Shaw enquanto os textos dos outros autores, mesmo que misturados, ficam em regiões preferenciais. Visualmente, a projeção bidimensional promete uma alta taxa de acerto, porém, encontrou-se que existem resultados melhores explorando outros valores dos parâmetros, resultado que apresentamos em detalhe no apêndice D. Como o Isomap tem um parâmetro livre, o número de vizinhos[†], variamos este parâmetro e o número de dimensões (também chamado de número de componentes) até um valor máximo de quinze cada. Na primeira parte do apêndice apresentamos as

[†] Usamos a variante K-Isomap.

taxas de acerto para cada algoritmo de classificação em forma de histograma. Observamos altos valores em todos os casos e mais importante ainda, que as distribuições têm uma assimetria para a direita, ou seja, a maioria das taxas de acerto tem valores altos.

Figura 37 – Projeção Isomap bidimensional.

Fonte: Elaborada pelo autor.

A maior taxa de acerto neste trabalho foi obtida usando o Isomap e o algoritmo KNN e fixando $n_{vizinhos} = 10$ e $n_{componentes} = 10$ e tem um valor de 88.75%. (20) Os resultados do KNN para outros valores estão em detalhe no apêndice D. A taxa de acerto é superior àquelas encontradas na literatura para métodos baseados em redes. Cabe esclarecer que para poder comparar diretamente as taxas de acerto, coleções com as mesmas características devem ser consideradas, e felizmente as coleções usadas nos trabalhos prévios têm características similares à nossa. Em (18) a máxima taxa de acerto foi 65% obtida com uma coleção composta de 40 livros de 8 autores, i.e. o mesmo número de autores que a nossa com a metade dos livros. Em outro trabalho (19), 36 livros de 5 autores na língua persa são analisados conseguindo classificar 28 corretamente, i.e. uma taxa de acerto de 77%. As taxas do presente método também são comparáveis às dos métodos tradicionais (não baseados em redes), por exemplo (75) reporta métodos com taxas máximas de 87% para coleções com 5 autores e 80% para coleções com 10 autores.

Concluimos que, embora seja possível atingir taxas de acerto maiores do que a da coleção original aplicando uma seleção de atributos, não existe uma maneira consistente de garantir um resultado exitoso, a não ser testando exaustivamente todas as combinações.

Por outro lado, uma redução não linear da dimensionalidade oferece um procedimento padrão com ótimos resultados. O êxito do uso do Isomap mostra que as medidas propostas encontram-se numa hiper-superfície curva onde os livros do mesmo autor estão geralmente juntos.

7 O PAPEL DAS PALAVRAS

A representação em forma de redes dos textos tem tido muito sucesso devido à natureza complexa da linguagem. No entanto, para a maioria das tarefas de análise automática incluindo reconhecimento de autoria continua se preferindo o uso das técnicas de análise de elementos lexicais descritas na seção 2.2. A grande maioria destas técnicas desconsidera a estrutura do texto e poderiam ser chamadas de abordagens frequenciais já que baseiam-se na contagem dos elementos lexicais. Mesmo assim algumas conseguem taxas de acerto competitivas, o que é a causa da sua popularidade.

Por outro lado, na representação de redes o texto é analisado desde o ponto de vista topológico: quando uma rede de co-ocorrência é construída, por exemplo para a tarefa de reconhecimento de autoria, as palavras que compõem o texto são esquecidas, e no seu lugar os nós aos que deram origem passam ser o objeto de estudo. Enquanto as abordagens frequenciais tradicionais concentram-se justamente em quais palavras são usadas sem dar importância à estrutura do texto além da mesma frase, os trabalhos de redes têm ignorado um aspecto importantíssimo da linguagem: as palavras; ou em termos mais gerais, os elementos lexicais. Apresentamos aqui uma proposta que reconcilia ambas as abordagens visando aproveitar as vantagens de cada. Mostraremos também que a nossa abordagem permite ser generalizada para aproveitar a variedade de medidas de rede disponíveis.

Propomos uma representação híbrida dos textos que inclui elementos da representação de redes e elementos lexicais dos mesmos. Posteriormente aplicaremos a representação para definir medidas de similaridade entre textos para ser aplicadas na tarefa de reconhecimento de autoria em coleções de textos. A metodologia utilizada é a seguinte: constroem-se as redes de co-ocorrência da coleção de textos e adiciona-se a informação lexical (i.e. as palavras) para obter a representação proposta, seguidamente define-se uma família de medidas de similaridade; gera-se uma matriz de distâncias para cada medida de similaridade; e finalmente juntam-se as diferentes matrizes de distâncias e os dados resultantes usam-se como entrada para os algoritmos de classificação. (76)

7.1 Representação híbrida dos textos

As redes de co-ocorrência dos textos são construídas segundo o descrito na seção 3.3. Posteriormente escolhe-se um conjunto de métricas de rede locais aos nós, i.e. a métrica deve ter um valor para cada nó da rede. Escolhe-se um número N_p de palavras a serem consideradas principais das métricas de rede, o resto das palavras será desconsiderado da análise. Neste trabalho encontramos que na faixa de $N_p \in [50, 150]$ palavras se obtêm os resultados mais promissórios e portanto fixamos o número de palavras em $N_p = 100$. As palavras principais escolhem-se de acordo com a métrica, especificamente, conforme o

perfil da distribuição sobre os nós da métrica.

O método é bem sucedido se escolhermos os nós mais particulares, i.e. aqueles que apresentam um valor incomum da métrica. Nas figuras 38 e 39 apresentam-se as distribuições das centralidades de intermediação com $N_p = 20$ palavras principais sinalizadas para os livros *The Memoirs of Sherlock Holmes* e *The Return of Sherlock Holmes* do Arthur Conan Doyle. Note-se que as palavras são quase as mesmas e ocupam posições próximas e neste sentido os livros são similares.

Figura 38 – Distribuição da centralidade de intermediação e palavras principais do livro *The Memoirs of Sherlock Holmes*.

Fonte: Elaborada pelo autor.

7.2 Medidas de similaridade

Na representação introduzida, um livro é descrito unicamente por uma lista organizada de palavras para cada métrica de rede considerada. Por exemplo os livros do Arthur Conan Doyle estariam definidos pelas colunas primeira e terceira da tabela 5. Se da uma qualificação a cada palavra de acordo com a seguinte regra: a primeira palavra recebe uma qualificação igual ao número de palavras principais N_p e cada palavra da sequência recebe uma qualificação uma unidade menor que a da palavra predecessora como mostrado nas colunas segunda e quarta da tabela 5.

A medida de similaridade é proporcional ao produto interno ponderado por estas

Figura 39 – Distribuição da centralidade de intermediação e palavras principais do livro *The Return of Sherlock Holmes*.

Fonte: Elaborada pelo autor.

qualificações:

$$A \cdot B = \sum_{w \in V_A \cap V_B} C_A(w)C_B(w), \quad (7.1)$$

onde A e B são os textos, V são os vocabulários (todas as palavras) dos textos e C é a qualificação de uma palavra em cada texto. Para o exemplo teríamos os valores listados na tabela 6. Devido à definição da qualificação das palavras, todos os livros possuem a mesma norma, i.e.

$$A \cdot A = B \cdot B = N_p(N_p + 1)(2N_p + 1)/6, \quad (7.2)$$

que no exemplo é igual a 2870, portanto os livros escolhidos têm uma alta similaridade. Especificamente, definimos a similaridade como sendo o produto interno entre dois vetores sobre a norma de qualquer deles:

$$s_{AB} \equiv \frac{A \cdot B}{A \cdot A} = \frac{A \cdot B}{B \cdot B}, \quad (7.3)$$

que para o exemplo é igual a 0.948. Da mesma forma definimos a distância* como $d_{AB} = 1 - s_{AB}$, e a matriz de distâncias de uma coleção de textos como aquela cuja

* Usamos aqui o termo distância como sinônimo de dissimilaridade (ver (9) pag.275).

Tabela 5 – Palavras principais dos livros de Arthur Conan Doyle.

Memoirs		Return	
Palavra	Qualificação	Palavra	Qualificação
say	20	one	20
one	19	holmes	19
come	18	say	18
upon	17	man	17
man	16	upon	16
see	15	would	15
holmes	14	come	14
would	13	know	13
find	12	see	12
could	11	could	11
go	10	take	10
hand	9	well	9
u	8	room	8
take	7	find	7
make	6	go	6
know	5	hand	5
room	4	u	4
little	3	make	3
well	2	face	2
face	1	look	1

Fonte: Elaborada pelo autor.

Tabela 6 – Produto interno dos livros de Arthur Conan Doyle.

Palavra	$C_A C_B$
one	380
say	360
upon	272
man	272
holmes	266
come	252
would	195
see	180
could	121
find	84
take	70
know	65
go	60
hand	45
u	32
room	32
make	18
well	18
face	2
little	0
look	0
$A \cdot B$	2722

Fonte: Elaborada pelo autor.

ij -ésima entrada é a distância entre o i -ésimo e o j -ésimo livro: $D^{(m)} = \{d_{ij}\}$, onde o super-índice indica a métrica de rede que foi usada. Aplicando este procedimento na primeira coleção do apêndice C usando a centralidade de intermediação obtemos a figura 40. A coleção de livros está organizada de tal forma que os dez primeiros livros pertencem ao primeiro autor, os dez seguintes pertencem ao segundo e assim sucessivamente, isto se reflete na matriz de distâncias onde podemos ver quadrados de tamanho 10 de cores mais cálidas ao longo da diagonal o que mostra que os livros daquele autor são mais similares entre eles.

7.3 União de matrizes de distâncias

Uma das vantagens do método é a obtenção de uma família de medidas de similaridade entre dois textos, uma similaridade por cada métrica de rede usada. Porém, é preciso juntar a informação de todas elas a fim de aproveitar esta multiplicidade. As distâncias não podem ser adicionadas diretamente e portanto é preciso procurar uma forma de obter variáveis que sejam compatíveis. A ferramenta MDS, usualmente usada para visualização

Figura 40 – Matriz de distâncias usando centralidade de intermediação.

