UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO
DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA

ALLEF PABLO ARAUJO DA SILVA

Técnicas de classificação textual utilizando grafos

Ribeirão Preto–SP

2019

ALLEF PABLO ARAUJO DA SILVA

Técnicas de classificação textual utilizando grafos

Versão Corrigida

Versão original encontra-se na FFCLRP/USP.

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) da Universidade de São Paulo (USP), como parte das exigências para a obtenção do título de Mestre em Ciências.

Área de Concentração: Computação Aplicada.

Orientador: Prof. Dr. Alexandre Souto Martinez

Ribeirão Preto–SP

2019

Allef Pablo Araujo da Silva
Técnicas de classificação textual utilizando grafos. Ribeirão Preto–SP, 2019.
79p. : il.; 30 cm.

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras
de Ribeirão Preto da USP, como parte das exigências para
a obtenção do título de Mestre em Ciências,
Área: Computação Aplicada.

Orientador: Prof. Dr. Alexandre Souto Martinez

1. Grafos. 2. Redes Complexas. 3. Classificação textual.

Allef Pablo Araujo da Silva

Técnicas de classificação textual utilizando grafos

Modelo canônico de trabalho monográfico acadêmico em conformidade com as normas ABNT.

Aprovado em: Ribeirão Preto–SP, 15 de março de 2019.

Prof. Dr. Evandro Eduardo Seron Ruiz
(Substituto)

Prof. Dr. Evandro Marcos Saidel Ribeiro

Prof. Dr. Odemir Martinez Bruno

Prof. Dr. César Henrique Comin

Ribeirão Preto–SP
2019

Este trabalho é dedicado às crianças adultas que,
quando pequenas, sonharam em se tornar cientistas.

Agradecimentos

Agradecimentos especiais são direcionados ao meu orientador Prof. Dr. Alexandre Souto Martinez pela paciência e pelo suporte oferecido durante a elaboração deste trabalho, aos meus colegas do Laboratório de Modelagem de Sistemas Complexos pelas discussões instigantes, ao Prof. Dr. Evandro Eduardo Seron Ruiz por ter me guiado nos fundamentos do Processamento de Língua Natural e a Innowatt Desenvolvimento e Serviços LTDA por ter permitido que eu me ausentasse das minhas funções profissionais em diversos momentos durante os primeiros meses deste trabalho.

"Nenhum homem realmente produtivo pensa como se estivesse escrevendo uma dissertação."
(Albert Einstein)

Resumo

O grande volume de informação textual sendo gerado a todo momento torna necessário o aprimoramento constante de sistemas capazes de classificar textos em categorias específicas. Essa categorização visa, por exemplo, separar notícias indexadas por mecanismos de buscas, identificar a autoria de livros e cartas antigas ou detectar plágio em artigos científicos. As técnicas de classificação textual existentes, baseadas em conteúdo, apesar de conseguirem uma boa performance quantitativamente, ainda apresentam dificuldades em lidar com aspectos semânticos presentes nos textos escritos em língua natural. Neste sentido, abordagens alternativas vem sendo propostas, como as baseadas em redes complexas, que levam em consideração apenas o relacionamento entre as palavras. Neste estudo, aplicamos a modelagem de textos como redes complexas e utilizamos as métricas extraídas como atributos para classificação, utilizando um problema de reconhecimento de autoria para ilustrar a aplicação das técnicas descritas ao longo deste texto.

Palavras-chave: Grafos. Redes complexas. Classificação textual.

Abstract

The large volume of textual information being generated at all times makes it necessary to constantly improve systems capable of classifying texts into specific categories. This categorization aims, for example, to separate news items indexed by search engines, identify authorship of old books and letters, or detect plagiarism in scientific articles. Existing textual classification techniques, based on content, despite achieving good quantitative performance, still present difficulties in dealing with semantic aspects present in texts written in natural language. In this sense, alternative approaches have been proposed, such as those based on complex networks, which take into account only the relationship between words. In this study, we applied text modeling as graphs and extracted metrics typically used in the study of complex networks to be used as classifier attributes. To illustrate these techniques, a problem of authorship recognition in small texts was chosen as an example.

Keywords: Graph. Complex networks. Text categorization.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

PLN	Processamento de Língua Natural
SVM	*Support Vector Machine*
NILC	Núcleo Interinstitucional de Linguística Computacional

Sumário

Introdução

A linguagem é o mecanismo pelo qual os seres humanos expressam seus pensamentos e é a base sobre a qual se constroem todas as relações sociais. O entendimento dos processos de comunicação em língua natural representa um enorme desafio para diversas áreas do conhecimento como a Linguística, a Matemática, a Física e a Ciência da Computação. Neste sentido, diversos desafios em Processamento de Língua Natural (PLN) têm despertado a atenção da comunidade científica, dentre eles, os problemas relacionados a classificação textual.

É possível supor que desde que o homem passou a produzir textos em larga escala e as primeiras bibliotecas começaram a surgir na antiguidade, a necessidade de separar textos em categorias predefinidas se tornou evidente. Com o passar dos anos, o volume de informação textual gerada tem crescido de forma vertiginosa, particularmente devido a duas grandes revoluções, primeiro a invenção da prensa de Gutenberg e depois o surgimento da *World Wide Web*, tornando a necessidade de classificação cada vez maior.

Modernamente, a classificação automática de textos tem se tornado uma das principais áreas de pesquisa em PLN e aprendizado de máquina. Isto se deve ao grande número de aplicações, que vão desde a categorização de notícias indexadas por mecanismos de buscas, passando pela detecção de mensagens de *spam*, até o reconhecimento de autoria e detecção de plágio.

Dentre as possíveis abordagens para a classificação automática de textos em língua natural é possível citar os métodos *"frequencistas"*. Eles são fortemente dependentes de conteúdo, pois caracterizam documentos a partir da frequência com que as palavras aparecem nos textos. Em algumas situações, é possível atribuir pesos para palavras específicas possibilitando a identificação de termos importantes no texto. Estes métodos têm sido amplamente empregados nos sistemas modernos de classificação com grande sucesso. Há também os métodos baseados em grafos que levam em conta apenas as relações de adjacência entre as palavras e sua organização estrutural. Neste sentido, tem-se destacado os métodos baseados em redes complexas, que fazem uso intensivo de conceitos provenientes da mecânica estatística (ANTIQUEIRA et al., 2005) e representam uma nova fronteira ainda pouco estudada em PLN.Neste tipo de abordagem, usualmente os vértices da rede são palavras e as arestas são estabelecidas entre palavras que são vizinhas no texto.

Assim, a técnica possibilita a captura de informações relacionadas a maneira com que as palavras interagem no texto.

A motivação para a aplicação de técnicas de redes complexas vem do fato de que as redes formadas por palavras apresentam várias propriedades topológicas que são compartilhadas por diversas redes do mundo real. Logo, as métricas utilizadas comumente para a caracterização de redes complexas, podem ser empregadas como atributos para a caracterização de textos em língua natural. Além disso, a modelagem de redes permite a captura da inter-relação entre as palavras, ao contrário de algumas abordagens tradicionais que utilizam o conceito de *bag of words*. Estes métodos não levam em consideração os relacionamentos entre as palavras, apenas a frequência com que são usadas.

Amancio (AMANCIO, 2013) demonstrou que a aplicação da metodologia baseada em redes produz resultados promissores quando aplicada a problemas de estilografia, diferenciação entre textos e sequências aleatórias de caracteres, e outros. Neste trabalho utilizaremos esta técnica para o problema de reconhecimento de autoria em discursos presidenciais escritos em língua portuguesa.

Discursos presidenciais possuem particularidades, como por exemplo, o fato de que nem sempre eles são redigidos pelo presidente em questão e muitas vezes vários autores trabalham no mesmo texto, porém estes textos devem guardar uma consistência de estilo para cada um dos presidentes analisados. Um dos objetivos deste texto é verificar se a técnica é aplicável neste tipo de tarefa, além de verificar se variações na forma como a rede é construída trazem alguma melhora em relação ao desempenho de classificadores.

Também são realizados experimentos incluindo autovalores do espectro Laplaciano do grafo como atributos de classificação. Este método foi utilizado com sucesso para a geração de atributos para a classificação de imagens (HUMARI, 2016).

No capítulo 1 faremos uma revisão da teoria básica de grafos e redes complexas apresentando as principais definições utilizadas ao longo do texto. Neste capítulo apresentaremos também as formas usuais de representação de grafos em computador e alguns fudamentos da teoria espectral de grafos. Também será introduzido o conceito de redes complexas, com os modelos básicos de redes e a apresentação das métricas utilizadas para a sua caracterização.

No capítulo 2 faremos uma introdução ao Processamento de Linguagem Natural, com foco em categorização/classificação. Apresentaremos algumas formas clássicas de representação de textos e um método de classificação básico baseado no teorema de Bayes. Este capítulo é encerrado com a representação de textos utilizando grafos e algumas de suas propriedades.

No capítulo 3 é apresentada a metodologia aplicada em detalhes. Aqui são especificadas as etapas de obtenção e pré-processamento dos dados, construção da rede,

classificação e avaliação dos modelos preditivos.

No capítulo 4 são apresentados os resultados obtidos em todos os experimentos, com o desempenho de cada um dos classificadores para cada um dos conjuntos de dados.

No capítulo 5 discutiremos os resultados obtidos no capítulo 4.

1

Redes Complexas

Neste capítulo apresentamos alguns conceitos fundamentais relacionados a grafos e redes complexas. Nas seções a seguir apresentamos uma breve nota histórica a respeito do surgimento da teoria dos grafos seguida pelas definições básicas utilizadas neste texto. Também são discutidas as principais formas de representação computacional de grafos e são definidos os conceitos de **espectro** e **espectro laplaciano** de um grafo. Após a apresentação dos conceitos gerais, são discutidos os aspectos fundamentais do estudo de redes complexas, incluindo modelos de redes e métricas. O capítulo é finalizado com a apresentação das redes construídas a partir de textos em língua natural, que são o foco deste estudo.

1.1 Conceitos fundamentais sobre grafos

O primeiro estudo do que se conhece hoje como **Teoria dos Grafos** foi publicado em 1741 pelo matemático Leonhard Euler em um artigo intitulado "*Solutio problematis ad geometriam situs pertinentis*" (EULER, 1741). Nesse estudo, Euler apresenta uma solução para o famoso problema das sete pontes de Königsberg. Ele mostra que é impossível para um caminhante encontrar um trajeto que passe por cada uma das sete pontes da cidade apenas uma vez. Apesar de não fazer nenhuma menção ao termo **grafo**, este artigo motivou todo o desenvolvimento da teoria nos séculos seguintes, levando a avanços em diversas áreas do conhecimento como Química, Física, Computação etc.