Fonte: Elaborada pelo autor.

e apresentada na seção 5.3, tem como entrada um conjunto de distâncias entre elementos e como saída um conjunto de posições desses elementos num espaço de dimensão predefinida tal que as distâncias no espaço reproduzam da melhor maneira possível as distâncias originais. Na figura 41 mostra-se o mapeamento para um espaço bidimensional da matriz de distâncias da centralidade de intermediação da terceira coleção no apêndice C.

Figura 41 – Mapeamento de uma matriz de distâncias num espaço bidimensional usando MDS. As cores representam autores e os números correspondem aos índices que identificam os livros.

Fonte: Elaborada pelo autor.

Esta ferramenta se adapta justamente à nossa necessidade. A ideia geral é converter as matrizes de distâncias em posições em espaços (um espaço por cada matriz de distâncias)

e posteriormente juntar, para cada elemento, essas diferentes posições num único espaço conjunto cuja dimensão seja a soma das dimensões dos espaços anteriores. Vamos ilustrar o procedimento com um exemplo: considere três elementos (livros) A , B , e C e duas métricas de rede m_1 e m_2 . Aplicando o método apresentado até agora obtemos duas matrizes de distâncias entre A , B e C de tamanho 3×3 , $D^{(m_1)}$ e $D^{(m_2)}$. Em seguida usa-se o MDS para converter as matrizes de distâncias em posições e vamos supor que a matriz $D^{(m_1)}$ foi mapeada num espaço tridimensional enquanto a matriz $D^{(m_2)}$ foi mapeada num espaço bidimensional. Chamando por exemplo $\vec{A}_1$ à posição correspondente ao elemento A no espaço correspondente a m_1 teríamos:

$$D^{(m_1)} \xrightarrow{MDS} \begin{aligned} \vec{A}_1 &= (x_1^A, y_1^A, z_1^A) \\ \vec{B}_1 &= (x_1^B, y_1^B, z_1^B) \\ \vec{C}_1 &= (x_1^C, y_1^C, z_1^C) \end{aligned}, \quad D^{(m_2)} \xrightarrow{MDS} \begin{aligned} \vec{A}_2 &= (x_2^A, y_2^A) \\ \vec{B}_2 &= (x_2^B, y_2^B) \\ \vec{C}_2 &= (x_2^C, y_2^C) \end{aligned} \quad (7.4)$$

a posição no espaço conjunto estará dada por:

$$\begin{aligned} \vec{A} &= (x_1^A, y_1^A, z_1^A, x_2^A, y_2^A) \\ \vec{B} &= (x_1^B, y_1^B, z_1^B, x_2^B, y_2^B) \\ \vec{C} &= (x_1^C, y_1^C, z_1^C, x_2^C, y_2^C) \end{aligned} \quad (7.5)$$

Interpretando o espaço conjunto como um espaço de atributos, utilizamos diretamente os dados em 7.5 nos algoritmos de classificação. Para decidir o número de dimensões de um mapeamento existe uma regra de ouro chamada método do cotovelo. O estresse define-se em termos das diferenças entre as distâncias de entrada e de saída. Usualmente quanto maior é o número de dimensões menor é o estresse e teremos uma curva monotonicamente decrescente como ilustrado[†] na figura 42. A filosofia é que será escolhido o número de dimensões tal que uma dimensão extra não vai beneficiar significativamente os resultados. No exemplo da figura poderíamos conservar somente duas dimensões. No entanto existem figuras onde o “cotovelo” não está tão claramente definido, nesse caso podemos escolher um limiar como uma porcentagem do valor inicial e escolher o primeiro número de dimensões com estresse abaixo desse limiar, por exemplo 10% indicado pela linha a traços. Escolhemos a segunda alternativa já que permite automatizar o processo de decisão.

7.4 Resultados

Para avaliar o modelo proposto foram usadas 3 coleções, cada uma composta de textos de 8 autores do século 19 com 10 textos por autor, uma das quais já foi usada

[†] A figura não corresponde a dados reais, a relação é $y = x^{-3}$.

Figura 42 – Exemplo ilustrativo do método do cotovelo.

Fonte: Elaborada pelo autor.

Tabela 7 – Taxas de acerto (em percentagem) obtidas com o método proposto e com TF-IDF.

	J48	KNN	NB	RBFN
Redes				
Coleção 1	63.75	90.00	93.75	96.25
Coleção 2	72.50	83.75	92.50	90.00
Coleção 3	73.75	98.75	92.50	95.00
TF-IDF				
Coleção 1	41.25	72.50	71.25	70.00
Coleção 2	40.00	73.75	67.50	67.50
Coleção 3	28.75	52.50	52.50	57.50

Fonte: Elaborada pelo autor.

no capítulo anterior. 22 dos 24 autores foram nativos falantes da língua Inglesa como descrito em detalhe no apêndice C. Para coleções com estas características a taxa de acerto do ZeroR (i.e. por mera chance) é de $1/8 = 12.5\%$. Foram usadas quatro métricas de rede: conectividade, centralidade de intermediação, menor caminho, e intermitência. Também foram descartadas duas métricas que não apresentaram o desempenho desejado (possivelmente devido aos perfis das distribuições): o coeficiente de clusterização e a centralidade de autovalor. O método mostrou um desempenho excepcional como mostrado na tabela 7 e na figura 43. Estas taxas ultrapassam amplamente àquelas obtidas com o método TF-IDF para o qual a máxima taxa de acerto foi de 73.75%. Admiravelmente, o método proposto atingiu uma taxa de 98.75% na terceira coleção com o algoritmo de classificação KNN, o que é uma classificação quase perfeita considerando que corresponde a 79 dos 80 livros da coleção corretamente classificados. Este desempenho é robusto

considerando que as taxas se mantêm acima de 90% na maioria dos casos. As taxas são superiores às obtidas tanto com a aproximação frequencial quanto com a aproximação de redes quando usadas individualmente demonstrando que uma ótima caracterização dos textos requer das duas abordagens em conjunto.

Figura 43 – Taxas de acerto com o método introduzido e com TF-IDF.

Fonte: Elaborada pelo autor.

8 CONCLUSÕES

8.1 Conclusões gerais

A principal conclusão inferida dos resultados desta tese é a de que as métricas propostas dependem do autor do texto, e portanto tais métricas podem ser usadas para reconhecimento de autoria com ótimo desempenho.

O estudo das métricas de redes continua numa dicotomia: existe uma relação entre as diferentes métricas de redes reais? Se existir é preciso descobrir qual é; se não existir é preciso descobrir o que cada métrica significa para o sistema. Neste trabalho descobrimos que o uso conjunto das métricas melhora a diferenciação entre autores, e assim a segunda opção parece ser a correta. Caso contrário, não teríamos tal melhora. Podemos também afirmar que o comportamento é diferente entre as métricas como mostrado na figura 30, e entre os autores como mostrado na figura 31.

Seria útil perseguir nesse trabalho com redes de co-ocorrência em vista da importância do reconhecimento de autoria e dos resultados promissores apresentados aqui. Embora as taxas de acerto nesta tese sejam superiores às de trabalhos anteriores (18, 19), especialmente com o segundo método em que os rótulos dos nós (palavras) foram levados em conta, ainda é possível refinar os métodos. Em geral observou-se clara variabilidade entre os algoritmos de classificação, o que pode afetar o desempenho. A seguir mostramos as conclusões para cada método.

8.2 Dinâmica de redes

Para uma simples seleção de atributos, as taxas obtidas são razoáveis com uma taxa máxima de 78.75%. Mais importante, ela permitiu distinguir entre os diferentes atributos. Em geral os atributos têm variância similar e a centralidade de intermediação não apareceu nas melhores taxas, a despeito de ser semelhante à centralidade de intermediação que pelo contrário foi bastante escolhida. Os outros dois subconjuntos demonstraram a importância das medidas “dinâmicas” μ_2 , μ_3 , e μ_4 , com desempenho em geral melhor que a média. μ_1 é duas vezes melhor que o conjunto original de atributos.

No caso da extração de atributos os resultados melhoraram com um máximo de 88.75%. O contraste entre os resultados obtidos aplicando PCA, ISOMAP e sem usar nenhum deles, como mostrado na figura 36, permite compreender a natureza das medidas propostas. De um lado, a deterioração das taxas de acerto ao usar o PCA mostra que, embora o PCA seja ferramenta útil para certos casos, estas condições não são aplicáveis para os nossos dados. O PCA serve para tratar dados de alta dimensionalidade e remover dados sem importância, mas o número de atributos originais não apresentou problemas para

os algoritmos de classificação usados, além de que todos esses atributos foram necessários. Outra forma de ver isto é lembrar que o PCA é uma rotação no espaço de atributos, removendo as dimensões ortogonais àquelas consideradas principais. O fato de perder poder de classificação mostra que não foi conveniente remover as ditas dimensões. Por outro lado, o sucesso do ISOMAP mostra que os atributos encontram-se numa hiper-superfície, mas a relação de proximidade entre textos do mesmo autor se mantém.

8.3 O rol das palavras

O ótimo resultado obtido para as três coleções abre novas possibilidades e suscita novas questões. Considere por exemplo a intermitência, ou o seu inverso, a periodicidade. Uma palavra pode ter baixa intermitência e portanto alta pontuação em dois textos. No entanto, ela pode aparecer poucas vezes no texto, o que mostra que a estrutura macroscópica do texto é controlada pelo autor, inadvertidamente ou não. O método também vai um passo além das abordagens tradicionais, que somente estudam o papel das palavras ao nível de frases.

O maior problema deste método está na necessidade de usar métricas de nó, o que limita sua generalização. Além disto, nem todas as métricas levaram a bom resultado. Por exemplo, o coeficiente de clusterização, que tem uma distribuição em forma de sino, não foi útil. As métricas que têm funcionado possuem distribuições em forma de leis de potência, o que abre a questão do papel da complexidade nas propriedades únicas da linguagem.

8.4 Perspectivas para trabalhos futuros

- Seria interessante descobrir se é possível enganar nossa abordagem na classificação, i.e. se ela é resistente a um ataque usando estilometria do adversário. (77, 78)
- É possível aplicar as técnicas a outras coleções e línguas, e seria interessante ver as diferenças para comparar com a literatura.
- Flexibilizar o tipo de coleções seria útil. Os métodos poderiam ser testados em coleções com outras características, como diferente número de textos entre autores ou textos menores.
- Uma linha alternativa de pesquisa envolveria a comparação com os modelos teóricos e ver se é possível ajustar os parâmetros para obter resultados similares aos nossos.
- Caos é uma alternativa que ainda não se pode descartar e existem na literatura métodos mais sofisticados que os retratos de fase para analisá-lo. (54, 55)
- Os algoritmos de classificação se comportaram de diferentes formas. Considerando que os algoritmos usados foram os mais simples, poderíamos usar algoritmos mais sofisticados.

- O trabalho feito nesta tese é chamado de reconhecimento de autoria do primeiro tipo. O reconhecimento de segundo tipo consiste em usar um número de autores conhecido e responder se um texto dado é de um deles ou de um autor externo. O reconhecimento de terceiro tipo consiste em saber de quem é um texto para uma coleção aberta. (25) Propor métodos para abordar reconhecimento de segundo e terceiro tipos seria de grande interesse.
- Por último e mais importante, seria interessante descobrir se há relação entre as diferentes métricas. Se existir, identificar a relação, e como isso pode afetar o reconhecimento de autoria.

8.5 Produção

O trabalho de doutorado gerou dois artigos e duas participações em congressos internacionais:

Artigos

- Akimushkin C, Amancio DR, de Oliveira ON, On the role of words in the network structure of texts: application to authorship attribution, *arXiv preprint* arxiv.org/abs/1705.04187 11 May 2017.
- Akimushkin C, Amancio DR, de Oliveira ON, Text Authorship Identified Using the Dynamics of Word Co-Occurrence Networks, *PLOS ONE* [dx.doi.org/10.1371/journal.pone.0170527](https://doi.org/10.1371/journal.pone.0170527) 26 Jan 2017.

Participação em congressos

- Authorship recognition using the dynamics of co-occurrence networks *SoFiA: Latin American School and Workshop on Data Analysis and Mathematical Modelling of Social Sciences*, Buenos Aires, Argentina, 7-11 Novembro 2016.
- Time series from text co-occurrence networks, *6th International Conference on Nonlinear Science and Complexity*, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Brasil, Maio 2016.

REFERÊNCIAS

- 1 SMITH, K.; BRIGHTON, H.; KIRBY, S. Complex systems in language evolution: the cultural emergence of compositional structure. **Advances in Complex Systems**, v. 6, n. 4, p. 537–558, 2003.
- 2 STEYVERS, M.; TENENBAUM, J. B. The large-scale structure of semantic networks: statistical analyses and a model for semantic growth. **Cognitive Science**, v. 29, n. 1, p. 41–78, 2005.
- 3 GELL-MANN, M. Complexity and complex adaptive systems. In: COWAN, G.; PINES, D.; METZER, D.(Ed.) **Complexity: methaphors, models and reality**. New York: Addison Wesley, 1994.
- 4 GELL-MANN, M.; RUHLEN, M. The origin and evolution of word order. **Proceedings of the National Academy of Sciences**, v. 108, n. 42, p. 17290–17295, 2011.
- 5 ESTOUP, J. B. **Gammes stenographiques**. Paris: Institut Stenographique, 1916.
- 6 ZIPF, G. K. **The psycho-biology of language**. Oxon: George Routledge Ltd, 1935.
- 7 SORNETTE, D. **Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder**. Berlin: Springer Science & Business Media, 2006.
- 8 HERDAN, G. **The advanced theory of language as choice and chance**. New York: Springer-Verlag, 1966.
- 9 JUOLA, P. Authorship attribution. **Foundations and Trends in Information Retrieval**, v. 1, n. 3, p. 233–334, 2006.
- 10 MOSTELLER, F.; WALLACE, D. **Inference and disputed authorship: the federalist**. New York: Addison-Wesley, 1964.
- 11 TWEEDIE, F. J.; SINGH, S.; HOLMES, D. I. Neural network applications in stylometry: the federalist papers. **Computers and the Humanities**, v. 30, n. 1, p. 1–10, 1996.
- 12 ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. **Reviews of Modern Physics**, v. 74, n. 1, p. 47, 2002.
- 13 CANCHO, R. Ferrer-i; SOLÉ, R. V. Two regimes in the frequency of words and the origins of complex lexicons: Zipf’s law revisited. **Journal of Quantitative Linguistics**, v. 8, n. 3, p. 165–173, 2001.
- 14 DOROGOVTSSEV, S. N.; MENDES, J. F. Language as an evolving word web. **Proceedings of the Royal Society of London. Series B**, v. 268, n. 1485, p. 2603–2606, 2001.
- 15 GAO, Y.; LIANG, W.; SHI, Y.; HUANG, Q. Comparison of directed and weighted co-occurrence networks of six languages. **Physica A**, v. 393, p. 579–589, 2014.

- 16 AMANCIO, D. R.; OLIVEIRA, O. N. J.; COSTA, L. F. Identification of literary movements using complex networks to represent texts. **New Journal of Physics**, v. 14, n. 4, p. 043029, 2012.
- 17 AMANCIO, D. R. A complex network approach to stylometry. **PLoS One**, v. 10, n. 8, p. e0136076, 2015.
- 18 AMANCIO, D. R.; ALTMANN, E. G.; OLIVEIRA, O. N. J.; COSTA, L. da F. Comparing intermittency and network measurements of words and their dependence on authorship. **New Journal of Physics**, v. 13, n. 12, p. 123024, 2011.
- 19 MEHRI, A.; DAROONEH, A. H.; SHARIATI, A. The complex networks approach for authorship attribution of books. **Physica A**, v. 391, n. 7, p. 2429–2437, 2012.
- 20 AKIMUSHKIN, C.; AMANCIO, D. R.; OLIVEIRA, O. N. J. Text authorship identified using the dynamics of word co-occurrence networks. **PLoS ONE**, v. 12, n. 1, p. e0170527, 2017.
- 21 MEHLER, A.; KÖHLER, R. **Aspects of automatic text analysis**. London: Springer Science & Business Media, 2007. v. 209.
- 22 COSTA, L. F. et al. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. **Advances in Physics**, v. 60, n. 3, p. 329–412, 2011.
- 23 MITKOV, R. **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2005.
- 24 RUDMAN, J. The state of authorship attribution studies: Some problems and solutions. **Computers and the Humanities**, v. 31, n. 4, p. 351–365, 1997.
- 25 JUOLA, P.; STAMATATOS, E. **Overview of the author identification task at pan 2013**. Disponível em: <<http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-JuolaEt2013.pdf>>. Acesso em: 27 mar. 2017.
- 26 STAMATATOS, E. et al. **Overview of the author identification task at PAN 2014**. 2017. Disponível em: <<http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-JuolaEt2013.pdf>>. Acesso em: 27 mar. 2017.
- 27 PENG, F.; SCHUURMANS, D.; WANG, S.; KESELJ, V. Language independent authorship attribution using character level language models. In: CONFERENCE EUROPEAN OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 10, EACL, 2003. **Proceedings...** Budapest, 2003. v. 1, p. 267–274.
- 28 ESCALANTE, H. J.; SOLORIO, T.; GÓMEZ, M. Montes-y. **Local histograms of character n-grams for authorship attribution**. Disponível em: <<http://www.aclweb.org/anthology/P/P11/P11-1030.pdf>>. Acesso em: 27 mar. 2017.
- 29 FORSTALL, C.; SCHEIRER, W. Features from frequency: authorship and stylistic analysis using repetitive sound. **Journal of the Chicago Colloquium on Digital Humanities and Computer Science**, v. 1, n. 2, p. 1–23, 2010.
- 30 KUKUSHKINA, O. V.; POLIKARPOV, A. A.; KHMELEV, D. V. Using literal and grammatical statistics for authorship attribution. **Problems of Information Transmission**, v. 37, n. 2, p. 172–184, 2001.

- 31 CHASKI, C. E. Who's at the keyboard? authorship attribution in digital evidence investigations. **International Journal of Digital Evidence**, v. 4, n. 1, p. 1–13, 2005.
- 32 KEŠELJ, V.; PENG, F.; CERCONE, N.; THOMAS, C. **N-gram-based author profiles for authorship attribution**. 2003. Disponível em: <<http://web.cs.dal.ca/~vlado/papers/pacling03.pdf>>. Acesso em: 28 mar. 2017.
- 33 CLEMENT, R.; SHARP, D. Ngram and bayesian classification of documents for topic and authorship. **Literary and Linguistic Computing**, v. 18, n. 4, p. 423–447, 2003.
- 34 BAAYEN, H.; HALTEREN, H. V.; TWEEDIE, F. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. **Literary and Linguistic Computing**, v. 11, n. 3, p. 121–132, 1996.
- 35 RYGL, J.; ZEMKOVÁ, K.; KOVÁR, V. Authorship verification based on syntax features. In: WORKSHOP ON RECENTE ADVANCES IN SLAVONIC NATURAL LANGUAGE, 6, RASLA, 2012. **Proceedings...** Tibun EU: [s.n.], 2012. p. 111–119.
- 36 ZELLIG, H. Distributional structure. **Word**, v. 10, n. 2–3, p. 146–62, 1954.
- 37 SPARCK-JONES, K. A statistical interpretation of term specificity and its application in retrieval. **Journal of Documentation**, v. 28, n. 1, p. 11–21, 1972.
- 38 MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. et al. **Introduction to information retrieval**. Cambridge: Cambridge University Press, 2008. v. 1.
- 39 ZHANG, J.; MARSZALEK, M.; LAZEBNIK, S.; SCHMID, C. Local features and kernels for classification of texture and object categories: a comprehensive study. **International Journal of Computer Vision**, v. 73, n. 2, p. 213–238, 2007.
- 40 VOUTILAINEN, A. Part-of-speech tagging. In: **The Oxford handbook of computational linguistics**. Oxford: Oxford University Press, 2003. p. 219–232.
- 41 SAPKOTA, U.; SOLORIO, T.; MONTES, M.; BETHARD, S. **Domain adaptation for authorship attribution: improved structural correspondence learning**. 2016. Disponível em: <<http://www.aclweb.org/anthology/P16-1210>>. Acesso em: 20 jul. 2017.
- 42 GREENE, B. B.; RUBIN, G. M. **Automatic grammatical tagging of English**. Providence, Rhode Island: Department of Linguistics, Brown University, 1971.
- 43 ERDOS, P.; RÉNYI, A. On the evolution of random graphs. **Mathematical Institute Hungarian Academy Science**, v. 5, n. 1, p. 17–60, 1960.
- 44 WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. **Nature**, v. 393, n. 6684, p. 440–442, 1998.
- 45 BAK, P.; TANG, C.; WIESENFELD, K. Self-organized criticality: an explanation of the 1/f noise. **Physical Review Letters**, v. 59, n. 4, p. 381, 1987.
- 46 BOCCARA, N. **Modeling complex systems**. New York: Springer Science & Business Media, 2010.
- 47 LEE, T. I. et al. Transcriptional regulatory networks in *saccharomyces cerevisiae*. **Science**, American Association for the Advancement of Science, v. 298, n. 5594, p. 799–804, 2002.