A seguir, fazemos um breve resumo dos principais conceitos em teoria dos grafos que serão utilizados neste texto. Um grafo é uma estrutura discreta $G = (V, E)$, onde V representa um conjunto não vazio de vértices e E representa um conjunto de pares de elementos de V denotados por $(u, w) \in VxV$, chamados de arestas. Denotamos por $n = |V|$ o número de vértices presentes no grafo e por $m = |E|$ o número de arestas.

Alguns termos são definidos a seguir.

1. **Grafo não direcionado**: Um grafo é dito não direcionado quando não há um sentido de fluxo associado às suas arestas.
2. **Grafo direcionado ou dígrafo**: Um grafo é direcionado quando há um sentido de fluxo associado às suas arestas.
3. **Grafo ponderado**: Um grafo é ponderado quando existem pesos associados às suas arestas. Grafos **não ponderados** são equivalentes a grafos com o peso de todas as arestas igual a 1.
4. **Grafo conexo**: Se existe um caminho ente quaisquer dois vértices de um grafo, ele é chamado de **conexo**.
5. **Subgrafo**: Um subgrafo $H(Y, W)$ de $G(V, E)$ é qualquer grafo em que $Y \subseteq V$ e $W \subseteq E$.
6. **Árvore geradora**: É um subgrafo de G que contém os mesmos vértices, porém não possui ciclos (triângulos). Um grafo pode ter mais do que uma árvore geradora e todo grafo conexo tem pelo menos uma árvore geradora.
7. **Isomorfismo**: Dois grafos $G(V, E)$ e $H(Y, W)$ são isomorfos se há uma bijeção f de V em Y tal que dois vértices u e v são adjacentes em G se e somente se $f(v)$ e $f(w)$ são adjacentes em H.
8. **Vizinhança ou adjacência**: Dois vértices u e v são vizinhos ou adjacentes se existe uma aresta (u, v).
9. **Grau de um vértice**: O grau de um vértice é igual ao número de arestas incidentes sobre ele. Para grafos direcionados, o grau de um vértice u é a soma do **grau de entrada** com o **grau de saída**.

   O **grau de entrada** de um vértice u é igual ao número de arestas que possuem como destino o vértice u. O **grau de saída** é igual ao número de arestas que possuem como origem o vértice u.

Outras definições podem ser encontradas nas referências (NETTO, 2003; CORMEN et al., 2012; ROSEN, 2002; FEOFILOFF; KOHAYAKAWA; WAKABAYASHI, 2011).

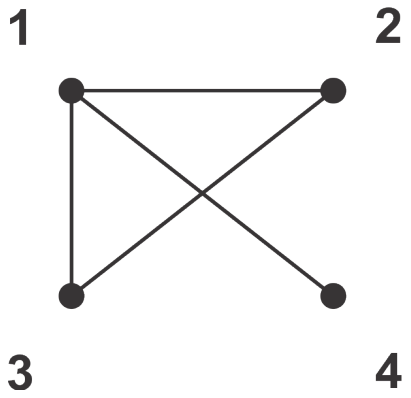
1.2 Representação de grafos

1.2.1 Lista de adjacências

O modo mais simples e compacto de representar um grafo é com listas de adjacência. Este é o método mais utilizado para representar grafos esparsos - onde m é muito menor que

n^2. Esta forma de representação consiste em um conjunto de n listas, uma para cada vértices. Cada lista representa um vértice e seus elementos representam os vértices que compartilham arestas com ele.

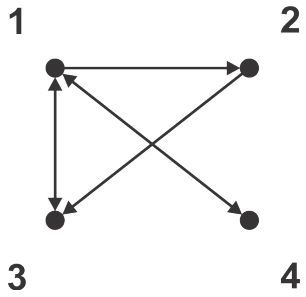
Para grafos não direcionados, a representação pode ser feita de acordo com a Figura 1. Note que cada ligação é indicada por apenas um sentido, evitando-se repetições. Neste caso é comum representar na lista de cada vértice apenas os vértices com índice superior ao seu. Observe também que uma lista associada a um vértice pode ser vazia, como as listas que representam os vértices 3 e 4 (NETTO, 2003). É importante salientar que em algumas referências os dois sentidos de ligação são representados na lista, mesmo em grafos não direcionados. Nesta segunda forma, a quantidade de memória necessária para armazenar a lista é maior, porém pode ser vantajosa em casos onde se deseja determinar se existe uma aresta (u, v). Na primeira forma de representação, devemos procurar por v na lista de adjacências de u e, caso a referência ao vértice não exista, devemos procurar por u na lista de v. Na segunda forma, basta procurar em uma das listas.



Vértice	Vértices Adjacentes
1	2 3 4
2	3
3	
4	

Figura 1 – Listas de adjacência para grafo não direcionado.

Em grafos direcionados, existem duas formas equivalentes de construção da lista de adjacências. A representação pode ser feita a partir dos vértices de origem de cada aresta ou a partir dos vértices de destino. Veja a Figura 2.



Vértices	
Origem	Destino
1	2 3 4
2	3
3	1
4	1

Vértices	
Destino	Origem
1	3 4
2	1
3	1 2
4	1

Figura 2 – Listas de adjacência para grafo direcionado.

De forma geral, listas de adjacência são eficientes em consumo de memória, porém existem algumas desvantagens, como o fato de que não há nenhuma forma mais rápida de

se determinar a existência de uma aresta (u, v) do que procurar por v na lista de u. Para este tipo de problema as matrizes de adjacência, descritas na próxima seção, fornecem métodos mais rápidos. A quantidade de memória exigida pelas listas de adjacência é $\Theta(V + E)$ (CORMEN et al., 2012).

1.2.2 Matriz de Adjacências

Nesta forma de representação, cada linha e cada coluna de uma matriz $n \times n$ são associadas a um vértice do grafo. Considerando que os vértices são numerados de forma arbitrária, com valores $1, 2, 3, .., n$, a matriz de adjacências associada a este grafo é $A = (a_{ij})$ tal que

$$a_{ij} = \begin{cases} 1, & \text{se } v_i \text{ e } v_j \text{ são adjacentes;} \\ 0, & \text{caso contrário.} \end{cases}$$

para arestas sem pesos. Caso as arestas sejam ponderadas, a posição (a_{ij}) recebe o peso da aresta. Na Figura 3 é possível ver a construção da matriz de adjacência de um grafo não direcionado. Observe que neste caso a matriz é simétrica.

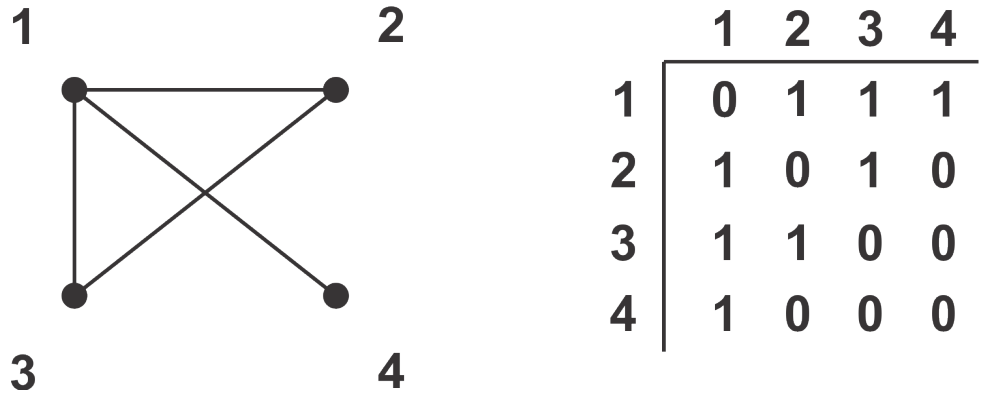
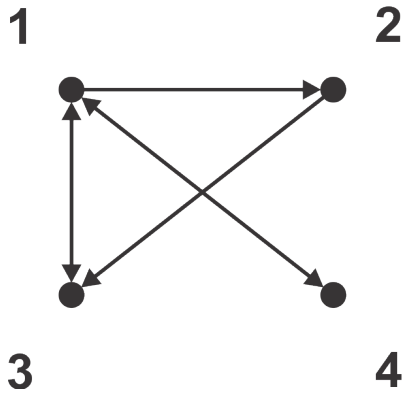


Figura 3 – Matriz de adjacência para grafo não direcionado.

Na Figura 4 é possível ver a construção da matriz de adjacência para o grafo direcionado.

A matriz de adjacência não é o método mais econômico para a representação de grafos, a quantidade de memória exigida é $\Theta(V^2)$ independente do número de arestas no grafo (CORMEN et al., 2012). Porém, esta forma de representação é preferível para a representação de grafos pequenos ou de grafos que não são esparsos.

Existem outras formas de representação de grafos que são menos utilizadas, como as **matrizes de incidência** e as **matrizes figurativas**. Para uma revisão, consulte o livro de Boaventura Netto (NETTO, 2003).



	1	2	3	4
1	0	1	1	1
2	0	0	1	0
3	1	0	0	0
4	1	0	0	0

Figura 4 – Matriz de adjacência para grafo direcionado.

1.3 O espectro de um grafo

Seja A a matriz de adjacência de um grafo G com n nodos, um escalar λ é um ***autovalor*** de A se satisfaz a expressão $Ax = \lambda x$, para $x \neq 0$, $x \in \mathbb{R}^n$. O conjunto dos autovalores de A é definido como o conjunto das $|V|$ raízes associadas ao polinômio

$$p_A = det(A - \lambda I_n) \tag{1.1}$$

chamado de **polinômio característico** de A, onde I_n é a matriz identidade de ordem n. O ***espectro*** do grafo G, $spect(G)$, é uma matriz $2 \times n$ em que os n autovalores de G são dispostos na primeira linha ordenados em ordem decrescente e na segunda linha suas respectivas multiplicidades algébricas. O maior autovalor da matriz de adjacência é chamado de **raio espectral** ou **índice** do grafo (SANTOS; RANGEL; BOERES, 2010). Ele está associado a uma variedade de processsos dinâmicos e é utilizado para calcular um índice de centralidade chamado *importância dinâmica*, definida na seção 2.6.3 sobre *centralidade de autovalor*. Em grafos não-direcionados, onde a matriz de adjacência é simétrica, todos os autovalores são reais.