- 48 COSTA, L. d. F.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. R. V. Characterization of complex networks: a survey of measurements. **Advances in physics**, Taylor & Francis, v. 56, n. 1, p. 167–242, 2007.
- 49 COMPRESSIVE processing of manifold-modeled data. 2017. Disponível em: <https://en.wikipedia.org/wiki/Centrality>. Acesso em: 21 fev. 2017.
- 50 BIEMANN, C.; QUASTHOFF, U. Networks generated from natural language text. In: In GANGOLY, N. et al. (Ed.) **Dynamics on and of complex networks**. Birhhauser: Springer, 2009. p. 167–185.
- 51 IZHIKEVICH, E. M. **Dynamical systems in neuroscience**. Massachusetts: MIT Press, 2007.
- 52 KINOUCI, O.; COPELLI, M. Optimal dynamical range of excitable networks at criticality. **Nature Physics**, v. 2, n. 5, p. 348–351, 2006.
- 53 KURAMOTO, Y. **Chemical oscillations, waves, and turbulence**. New York: Springer Science & Business Media, 2012. v. 19.
- 54 ANISHCHENKO, V. S.; VADIVASOVA, T.; STRELKOVA, G. **Deterministic nonlinear systems**. New York: Springer, 2014.
- 55 KANTZ, H.; SCHREIBER, T. **Nonlinear time series analysis**. Cambridge: Cambridge University Press, 2004.
- 56 WEISSTEIN, E. W. **Wiener-Khinchin Theorem**. 2017. Disponível em: <http://mathworld.wolfram.com/Wiener-KhinchinTheorem.html>. Acesso em: 21 fev. 2017.
- 57 KWIATKOWSKI, D.; PHILLIPS, P. C. B.; SCHMIDT, P.; SHIN, Y. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? **Journal of Econometrics**, v. 54, n. 1-3, p. 159–178, 1992.
- 58 PHILLIPS, P. C. B.; PERRON, P. Testing for a unit root in time series regression. **Biometrika**, v. 75, n. 2, p. 335–346, 1988.
- 59 MACKINNON, J. G. Numerical distribution functions for unit root and cointegration tests. **Journal of Applied Econometrics**, v. 11, n. 6, p. 601–618, 1996.
- 60 SCHMIDT, A. B. **Quantitative finance for physicists: an introduction**. New York: Academic Press, 2010.
- 61 BOX, G. E. P.; JENKINS, G. M. **Time series analysis: forecasting and control**. New York: John Wiley, 1976.
- 62 SKILLICORN, D. **Understanding complex datasets: data mining with matrix decompositions**. Boca Raton: CRC Press, 2007.
- 63 MITCHELL, T. M. Machine learning (mcgraw-hill international editions computer science series). McGraw-Hill, 1997.

- 64 COMPRESSIVE processing of manifold-modeled data. 2011. Disponível em: <http://archive.cnx.org/contents/24ab633d-e8e2-46d3-8ad8-429220d9802a\spacefactor\@m{}6/compressive-processing-of-manifold-modeled-data>. Acesso em: 21 fev. 2017.
- 65 TENENBAUM, J. B.; SILVA, V. D.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **Science**, v. 290, n. 5500, p. 2319–2323, 2000.
- 66 ONER. 2017. Disponível em: <http://www.saedsayad.com/oner.htm>. Acesso em: 21 fev. 2017.
- 67 QUINLAN, J. **C4. 5: programs for machine learning**. c4. 5-programs for machine learning/j. ross quinlan. In: SALZBERG, S. **Machine learning**. Boston: Kluwer Academic Publishing, 1994. v. 16, p. 236–240.
- 68 SHANNON, C. E. A mathematical theory of communication. **Bell System Technical Journal**, v. 27, n. 3, p. 379423, 1948.
- 69 KNUTH, D. E. **The art of computer programming: sorting and searching**. Boston: Pearson Education, 1998. v. 3.
- 70 RADIAL basis function network. 2017. Disponível em: http://www.saedsayad.com/artificial_neural_network_rbf.htm. Acesso em: 21 fev. 2017.
- 71 RADIAL Basis Function Networks: Applications. 2017. Disponível em: <http://www.cs.bham.ac.uk/~jxb/INC/l15.pdf>. Acesso em: 21 fev. 2017.
- 72 NOEL O’Blog: Convert distance matrix to 2D projection with Python. 2017. Disponível em: <http://baoilleach.blogspot.com.br/2014/01/convert-distance-matrix-to-2d.html>. Acesso em: 21 fev. 2017.
- 73 IZENMAN, A. J. **Modern multivariate statistical techniques**. New York: Springer, 2008.
- 74 PAULOVICH, F. V.; OLIVEIRA, M. C. F.; MINGHIM, R. The projection explorer: A flexible tool for projection-based multidimensional visualization. In: BRAZILIAN SYMPOSIUM ON COMPUTER GRAPHICS AND IMAGE PROCESSING, 20, SIBGRAPH, 2007 **Proceedings...** Belo Horizonte: UFMG, 2007. p. 27–36.
- 75 GRIEVE, J. Quantitative authorship attribution: an evaluation of techniques. **Literary and Linguistic Computing**, v. 22, n. 3, p. 251–270, 2007.
- 76 AKIMUSHKIN, C.; AMANCIO, D. R.; OLIVEIRA, O. N. J. On the role of words in the network structure of texts: application to authorship attribution. **arXiv**, 2017. Disponível em: <http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-JuolaEt2013.pdf>. Acesso em: 27 mar. 2017.
- 77 BRENNAN, M.; AFROZ, S.; GREENSTADT, R. Adversarial stylometry: circumventing authorship recognition to preserve privacy and anonymity. **ACM Transactions on Information and System Security (TISSEC)**, v. 15, n. 3, p. 12, 2012.

78 AFROZ, S.; BRENNAN, M.; GREENSTADT, R. **Detecting hoaxes, frauds, and deception in writing style online**. Disponível em: <<http://www1.icsi.berkeley.edu/~sadia/papers/oakland-deception.pdf>>. Acesso em: 27 mar. 2017.

Apêndices

APÊNDICE A – CORPORA DE TEXTOS

A continuação listam-se os textos estudados no trabalho organizados pelo autor junto com um hiper-vínculo para a sua consulta da base de dados aberta www.gutenberg.org. Na primeira parte do trabalho (capítulo 6) somente o primeiro corpus foi usado enquanto na segunda parte (capítulo 7) foram usados os três corpora. Cada corpus consta de 80 textos tendo 8 autores e 10 textos por autor.

A.1 Primeiro corpus

Arthur Conan Doyle

A Study in Scarlet	gutenberg.org/ebooks/244
His Last Bow: An Epilogue of Sherlock Holmes	gutenberg.org/ebooks/2350
The Adventures of Sherlock Holmes	gutenberg.org/ebooks/1661
The Gully of Bluemansdyke	gutenberg.org/ebooks/40848
The Hound of the Baskervilles	gutenberg.org/ebooks/2852
The Lost World	gutenberg.org/ebooks/139
The Memoirs of Sherlock Holmes	gutenberg.org/ebooks/834
The Return of Sherlock Holmes	gutenberg.org/ebooks/221
The Sign of the Four	gutenberg.org/ebooks/2097
The Valley of Fear	gutenberg.org/ebooks/3289

Bernard Shaw

Major Barbara	gutenberg.org/ebooks/3790
Man and Superman: A Comedy and a Philosophy	gutenberg.org/ebooks/3328
Mrs. Warren's Profession	gutenberg.org/ebooks/1097
Pygmalion	gutenberg.org/ebooks/3825
Caesar and Cleopatra	gutenberg.org/ebooks/3329
Cashel Byron's Profession	gutenberg.org/ebooks/5872
John Bull's Other Island	gutenberg.org/ebooks/3612
An Unsocial Socialist	gutenberg.org/ebooks/1654
On the Prospects of Christianity	gutenberg.org/ebooks/4004
Treatise on Parents and Children	gutenberg.org/ebooks/908

Fonte: Elaborada pelo autor.

Fyodor Dostoyevsky

Short Stories	gutenberg.org/ebooks/40745
Crime and Punishment	gutenberg.org/ebooks/2554
Notes from the Underground	gutenberg.org/ebooks/600
Poor Folk	gutenberg.org/ebooks/2302
The Brothers Karamazov	gutenberg.org/ebooks/28054
The Gambler	gutenberg.org/ebooks/2197
The Idiot	gutenberg.org/ebooks/2638
The Possessed (The Devils)	gutenberg.org/ebooks/8117
Uncle's Dream; and The Permanent Husband	gutenberg.org/ebooks/38241
White Nights and Other Stories	gutenberg.org/ebooks/36034

Herman Melville

Mardi: and A Voyage Thither, Vol. I	gutenberg.org/ebooks/13720
Mardi: and A Voyage Thither, Vol. II	gutenberg.org/ebooks/13721
Moby Dick; Or, The Whale	gutenberg.org/ebooks/2701
Omoo: Adventures in the South Seas	gutenberg.org/ebooks/4045
Pierre; or The Ambiguities	gutenberg.org/ebooks/34970
Redburn. His First Voyage	gutenberg.org/ebooks/8118
The Confidence-Man: His Masquerade	gutenberg.org/ebooks/21816
The Piazza Tales	gutenberg.org/ebooks/15859
Typee: A Romance of the South Seas	gutenberg.org/ebooks/1900
White Jacket; Or, The World on a Man-of-War	gutenberg.org/ebooks/10712

Jack London

Burning Daylight	gutenberg.org/ebooks/746
Martin Eden	gutenberg.org/ebooks/1056
Michael, Brother of Jerry	gutenberg.org/ebooks/1730
Smoke Bellew	gutenberg.org/ebooks/1596
The Iron Heel	gutenberg.org/ebooks/1164
The Jacket (The Star-Rover)	gutenberg.org/ebooks/1162
The Little Lady of the Big House	gutenberg.org/ebooks/6455
The Mutiny of the Elsinore	gutenberg.org/ebooks/2415
The Sea-Wolf	gutenberg.org/ebooks/1074
The Valley of the Moon	gutenberg.org/ebooks/1449

Fonte: Elaborada pelo autor.