Muitas características topológicas de um grafo podem ser determinadas a partir de seu espectro conforme apresentado em Santos (SANTOS; RANGEL; BOERES, 2010). A seguir, citamos algumas destas características.

1. O **número de arestas no grafo** é a soma dos quadrados dos autovalores dividido por dois.

2. O **número de triângulos no grafo** é soma dos cubos dos autovalores dividido por seis.

É importante mencionar que se dois grafos são isomorfos, eles têm o mesmo espectro, porém a recíproca não é verdadeira. Grafos com o mesmo espectro são chamados de *co-espectrais*.

Nas seções a seguir discutiremos outro conceito importante no estudo da teoria espectral de grafo, o *espectro do laplaciano*.

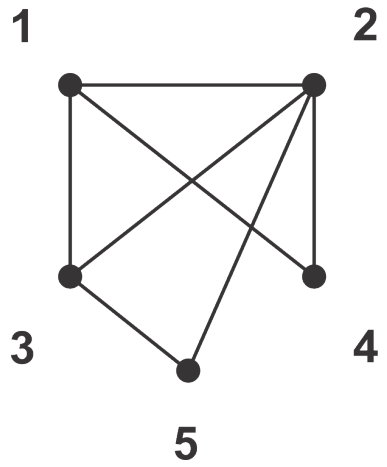
1.3.1 O espectro laplaciano

Dentre as diversas matrizes associadas a um grafo, as de maior interesse são a **matriz de graus** e a **matriz laplaciana**.

Definimos a **matriz de graus** de um grafo como a matriz diagonal $D = (d_{ij})$, tal que

$$d_{ij} = \begin{cases} g(v_i), & \text{para } i = j; \\ 0, & \text{caso contrário.} \end{cases}$$

onde $g(v_i)$ é o grau do i-ésimo vértice do grafo G. Na figura 5 é possível ver a construção da matriz de graus para um grafo de 5 vértices.



	1	2	3	4	5
1	3	0	0	0	0
2	0	4	0	0	0
3	0	0	3	0	0
4	0	0	0	2	0
5	0	0	0	0	2

Figura 5 – Matriz de graus.

A **matriz laplaciana** de um grafo é definida como $L = D - A$. Para o grafo da Figura 5, temos que sua matriz de adjacência é

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

logo, a matriz laplaciana L associada a este grafo é

$$L = \begin{bmatrix} 3 & -1 & -1 & -1 & 0 \\ -1 & 4 & -1 & -1 & -1 \\ -1 & -1 & 3 & 0 & -1 \\ -1 & -1 & 0 & 2 & 0 \\ 0 & -1 & -1 & 0 & 2 \end{bmatrix}$$

O espectro desta matriz é chamado de **espectro laplaciano**. Para a matriz laplaciana acima, temos o espectro $\zeta(G)$:

$$\zeta(G) = \begin{bmatrix} 5 & 4.4142 & 3 & 1.5858 & 7.3915e-16 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

A partir do espectro laplaciano, diversas propriedades estruturais de um grafo podem ser determinadas como, por exemplo, o número de componentes conexas é igual a quantidade de autovalores nulos (FRITSCHER, 2011).

O **número de arestas** também pode ser determinado pelo espectro laplaciano:

$$\sum_{i=1}^{n} \mu_i = 2m \tag{1.2}$$

onde $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$ são os autovalores do laplaciano e m é o número de arestas.

Para ilustrar a aplicação desta fórmula, considere o grafo da Figura 6.

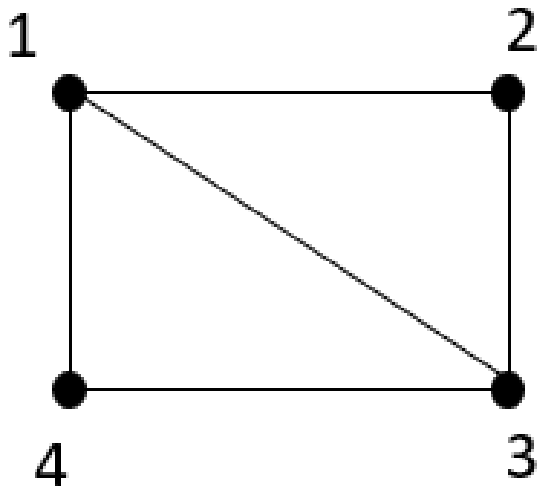


Figura 6 – Grafo de exemplo para o cálculo do número de arestas.

Para o cálculo do espectro laplaciano, consideremos a representação deste grafo por meio de sua matriz de adjacências A e sua matriz de graus D.

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

$$D = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

Aplicando a definição de matriz laplaciana, $L = D - A$, obtemos a matriz L, definida abaixo.

$$L = \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & -1 \\ -1 & 0 & -1 & 2 \end{bmatrix}$$

Ao calcular os autovalores desta matriz obtemos o seguinte espectro.

$$\zeta(G) = \begin{bmatrix} 0 & 2 & 4 & 4 \\ 1 & 1 & 2 & 2 \end{bmatrix}$$

De acordo com a equação 1.2, a soma dos autovalores da matriz laplaciana é igual ao dobro do número de arestas. Observamos que a soma dos autovalores calculados acima é igual a 10, que é exatamento o dobro do número de arestas do grafo da Figura 6.