Jonathan Swift

Gulliver's Travels into Several Remote Nations of the World	gutenberg.org/ebooks/829
The Journal to Stella	gutenberg.org/ebooks/4208
The Poems of Jonathan Swift, D.D., Volume 1	gutenberg.org/ebooks/14353
The Poems of Jonathan Swift, D.D., Volume 2	gutenberg.org/ebooks/13621
The Prose Works of Jonathan Swift, D.D. Volume 03	gutenberg.org/ebooks/12252
The Prose Works of Jonathan Swift, D.D. Volume 04	gutenberg.org/ebooks/12746
The Prose Works of Jonathan Swift, D.D. Volume 06	gutenberg.org/ebooks/12784
The Prose Works of Jonathan Swift, D.D. Volume 07	gutenberg.org/ebooks/18250
The Prose Works of Jonathan Swift, D.D. Volume 09	gutenberg.org/ebooks/13169
The Prose Works of Jonathan Swift, D.D. Volume 10	gutenberg.org/ebooks/13040

Leo Tolstoy

Anna Karenina	gutenberg.org/ebooks/1399
Redemption and two other plays	gutenberg.org/ebooks/9792
Resurrection	gutenberg.org/ebooks/1938
The Awakening	gutenberg.org/ebooks/17352
The Cossacks: A Tale of 1852	gutenberg.org/ebooks/4761
The Kingdom of God Is Within You	gutenberg.org/ebooks/4602
War and Peace	gutenberg.org/ebooks/2600
What Shall We Do?	gutenberg.org/ebooks/38690
What to Do? Thoughts Evoked by the Census of Moscow	gutenberg.org/ebooks/3630
Youth	gutenberg.org/ebooks/2637

Nathaniel Hawthorne

True Stories of History and Biography	gutenberg.org/ebooks/15697
The Scarlet Letter	gutenberg.org/ebooks/33
The Marble Faun; Or, The Romance of Monte Beni - Volume 2	gutenberg.org/ebooks/2182
Doctor Grimshawe's Secret	gutenberg.org/ebooks/7183
Passages from the English Notebooks, Complete	gutenberg.org/ebooks/7878
Passages from the French and Italian Notebooks, Volume 1	gutenberg.org/ebooks/7879
Passages from the French and Italian Notebooks, Volume 2	gutenberg.org/ebooks/7880
Our Old Home: A Series of English Sketches	gutenberg.org/ebooks/8090
Sketches and Studies	gutenberg.org/ebooks/8091
Twice Told Tales	gutenberg.org/ebooks/13707

Fonte: Elaborada pelo autor.

A.2 Segundo corpus

Charles Dickens

Great Expectations	gutenberg.org/ebooks/1400
Oliver Twist	gutenberg.org/ebooks/730
The Uncommercial Traveller	gutenberg.org/ebooks/914
A Tale of Two Cities	gutenberg.org/ebooks/98
Hard Times	gutenberg.org/ebooks/786
The Mystery of Edwin Drood	gutenberg.org/ebooks/564
Master Humphrey's Clock	gutenberg.org/ebooks/588
The Haunted Man and the Ghost's Bargain	gutenberg.org/ebooks/644
A Christmas Carol	gutenberg.org/ebooks/19337
Barnaby Rudge: A Tale of the Riots of 'Eighty	gutenberg.org/ebooks/917

Henry James

The Golden Bowl	gutenberg.org/ebooks/4264
The Ambassadors	gutenberg.org/ebooks/432
The American	gutenberg.org/ebooks/177
What Maisie Knew	gutenberg.org/ebooks/7118
Washington Square	gutenberg.org/ebooks/2870
The Turn of the Screw	gutenberg.org/ebooks/209
The Aspern Papers	gutenberg.org/ebooks/211
In the Cage	gutenberg.org/ebooks/1144
Tragic Muse	gutenberg.org/ebooks/20085
Awkward Age	gutenberg.org/ebooks/7433

Herbert George Wells

Tono-Bungay	gutenberg.org/ebooks/718
The War in the Air	gutenberg.org/ebooks/780
A Modern Utopia	gutenberg.org/ebooks/6424
When the Sleeper Wakes	gutenberg.org/ebooks/775
The First Men in the Moon	gutenberg.org/ebooks/1013
The World Set	gutenberg.org/ebooks/1059
The War of the Worlds	gutenberg.org/ebooks/36
The Island of Doctor Moreau	gutenberg.org/ebooks/159
The Time Machine	gutenberg.org/ebooks/35

Fonte: Elaborada pelo autor.

Louisa May Alcott

Little Women	gutenberg.org/ebooks/514
Work: A Story of Experience	gutenberg.org/ebooks/4770
Little Men	gutenberg.org/ebooks/2788
An Old-Fashioned Girl	gutenberg.org/ebooks/2787
Jo's Boys	gutenberg.org/ebooks/3499
Rose in Bloom	gutenberg.org/ebooks/2804
Jack and Jill	gutenberg.org/ebooks/2786
Moods	gutenberg.org/ebooks/28203
Silver Pitchers: and Independence, a Centennial Love Story	gutenberg.org/ebooks/34920
Under the Lilacs	gutenberg.org/ebooks/3795

Mark Twain

Roughing It	gutenberg.org/ebooks/3177
A Tramp Abroad	gutenberg.org/ebooks/119
Life on the Mississippi	gutenberg.org/ebooks/245
A Connecticut Yankee in King Arthur's Court	gutenberg.org/ebooks/86
Adventures of Huckleberry Finn	gutenberg.org/ebooks/76
The Prince and the Pauper	gutenberg.org/ebooks/1837
The Adventures of Tom Sawyer	gutenberg.org/ebooks/74
The American Claimant	gutenberg.org/ebooks/3179
The Innocents Abroad	gutenberg.org/ebooks/3176
Following the Equator: A Journey Around the World	gutenberg.org/ebooks/2895

Wilkie Collins

Hide and Seek	gutenberg.org/ebooks/7893
Poor Miss Finch	gutenberg.org/ebooks/3632
The Dead Secret: A Novel	gutenberg.org/ebooks/43092
The Queen of Hearts	gutenberg.org/ebooks/1917
The Law and the Lady	gutenberg.org/ebooks/1622
After Dark	gutenberg.org/ebooks/1626
The Fallen Leaves	gutenberg.org/ebooks/7894
Blind Love	gutenberg.org/ebooks/7890
Basil	gutenberg.org/ebooks/4605

Fonte: Elaborada pelo autor.

William Henry Hudson

Green Mansions: A Romance of the Tropical Forest	gutenberg.org/ebooks/942
A Shepherd's Life: Impressions of the South Wiltshire Downs	gutenberg.org/ebooks/7415
Afoot in England	gutenberg.org/ebooks/5406
Birds and Man	gutenberg.org/ebooks/37787
Birds in Town & Village	gutenberg.org/ebooks/7353
A Traveller in Little Things	gutenberg.org/ebooks/7982
A Crystal Age	gutenberg.org/ebooks/7401
Fan : The Story of a Young Girl's Life	gutenberg.org/ebooks/7827
Far Away and Long Ago: A History of My Early Life	gutenberg.org/ebooks/6093
The Land's End	gutenberg.org/ebooks/47990

William Makepeace Thackeray

Burlesques	gutenberg.org/ebooks/2675
Barry Lyndon	gutenberg.org/ebooks/4558
Roundabout Papers	gutenberg.org/ebooks/2608
The Christmas Books of Mr. M.A. Titmarsh	gutenberg.org/ebooks/2731
Catherine: A Story	gutenberg.org/ebooks/1969
Men's Wives	gutenberg.org/ebooks/1985
The Book of Snobs	gutenberg.org/ebooks/2686
Notes on a Journey from Cornhill to Grand Cairo	gutenberg.org/ebooks/1863
The Memoirs of Mr. Charles J. Yellowplush	gutenberg.org/ebooks/2796
Vanity Fair	gutenberg.org/ebooks/599

Fonte: Elaborada pelo autor.

A.3 Terceiro corpus

Edgar Rice Burroughs

Tarzan of the Apes	.gutenberg.org/ebooks/78
The Return of Tarzan	.gutenberg.org/ebooks/81
The Beasts of Tarzan	.gutenberg.org/ebooks/85
The Son of Tarzan	.gutenberg.org/ebooks/90
A Princess of Mars	.gutenberg.org/ebooks/62
The Gods of Mars	.gutenberg.org/ebooks/64
The Chessmen of Mars	.gutenberg.org/ebooks/1153
The Mucker	.gutenberg.org/ebooks/331
The Mad King	.gutenberg.org/ebooks/364
The Outlaw of Torn	.gutenberg.org/ebooks/369

Edith Wharton

The Valley of Decision	gutenberg.org/ebooks/4327
The Fruit of the Tree	gutenberg.org/ebooks/19191
The Custom of the Country	gutenberg.org/ebooks/11052
The House of Mirth	gutenberg.org/ebooks/284
The Age of Innocence	gutenberg.org/ebooks/541
Tales of Men and Ghosts	gutenberg.org/ebooks/4514
The Reef	gutenberg.org/ebooks/283
The Glimpses of the Moon	gutenberg.org/ebooks/1263
The Descent of Man and Other Stories	gutenberg.org/ebooks/4519
The Hermit and the Wild Woman, and Other Stories	gutenberg.org/ebooks/4533

George Eliot

Middlemarch	gutenberg.org/ebooks/145
Daniel Deronda	gutenberg.org/ebooks/7469
Romola	gutenberg.org/ebooks/24020
Adam Bede	gutenberg.org/ebooks/507
The Mill on the Floss	gutenberg.org/ebooks/6688
Felix Holt, the Radical	gutenberg.org/ebooks/40882
Scenes of Clerical Life	gutenberg.org/ebooks/17780
Silas Marner	gutenberg.org/ebooks/550
Impressions of Theophrastus Such	gutenberg.org/ebooks/10762

Fonte: Elaborada pelo autor.

Joseph Conrad

Nostromo: A Tale of the Seaboard	gutenberg.org/ebooks/2021
Chance: A Tale in Two Parts	gutenberg.org/ebooks/1476
Lord Jim	gutenberg.org/ebooks/5658
Victory: An Island Tale	gutenberg.org/ebooks/6378
Under Western Eyes	gutenberg.org/ebooks/2480
An Outcast of the Islands	gutenberg.org/ebooks/638
The Secret Agent: A Simple Tale	gutenberg.org/ebooks/974
A Set of Six	gutenberg.org/ebooks/2305
Almayer's Folly: A Story of an Eastern River	gutenberg.org/ebooks/720
Tales of Unrest	gutenberg.org/ebooks/1202

Rudyard Kipling

From Sea to Sea; Letters of Travel	gutenberg.org/ebooks/32977
Indian Tales	gutenberg.org/ebooks/8649
A Diversity of Creatures	gutenberg.org/ebooks/13085
The Day's Work - Volume 1	gutenberg.org/ebooks/2569
Kim	gutenberg.org/ebooks/2226
Life's Handicap: Being Stories of Mine Own People	gutenberg.org/ebooks/5777
Traffics and Discoveries	gutenberg.org/ebooks/9790
Soldiers Three	gutenberg.org/ebooks/6120
Plain Tales from the Hills	gutenberg.org/ebooks/1858
Rewards and Fairies	gutenberg.org/ebooks/556

Thomas Hardy

The Dynasts: An Epic-Drama of the War with Napoleon	gutenberg.org/ebooks/4043
Tess of the d'Urbervilles: A Pure Woman	gutenberg.org/ebooks/110
Jude the Obscure	gutenberg.org/ebooks/153
Desperate Remedies	gutenberg.org/ebooks/3044
The Hand of Ethelberta: A Comedy in Chapters	gutenberg.org/ebooks/3469
The Return of the Native	gutenberg.org/ebooks/122
A Laodicean : A Story of To-day	gutenberg.org/ebooks/3258
Far from the Madding Crowd	gutenberg.org/ebooks/27
The Woodlanders	gutenberg.org/ebooks/482

Fonte: Elaborada pelo autor.

William Dean Howells

Familiar Spanish Travels	gutenberg.org/ebooks/7430
Venetian Life	gutenberg.org/ebooks/7083
Italian Journeys	gutenberg.org/ebooks/14276
The Rise of Silas Lapham	gutenberg.org/ebooks/154
A Hazard of New Fortunes	gutenberg.org/ebooks/4600
Literature and Life	gutenberg.org/ebooks/3389
The Landlord at Lion's Head	gutenberg.org/ebooks/4645
A Modern Instance	gutenberg.org/ebooks/8203
April Hopes	gutenberg.org/ebooks/3404
The Quality of Mercy	gutenberg.org/ebooks/30108

Walter Scott

The Heart of Mid-Lothian, Complete	gutenberg.org/ebooks/6944
Peveril of the Peak	gutenberg.org/ebooks/5959
Waverley; Or, 'Tis Sixty Years Since	gutenberg.org/ebooks/5998
The Pirate	gutenberg.org/ebooks/42389
The Fortunes of Nigel	gutenberg.org/ebooks/5950
Woodstock; or, the Cavalier	gutenberg.org/ebooks/9785
Rob Roy Complete	gutenberg.org/ebooks/7025
Ivanhoe: A Romance	gutenberg.org/ebooks/82
Kenilworth	gutenberg.org/ebooks/1606
Redgauntlet: A Tale Of The Eighteenth Century	gutenberg.org/ebooks/2516

Fonte: Elaborada pelo autor.

APÊNDICE B – RESULTADOS DOS TESTES

As características das séries temporais podem ser avaliadas com os chamados testes de raiz unitária. Estes testes assumem uma hipótese nula e avaliam se a hipótese pode ser rejeitada em cujo caso se considera uma hipótese alternativa. Cada um dos testes usados tem tanto hipóteses nula como hipóteses alternativas diferentes assim como diferentes considerações para aceitar ou rejeitar a hipótese. O teste de Kwiatkowski-Phillips-Schmidt-Shin (KPSS) é o único que considera estacionariedade com tendência como hipótese nula, sendo a presença de uma raiz unitária a hipótese alternativa. Para o teste de Phillips-Perron a hipótese nula é que a série é integrada de primeira ordem. Consideramos também o teste aumentado de Dickey-Fuller para o qual a hipótese nula é a presença de uma raiz unitária. Por último, McKinnon também considera a presença de uma raiz unitária como hipótese nula mas usa diferentes distribuições para estimar os p_{values} (colunas 5 e 6 na tabela 8 correspondem aos testes de amostra finita e assintótico respectivamente). Note-se que os resultados obtidos na tabela 8 são médias sobre a coleção de 80 livros ao invés de resultados para séries particulares.

Tabela 8 – Averages of p_{values} obtained with various tests.

	Phillips-Perron	KPSS	Dickey-Fuller	McKinnon
Clustering	0.010	0.071	0.017	0.008 0.167
Betweenness centrality	0.023	0.074	0.350	0.360 0.510
Cliques	0.010	0.086	0.377	0.393 0.521
Diameter	0.010	0.076	0.116	0.111 0.365
Intermittency	0.010	0.071	0.080	0.074 0.335
Load centrality	0.081	0.080	0.457	0.478 0.583
Degree	0.019	0.066	0.470	0.513 0.579
Radius	0.011	0.073	0.118	0.114 0.368
Shortest path	0.013	0.071	0.214	0.208 0.430
Edges	0.253	0.078	0.362	0.369 0.512
Nodes	0.022	0.067	0.368	0.378 0.514
Transitivity	0.010	0.083	0.014	0.005 0.126

Fonte: Elaborada pelo autor.

APÊNDICE C – AJUSTES COM O MODELO ARIMA

As $80 \times 12 = 960$ séries foram ajustadas como processos ARIMA(p,d,q). Cada tabela corresponde a um parâmetro e o valor em cada célula corresponde ao número de séries que foram ajustadas com o parâmetro sendo igual ao da coluna. A tabela 12 corresponde à soma $p + d + q$.

Tabela 9 – Número de séries ajustadas com um modelo ARIMA(p,d,q) para diferentes valores do parâmetro p .

Métrica de rede	Valor de p									
	0	1	2	3	4	5	6	7	8	9
Coefficiente de clusterização	36	27	11	5	1	0	0	0	0	0
Centralidade de intermediação	25	37	12	6	0	0	0	0	0	0
Cliques	59	11	7	2	1	0	0	0	0	0
Diâmetro	32	32	8	8	0	0	0	0	0	0
Intermitência	27	34	13	5	1	0	0	0	0	0
Centralidade de carga	35	28	9	7	0	1	0	0	0	0
Conetividade	28	36	11	3	1	1	0	0	0	0
Raio	27	32	14	7	0	0	0	0	0	0
Menor caminho	25	26	21	4	3	1	0	0	0	0
Número de arestas	40	25	8	6	1	0	0	0	0	0
Número de nós	27	33	13	5	1	1	0	0	0	0
Transitividade	51	20	6	1	2	0	0	0	0	0
Total	412	341	133	59	11	4	0	0	0	0

Fonte: Elaborada pelo autor.

Tabela 10 – Número de séries ajustadas com um modelo ARIMA(p,d,q) para diferentes valores do parâmetro d .

Métrica de rede	Valor de d									
	0	1	2	3	4	5	6	7	8	9
Coefficiente de clusterização	55	25	0	0	0	0	0	0	0	0
Centralidade de intermediação	57	23	0	0	0	0	0	0	0	0
Cliques	69	11	0	0	0	0	0	0	0	0
Diâmetro	60	20	0	0	0	0	0	0	0	0
Intermitência	56	24	0	0	0	0	0	0	0	0
Centralidade de carga	63	17	0	0	0	0	0	0	0	0
Conetividade	51	29	0	0	0	0	0	0	0	0
Raio	58	22	0	0	0	0	0	0	0	0
Menor caminho	55	25	0	0	0	0	0	0	0	0
Número de arestas	61	19	0	0	0	0	0	0	0	0
Número de nós	49	31	0	0	0	0	0	0	0	0
Transitividade	64	16	0	0	0	0	0	0	0	0
Total	698	262	0	0	0	0	0	0	0	0

Fonte: Elaborada pelo autor.

Tabela 11 – Número de séries ajustadas com um modelo ARIMA(p,d,q) para diferentes valores do parâmetro q .

Métrica de rede	Valor de q									
	0	1	2	3	4	5	6	7	8	9
Coefficiente de clusterização	31	29	13	6	1	0	0	0	0	0
Centralidade de intermediação	16	42	16	6	0	0	0	0	0	0
Cliques	53	16	10	1	0	0	0	0	0	0
Diâmetro	27	36	11	5	1	0	0	0	0	0
Intermitência	21	38	19	2	0	0	0	0	0	0
Centralidade de carga	32	33	10	4	1	0	0	0	0	0
Conetividade	21	44	12	2	0	1	0	0	0	0
Raio	27	32	18	3	0	0	0	0	0	0
Menor caminho	23	40	10	6	1	0	0	0	0	0
Número de arestas	35	29	10	4	1	1	0	0	0	0
Número de nós	26	40	11	3	0	0	0	0	0	0
Transitividade	39	27	13	1	0	0	0	0	0	0
Total	351	406	153	43	5	2	0	0	0	0

Fonte: Elaborada pelo autor.