Além disso, o segundo menor autovalor da matriz laplaciana, chamado de **conectividade algébrica** está relacionado a diversas invariantes do grafo, como o diâmetro, distância média, conectividade e é tido por Mohar e colaboradores (MOHAR et al., 1991) como a informação mais importante contida no espectro de um grafo.

Para um estudo mais aprofundado da teoria espectral de grafos, incluindo demonstrações das propriedades, consulte as referências (FRITSCHER, 2011; SANTOS; RANGEL; BOERES, 2010; MOHAR et al., 1991).

1.4 Redes complexas

Nos últimos anos a análise de diversos sistemas naturais e sociais vem sofrendo forte influência do conceito de redes complexas. Dentre eles é possível citar as redes neurais, redes de telefonia, a *World Wide Web* e até mesmo a linguagem humana. A modelagem destes sistemas como redes revelou propriedades importantes, como a emergência de leis de potência na distribuição de graus.

É possível considerar alguns estudos relacionados à análise de redes sociais como o início da aplicação de redes complexas, mesmo antes do termo ser popularizado e de técnicas serem formalizadas. Um desses estudos é o conhecido artigo do psicólogo Stanley Milgram publicado em 1967 (MILGRAM, 1967). No estudo Milgram enviou cartas a um determinado número de pessoas, que não eram os destinatários corretos e solicitou que cada pessoa que recebeu a correspondência enviasse a mesma para o destinatário, caso o conhecesse, ou enviasse para outra pessoa que poderia conhecer o destinatário. O objetivo

desse estudo era determinar o número médio (a rigor, a mediana) de pessoas pelas quais as cartas passariam antes de chegar ao seu destinatário. A partir deste experimento, o conceito de "*seis graus de separação*" foi popularizado e passou a receber a atenção da comunidade científica nas décadas seguintes.

Não existe uma definição clara que diferencie uma rede complexa de um grafo trivial. Um dos principais aspectos observados diz respeito a conectividade que para redes complexas não segue um padrão regular. Mesmo o conceito de padrão regular não é definido com clareza nas referências existentes. De uma forma geral, é necessário observar, além do padrão de conectividade, a forma como a rede evolui ao longo tempo, dentre muitos outros aspectos (METZ et al., 2007).

O estudo das características estruturais e dinâmicas das redes complexas é feito utilizando métricas, oriundas de análises estatísticas. Neste texto, utilizamos estas métricas, em conjunto com outros valores, para o estudo de grafos formados por textos escritos em língua natural. Este tipo de grafos possui propriedades importantes que são compartilhadas com outras redes do mundo real, como a distrubição de graus em lei de potência e o alto valor de coeficiente de agrupamento. Estas propriedades serão discutidas nas próximas seções deste capítulo.

1.5 Modelos de redes

1.5.1 Modelo de Erdös-Rényi

O estudo das redes aleatórias foi sistematizado por Paul Erdös e Alfred Rényi em artigo publicado no ano de 1959 (ERDÖS; RÉNYI, 1959). O modelo de Erdös-Rényi propõe uma técnica para a criação de grafos aleatórios com n vértices e m arestas. Iniciando com n vértices desconectados, pares de vértices são escolhidos aleatoriamente, de uma forma que evite múltiplas conexões e *self-loops*, então estabelece-se uma aresta entre os vértices escolhidos. O procedimento é realizado até que o número de arestas seja igual a m. Um grafo gerado por esta técnica é apenas uma de muitas possíveis combinações de arestas.

Uma técnica alternativa consiste em se conectar os pares de vértices com probabilidade $0 < p < 1$. Esta técnica produz um conjunto de grafos diferente do conjunto gerado pela técnica anterior e grafos com m arestas aparecerão no conjunto com probabilidade $p^m(1-p)^{n(n-1)/2-m}$ (BOCCALETTI et al., 2006).

Apesar de muito estudado, o modelo de Erdös-Rényi não reproduz a maioria das propriedades das redes do mundo real. Uma das principais características das redes aleatórias, pouco observadas em redes sociais ou da natureza é a distribuição de graus que se aproxima de uma distribuição de Poisson. A maioria das redes que estudamos, obedece

uma distribuição livre de escala, discutida nas seções seguintes.

Para este tipo de rede a probabilidade de conexão é a mesma para quaisquer dois pares de vértices e todos os vértices tem aproximadamente a mesma quantidade de arestas conectadas a ele. Muitos experimentos realizados mostraram que em redes reeais, estas características dificilmente são encontradas.

1.5.2 Modelo de Watts-Strogatz

Uma propriedade comum entre as redes aleatórias e as redes do mundo real é a propriedade de **mundo pequeno**, porém os dois tipos de rede diferem em relação ao **coeficiente de aglomeração** (*clustering coefficient*), discutido em detalhes na seção sobre métricas de rede. De forma geral, podemos entender o coeficiente de aglomeração como a probabilidade dos vizinhos de um determinado vértice estarem conectados entre si, formando grupos locais. Em redes aleatórias, este coeficiente tem um valor baixo, devido a probabilidade de conexão entre os vértices ser uniforme. Nas redes do mundo real, geralmente ocorre o oposto, ou seja, costumam apresentar altos valores de coeficiente de aglomeração.

Em 1998, Watts e Strogatz (WATTS; STROGATZ, 1998) propuseram um modelo para a construção de redes com a propriedade de mundo pequeno e alto coeficiente de agrupamento simultaneamente. A proposta parte da construção de um grafo regular, ilustrado na forma de uma anel de vértices, com cada vértice conectado aos seus k vizinhos mais próximos. Sendo $k/2$ conectados no sentido anti-horário e $k/2$ conectados no sentido horário, conforme apresentado na imagem da Fig. 7. A partir deste grafo, as arestas são modificadas de forma a inserir aleatoriedade na rede. Isto é feito reposicionando as arestas do grafo aleatoriamente com probabilidade p. Para $p = 0$, nenhuma aresta é reposicionada, mantendo-se o grafo regular originalmente construído. Esta rede possui alto coeficiente de agrupamento, porém não apresenta a propriedade de mundo pequeno. Para $p = 1$, todas as arestas são reposicionadas, transformando o grafo em uma rede aleatória. Neste tipo de rede, surge a propriedade de mundo pequeno, porém o coeficiente de aglomeração diminui conforme p aumenta. O modelo de Watts e Strogatz possibilita a construção de um grafo que combina estas duas características de forma a se aproximar mais do que o modelo de Erdös-Rényi das redes do mundo real, particularmente das redes sociais que foram a motivação inicial do trabalho de Watts em redes complexas, conforme discutido na Ref. (BARÁBASI, 2009).

Apesar dos avanços apresentados por este modelo, alguns aspectos encontrados em diversas redes do mundo real como a *World Wide Web* ou mesmo as redes sociais são difíceis de explicar a partir do modelo de Watts-Strogatz. Uma destas propriedades é a existência de *hubs* que são vértices com um grande número de vizinhos, observada em diversos experimentos com redes reais. Este problema, foi resolvido com o desenvolvimento

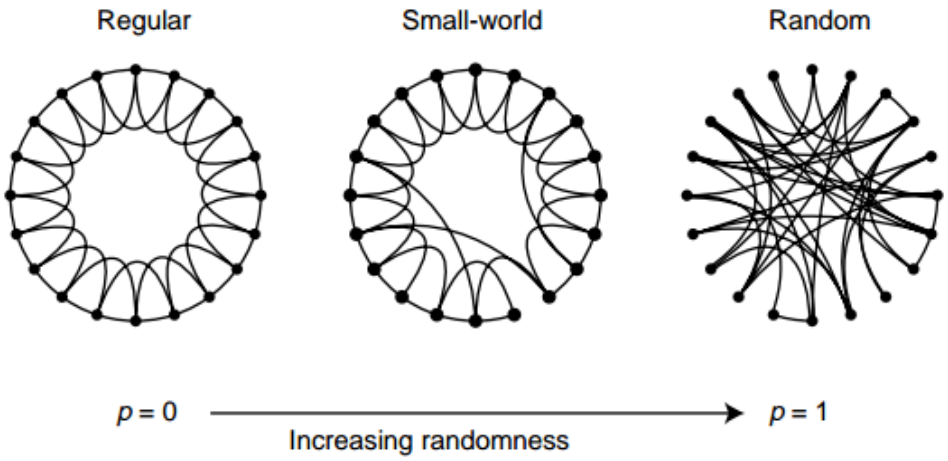


Figura 7 – Modelo de Watts-Strogatz. Imagem reproduzida de (WATTS; STROGATZ, 1998)

do modelo de Barabási-Albert, discutido a seguir.

1.5.3 Modelo de Barabási-Albert

Resultados empíricos mostraram que diversas redes do mundo real, como a internet, redes de citação, redes de interação de proteínas e as redes de palavras apresentam uma distribuição de graus que se aproxima de uma lei de potência. Estas redes foram chamadas de **redes livres de escala** (ALBERT; BARABÁSI, 2002). Isto significa que as conexões neste tipo de rede obedecem algum tipo de ordem e não ocorrem de forma completamente aleatória.

Para explicar o surgimento da lei de potência observada nestas redes, Barabási e Albert propuseram em 1999 (BARABÁSI; ALBERT, 1999) que dois aspectos precisam ser observados. O primeiro aspecto é o *crescimento*, que diz respeito a adição contínua de novos vértices durante a existência da rede. Esta característica é compartilhada pelas redes citadas no parágrafo anterior. O segundo aspecto, é o que foi chamado por Barabási e Albert de *conexões preferenciais*, que significa que a probabilidade de haver uma conexão para um determinado nó, depende do grau deste nó. Este comportamento pode ser descrito informalmente pela expressão *os ricos ficam mais ricos*.

O modelo de Barabási-Albert é descrito algoritmicamente, na referência ALBERT; BARABÁSI, da seguinte forma:

1. A rede é iniciada com um pequeno número de vértices (m_0);
2. A cada intervalo de tempo um novo vértice é adicionado com $m \ll m_0$ arestas que conectam o novo vértice a m diferentes vértices já existentes;

3. As conexões são feitas de acordo com a probabilidade de um novo nó se conectar a um nó i, que é dada pelo seu grau k_i, da seguinte forma

$$\Pi(k_i) = k_i / \sum_i k_j \ . \tag{1.3}$$

1.6 Métricas de redes

A seguir discutimos alguma métricas utilizadas para a caracterização e análise de redes complexas. É a partir dessas medidas que os atributos utilizados pelos algoritmos de aprendizado de máquina são derivados. Para uma revisão completa de diversas medidas de rede, consultar as referências (BRANDES; ERLEBACH, 2005; COSTA et al., 2007).