Tabela 12 – Número de séries ajustadas com um modelo ARIMA(p,d,q) para diferentes valores da soma $p + d + q$.

Métrica de rede	Valor de $p + d + q$									
	0	1	2	3	4	5	6	7	8	9
Coefficiente de clusterização	20	15	16	10	12	2	2	1	2	0
Centralidade de intermediação	9	9	27	19	9	4	3	0	0	0
Cliques	42	14	11	5	6	2	0	0	0	0
Diâmetro	13	20	20	11	9	4	1	2	0	0
Intermitência	8	20	17	17	11	6	1	0	0	0
Centralidade de carga	18	19	17	14	4	5	1	1	0	1
Conetividade	8	16	24	16	10	5	0	1	0	0
Raio	13	16	20	12	11	5	2	1	0	0
Menor caminho	8	18	20	11	12	3	6	2	0	0
Número de arestas	21	21	12	10	9	4	1	1	1	0
Número de nós	9	19	21	13	10	4	3	1	0	0
Transitividade	32	16	13	9	7	1	1	1	0	0
Total	201	203	218	147	110	45	21	11	3	1

Fonte: Elaborada pelo autor.

Tabela 13 – Séries ajustadas com um modelo ARIMA(p,d,q) que possuem o maior valor da soma $p + d + q$.

Livro	Medida	Soma	ARIMA		
			p	d	q
The Poems of Jonathan Swift, D.D., Volume 2	Cent. de carga	9	5	0	4
The Journal to Stella	Clusterização	8	2	1	5
The Iron Heel	Clusterização	8	3	1	4
Typee: A Romance of the South Seas	No. de arestas	8	4	1	3

Fonte: Elaborada pelo autor.

APÊNDICE D – TAXAS DE ACERTO USANDO ISOMAP

Os resultados da classificação após usar Isomap para os diferentes algoritmos de classificação são apresentados em forma de histogramas na figura 44. Cada figura contém os 196 casos obtidos variando $n_{componentes}$ e $n_{vizinhos}$ de 2 até 15. Para todos os algoritmos, os histogramas obtidos estão inclinados para o lado direito mostrando que para a maioria dos casos se obtém uma alta taxa de acerto. O valor máximo foi de 88.75% e foi obtido usando o KNN como mostrado na figura 44b. Para este algoritmo de classificação os resultados detalhados são apresentados na tabela 14 na página seguinte.

(a) Taxas de acerto do J48.(%).

(b) Taxas de acerto do KNN.(%).

(c) Taxas de acerto do Naive Bayes(%).

(d) Taxas de acerto do RBFNetwork(%).

Figura 44 – Taxas de acerto aplicando Isomap. Os parâmetros $n_{componentes}$ e $n_{vizinhos}$ do Isomap foram variados desde 2 até 15 gerando 196 taxas de acerto. Os histogramas mostram um comportamento similar para a maioria dos valores dos parâmetros com um pico no lado direito de cada figura e portanto em valores altos da taxa de acerto.

Fonte: Elaborada pelo autor.

Tabela 14 – Taxas de acerto obtidas com KNN aplicando Isomap para vários valores dos parâmetros $n_{componentes}$ e $n_{vizinhos}$.

n_c	n_n	Score	n_c	n_n	Score	n_c	n_n	Score	n_c	n_n	Score	n_c	n_n	Score
2	2	53.75	5	2	65.00	8	2	68.75	11	2	65.00	14	2	62.50
2	3	72.50	5	3	68.75	8	3	75.00	11	3	77.50	14	3	77.50
2	4	48.75	5	4	55.00	8	4	63.75	11	4	66.25	14	4	68.75
2	5	52.50	5	5	65.00	8	5	76.25	11	5	77.50	14	5	61.25
2	6	65.00	5	6	75.00	8	6	68.75	11	6	73.75	14	6	73.75
2	7	58.75	5	7	73.75	8	7	83.75	11	7	80.00	14	7	77.50
2	8	47.50	5	8	76.25	8	8	73.75	11	8	72.50	14	8	73.75
2	9	45.00	5	9	68.75	8	9	71.25	11	9	73.75	14	9	73.75
2	10	48.75	5	10	77.50	8	10	78.75	11	10	80.00	14	10	82.50
2	11	43.75	5	11	71.25	8	11	75.00	11	11	73.75	14	11	72.50
2	12	43.75	5	12	75.00	8	12	70.00	11	12	72.50	14	12	70.00
2	13	40.00	5	13	72.50	8	13	72.50	11	13	71.25	14	13	65.00
2	14	50.00	5	14	63.75	8	14	61.25	11	14	58.75	14	14	57.50
2	15	43.75	5	15	65.00	8	15	65.00	11	15	66.25	14	15	66.25
3	2	61.25	6	2	62.50	9	2	72.50	12	2	65.00	15	2	65.00
3	3	71.25	6	3	73.75	9	3	76.25	12	3	75.00	15	3	80.00
3	4	51.25	6	4	56.25	9	4	68.75	12	4	66.25	15	4	68.75
3	5	55.00	6	5	65.00	9	5	75.00	12	5	68.75	15	5	62.50
3	6	67.50	6	6	71.25	9	6	73.75	12	6	75.00	15	6	72.50
3	7	73.75	6	7	77.50	9	7	80.00	12	7	80.00	15	7	77.50
3	8	78.75	6	8	77.50	9	8	75.00	12	8	71.25	15	8	75.00
3	9	70.00	6	9	72.50	9	9	72.50	12	9	71.25	15	9	75.00
3	10	80.00	6	10	78.75	9	10	78.75	12	10	76.25	15	10	75.00
3	11	70.00	6	11	70.00	9	11	78.75	12	11	73.75	15	11	70.00
3	12	71.25	6	12	75.00	9	12	72.50	12	12	68.75	15	12	73.75
3	13	73.75	6	13	73.75	9	13	75.00	12	13	68.75	15	13	71.25
3	14	55.00	6	14	62.50	9	14	60.00	12	14	61.25	15	14	58.75
3	15	56.25	6	15	65.00	9	15	68.75	12	15	65.00	15	15	67.50
4	2	60.00	7	2	63.75	10	2	73.75	13	2	63.75			
4	3	68.75	7	3	73.75	10	3	77.50	13	3	77.50			
4	4	62.50	7	4	58.75	10	4	67.50	13	4	71.25			
4	5	66.25	7	5	75.00	10	5	73.75	13	5	61.25			
4	6	73.75	7	6	68.75	10	6	73.75	13	6	73.75			
4	7	76.25	7	7	78.75	10	7	80.00	13	7	77.50			
4	8	77.50	7	8	73.75	10	8	70.00	13	8	72.50			
4	9	71.25	7	9	72.50	10	9	72.50	13	9	73.75			
4	10	76.25	7	10	77.50	10	10	88.75	13	10	82.50			
4	11	67.50	7	11	68.75	10	11	78.75	13	11	73.75			
4	12	71.25	7	12	75.00	10	12	73.75	13	12	70.00			
4	13	72.50	7	13	73.75	10	13	73.75	13	13	66.25			
4	14	56.25	7	14	63.75	10	14	60.00	13	14	60.00			
4	15	60.00	7	15	65.00	10	15	68.75	13	15	66.25			

Fonte: Elaborada pelo autor.

APÊNDICE E – OBTENÇÃO DAS RELAÇÕES PARA MDS

Na condição 5.22, elevamos ao quadrado ambos lados da equação e considerando que $b_{ij} = \sum_k x_{ik}x_{jk}$

$$\begin{aligned}
 d_{ij}^2 &= |x_i - x_j|^2 \\
 &= x_i \cdot x_i - x_i \cdot x_j - x_j \cdot x_i + x_j \cdot x_j \\
 &= b_{ii} + b_{jj} - b_{ij} - b_{ji} \\
 &= b_{ii} + b_{jj} - 2b_{ij},
 \end{aligned} \tag{E.1}$$

obtendo a equação 5.24. Primeiro verificamos o que acontece com o último termo em 5.24 quando uma somatória é feita,

$$\begin{aligned}
 \sum_{i=1}^n b_{ij} &= \sum_{i=1}^n \sum_{k=1}^q x_{ik}x_{jk} \\
 &= \sum_{k=1}^q x_{jk} \sum_{i=1}^n x_{ik} = 0,
 \end{aligned} \tag{E.2}$$

para todo j , o mesmo resultado é obtido trocando i e j . Usando o resultado acima aplicamos somatórias sobre diferentes índices em 5.24 lembrando a definição da traça $T = \sum_{i=1}^n b_{ii}$. Primeiro somamos sobre i :

$$\begin{aligned}
 \sum_{i=1}^n d_{ij}^2 &= T + \sum_{i=1}^n b_{jj} \\
 &= T + nb_{jj}.
 \end{aligned} \tag{E.3}$$

Somando sobre j obtemos:

$$\begin{aligned}
 \sum_{j=1}^n d_{ij}^2 &= \sum_{j=1}^n b_{ii} + T \\
 &= nb_{ii} + T.
 \end{aligned} \tag{E.4}$$

E finalmente sobre ambos índices, para isto somamos sobre j o resultado de E.3,

$$\begin{aligned}
 \sum_{j=1}^n \sum_{i=1}^n d_{ij}^2 &= \sum_{j=1}^n T + \sum_{j=1}^n nb_{jj} \\
 &= nT + nT = 2nT.
 \end{aligned} \tag{E.5}$$

Juntando as equações E.3, E.4, e E.5 obtemos 5.25. Agora mostramos que a relação 5.26 permite obter uma expressão para o termo b_{ij} . Para isto desenvolvemos o ij -ésimo termo do lado direito de 5.26:

$$\begin{aligned}
 \left[-\frac{1}{2}HSH\right]_{ij} &= -\frac{1}{2} \sum_{kl} h_{ik} s_{kl} h_{lj} \\
 &= -\frac{1}{2} \sum_{kl} \left(\delta_{ik} - \frac{1}{n}\right) d_{kl}^2 \left(\delta_{lj} - \frac{1}{n}\right) \\
 &= -\frac{1}{2} \sum_{kl} \left(\delta_{ik} d_{kl}^2 \delta_{lj} - \delta_{ik} d_{kl}^2 \frac{1}{n} - \frac{1}{n} d_{kl}^2 \delta_{lj} + d_{kl}^2 \frac{1}{n^2}\right) \\
 &= -\frac{1}{2} \left(d_{ij}^2 - \frac{1}{n} \sum_l d_{il}^2 - \frac{1}{n} \sum_k d_{kj}^2 + \frac{1}{n^2} \sum_{kl} d_{kl}^2\right).
 \end{aligned} \tag{E.6}$$

Substituindo as expressões E.3, E.4, e E.5 obtemos:

$$\begin{aligned}
 \left[-\frac{1}{2}HSH\right]_{ij} &= -\frac{1}{2} \left(d_{ij}^2 - \frac{1}{n} (T + nb_{ii}) - \frac{1}{n} (T + nb_{jj}) + \frac{1}{n^2} 2nT\right) \\
 &= -\frac{1}{2} (d_{ij}^2 - b_{ii} - b_{jj}),
 \end{aligned} \tag{E.7}$$

que, segundo a equação 5.24 é igual a b_{ij} , demonstrando assim a equação 5.26.

Anexos

ANEXO A – LISTA DE PART-OF-SPEECH TAGS USADOS NO NLTK

Listam-se os tags usados pelo pacote Natural Language Toolkit–NLTK (www.nltk.org). Eles podem ser obtidos da linha de comandos de Python digitando:

```
nltk.help.upenn_tagset()
```

A.1 Part-of-Speech tags

```
$: dollar
    $ -$ --$ A$ C$ HK$ M$ NZ$ S$ U.S.$ US$
'': closing quotation mark
    , ''
(: opening parenthesis
    ( [ {
): closing parenthesis
    ) ] }
,: comma
    ,
--: dash
    --
.: sentence terminator
    . ! ?
:: colon or ellipsis
    : ; ...
CC: conjunction, coordinating
    & 'n and both but either et for less minus neither nor or plus so
    therefore times v. versus vs. whether yet
CD: numeral, cardinal
    mid-1890 nine-thirty forty-two one-tenth ten million 0.5 one forty-
    seven 1987 twenty '79 zero two 78-degrees eighty-four IX '60s .025
    fifteen 271,124 dozen quintillion DM2,000 ...
DT: determiner
    all an another any both del each either every half la many much nary
    neither no some such that the them these this those
EX: existential there
    there
FW: foreign word
```

gemeinschaft hund ich jeux habeas Haementeria Herr K'ang-si vous
lutihaw alai je jour objets salutaris fille quibusdam pas trop Monte
terram fiche oui corporis ...

IN: preposition or conjunction, subordinating

astride among uppon whether out inside pro despite on by throughout
below within for towards near behind atop around if like until below
next into if beside ...

JJ: adjective or numeral, ordinal

third ill-mannered pre-war regrettable oiled calamitous first separable
ectoplasmic battery-powered participatory fourth still-to-be-named
multilingual multi-disciplinary ...

JJR: adjective, comparative

bleaker braver breezier briefer brighter brisker broader bumper busier
calmer cheaper choosier cleaner clearer closer colder commoner costlier
cozier creamier crunchier cuter ...

JJS: adjective, superlative

calmest cheapest choicest classiest cleanest clearest closest commonest
corniest costliest crassest creepiest crudest cutest darkest deadliest
dearest deepest densest dinkiest ...

LS: list item marker

A A. B B. C C. D E F First G H I J K One SP-44001 SP-44002 SP-44005
SP-44007 Second Third Three Two * a b c d first five four one six three
two

MD: modal auxiliary

can cannot could couldn't dare may might must need ought shall should
shouldn't will would

NN: noun, common, singular or mass

common-carrier cabbage knuckle-duster Casino afghan shed thermostat
investment slide humour falloff slick wind hyena override subhumanity
machinist ...

NNP: noun, proper, singular

Motown Venneboerger Czestochwa Ranzer Conchita Trumplane Christos
Oceanside Escobar Kreisler Sawyer Cougar Yvette Ervin ODI Darryl CTCA
Shannon A.K.C. Meltex Liverpool ...

NNPS: noun, proper, plural

Americans Americas Amharas Amityvilles Amusements Anarcho-Syndicalists
Andalusians Andes Andruses Angels Animals Anthony Antilles Antiques
Apache Apaches Apocrypha ...

NNS: noun, common, plural

undergraduates scotches bric-a-brac products bodyguards facets coasts
divestitures storehouses designs clubs fragrances averages
subjectivists apprehensions muses factory-jobs ...

PDT: pre-determiner

all both half many quite such sure this

POS: genitive marker

’ ’ S

PRP: pronoun, personal

hers herself him himself hisself it itself me myself one oneself ours
ourselves ownself self she thee theirs them themselves they thou thy us

PRP\$: pronoun, possessive

her his mine my our ours their thy your

RB: adverb

occasionally unabatingly maddeningly adventurously professedly
stirringly prominently technologically magisterially predominately
swiftly fiscally pitilessly ...

RBR: adverb, comparative

further gloomier grander graver greater grimmer harder harsher
healthier heavier higher however larger later leaner lengthier less-
perfectly lesser lonelier longer louder lower more ...

RBS: adverb, superlative

best biggest bluntest earliest farthest first furthest hardest
heartiest highest largest least less most nearest second tightest worst

RP: particle

aboard about across along apart around aside at away back before behind
by crop down ever fast for forth from go high i.e. in into just later
low more off on open out over per pie raising start teeth that through
under unto up up-pp upon whole with you

SYM: symbol

```
% & ' ' ' ' . ) ). * + , . < = > @ A[fj] U.S U.S.S.R * ** ***
```

T0: "to" as preposition or infinitive marker

to

UH: interjection

Goodbye Goody Gosh Wow Jeepers Jee-sus Hubba Hey Kee-reist Oops amen
huh howdy uh dammit whammo shucks heck anyways whodunnit honey golly
man baby diddle hush sonuvabitch ...

VB: verb, base form

ask assemble assess assign assume atone attention avoid bake balkanize
bank begin behold believe bend benefit bevel beware bless boil bomb

boost brace break bring broil brush build ...

VBD: verb, past tense

dipped pleaded swiped regummed soaked tidied convened halted registered
cushioned exacted snubbed strode aimed adopted belied figgered
speculated wore appreciated contemplated ...

VBG: verb, present participle or gerund

telegraphing stirring focusing angering judging stalling lactating
hankerin' alleging veering capping approaching traveling besieging
encrypting interrupting erasing wincing ...

VCN: verb, past participle

multihulled dilapidated aerosolized chaired languished panelized used
experimented flourished imitated reunified factored condensed sheared
unsettled primed dubbed desired ...

VBP: verb, present tense, not 3rd person singular

predominate wrap resort sue twist spill cure lengthen brush terminate
appear tend stray glisten obtain comprise detest tease attract
emphasize mold postpone sever return wag ...

VBZ: verb, present tense, 3rd person singular

bases reconstructs marks mixes displeases seals carps weaves snatches
slumps stretches authorizes smolders pictures emerges stockpiles
seduces fizzes uses bolsters slaps speaks pleads ...

WDT: WH-determiner

that what whatever which whichever

WP: WH-pronoun

that what whatever whatsoever which who whom whosoever

WP\$: WH-pronoun, possessive

whose

WRB: Wh-adverb

how however whence whenever where whereby wherever wherein whereof why

“: opening quotation mark

‘ ‘

ANEXO B – ELEMENTOS REMOVIDOS

Tabela 15 – Palavras retiradas dos textos (primeira parte).

a	about	above	across	after	afterwards
again	against	all	almost	alone	along
already	also	although	always	am	among
amongst	amongst	amount	an	and	another
any	anyhow	anyone	anything	anyway	anywhere
are	around	as	at	back	be
became	because	become	becomes	becoming	been
before	beforehand	behind	being	below	beside
besides	between	beyond	bill	both	bottom
but	by	call	can	cannot	cant
co	con	could	couldnt	cry	de
describe	detail	do	done	down	due
during	each	eg	eight	either	eleven
else	elsewhere	empty	enough	etc	even
ever	every	everyone	everything	everywhere	except
few	fifteen	fifty	fill	find	fire
first	five	for	former	formerly	forty
found	four	from	front	full	further
get	give	go	had	has	hasnt
have	he	hence	her	here	hereafter
hereby	herein	hereupon	hers	herself	him
himself	his	how	however	hundred	i
ie	if	in	inc	indeed	interest
into	is	it	its	itself	keep
last	latter	latterly	least	less	ltd
made	many	may	me	meanwhile	might

Fonte: Elaborada pelo autor.

Tabela 16 – Palavras retiradas dos textos (segunda parte).

mill	mine	more	moreover	most	mostly
move	much	must	my	myself	name
namely	neither	never	nevertheless	next	nine
no	nobody	none	noone	nor	not
nothing	now	nowhere	of	off	often
on	once	one	only	onto	or
other	others	otherwise	our	ours	ourselves
out	over	own	part	per	perhaps
please	put	rather	re	same	see
seem	seemed	seeming	seems	serious	several
she	should	show	side	since	sincere
six	sixty	so	some	somehow	someone
something	sometime	sometimes	somewhere	still	such
system	take	ten	than	that	the
their	them	themselves	then	thence	there
thereafter	thereby	therefore	therein	thereupon	these
they	thick	thin	third	this	those
though	three	through	throughout	thru	thus
to	together	too	top	toward	towards
twelve	twenty	two	un	under	until
up	upon	us	very	via	was
we	well	were	what	whatever	when
whence	whenever	where	whereafter	whereas	whereby
wherein	whereupon	wherever	whether	which	while
whither	who	whoever	whole	whom	whose
why	will	with	within	without	would
yet	you	your	yours	yourself	yourselves

Fonte: Elaborada pelo autor.

Tabela 17 – Símbolos retirados dos textos.

.	,	;	:	!	?	(	)	[	]
{	}	<	>	~	+	-	*	/	=
#	\$	%	&	@	"	'	_		

Fonte: Elaborada pelo autor.