1.6.1 Medidas de distância

A seguir, são listadas as principais medidas baseadas em distância.

1. **Distância geodésica média** Esta medida é definida como a média dos menores caminhos entre todos os pares de vértices do grafo, calculada da seguinte forma.

$$l = \frac{1}{n(n-1)} \sum_{i \neq j} d_{ij} \ . \tag{1.4}$$

2. **Eficiência global**: Um problema com a distância geodésica média é que quando há vértices não conectados na rede, a soma diverge. Se a soma for feita desconsiderando os vértices desconectados, o resultado final é distorcido. Para evitar estes efeitos, foi proposta a medida de *eficiência global*.

$$E_f = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \tag{1.5}$$

   Esta medida quantifica a eficiência da rede em enviar informações de um vértice para outro.

1.6.2 Agrupamento

1. **Coeficiente de agrupamento (*Clustering coefficient*)**: Esta medida é muito importante na caracterização de redes complexas, pois quantifica a probabilidade dos vizinhos de um determinado vértice estarem conectados entre si e caracteriza a presença de triângulos na rede. Uma das formas de determinar o *coeficiente de*

agrupamento de uma rede é através da expressão a seguir, também chamada de *transitividade*.

$$C = \frac{3N_t}{N_3}, \tag{1.6}$$

onde N_t é o número de triângulos na rede e N_3 é o número de trios de vérrtices conectados. O número de triângulos e o número de trios na rede podem ser determinados a partir da matriz de adjacência da seguinte forma.

$$N_t = \sum_{k>j>i} a_{ij}a_{ik}a_{jk} \tag{1.7}$$

com

$$N_3 = \sum_{k>j>i} \left(a_{ij}a_{ik} + a_{ji}a_{jk} + a_{ki}a_{kj}\right) . \tag{1.8}$$

Também é possível calcular o coeficiente de agrupamento para um vértice específico da seguinte forma.

$$C_i = \frac{N_t(i)}{N_3(i)} \tag{1.9}$$

onde $N_t(i)$ é definido como o número de triângulos contendo i e $N_3(i)$ é o número de trios tenho i como vértice central.

$$N_t(i) = \sum_{k>j} a_{ij}a_{ik}a_{jk}, \tag{1.10}$$

com

$$N_3(i) = \sum_{k>j} a_{ij}aik \tag{1.11}$$

1.6.3 Medidas de centralidade

Os índices de centralidade são utilizados para quantificar a importância de vértices ou arestas específicas na rede. Existe uma infinidade de índices de centralidade, porém neste texto trataremos apenas de alguns mais conhecidos.

Antes de tratar dos índices especificamente é importante notar que a medida de *grau* definida anteriormente pode ser considerada uma medida de centralidade. Em certas aplicações, a quantidade de arestas que um vértice recebe pode ser considerada uma medida de importância daquele vértice, como por exemplo no caso das redes sociais.

Abaixo definimos os índices de interesse para esta monografia.

1. ***Betweenness***: A centralidade de *betweeness* pode ser definida tanto para vértices quanto para arestas. Esta medida depende da quantidade de menores caminhos que passam pelo vértice ou pela aresta. É uma medida relacionada ao fluxo na rede e um alto valor de betweenness indica que aquele elemento tem grande influência na

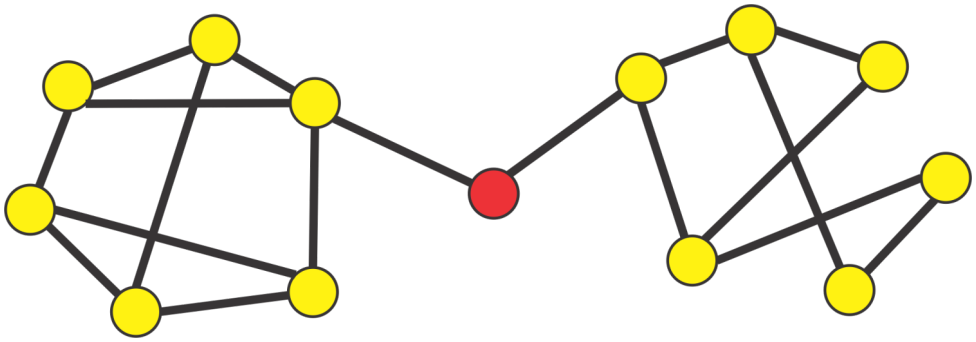


Figura 8 – *Betweenness*.

troca de informações na rede. É possível observar na Figura 8 abaixo que o nó em destaque é o nó com maior *betweenness*.

$$B_i = \sum_s \sum_t \frac{\eta_{st}^i}{g_{st}} \,, \tag{1.12}$$

onde η_{st}^i é o número de menores caminhos distintos entre os nós v_s e v_t que passam por v_i e g_{st} é o número total de menores caminhos entre v_s e v_t.

2. **Centralidade de autovalor**: O maior autovalor da matriz de adjacência de um grafo está associado a uma medida de centralidade chamada de ***importância dinâmica***. Ela é utilizada em situações onde se deseja, por exemplo, identificar vértices ou arestas chave para a manutenção da integridade da rede. Uma possível situação onde esta medida é útil é quando se deseja identificar vértices que ao serem removidos fazem a rede colapsar. A importância dinâmica de uma aresta (i,j), $I_{i,j}$, é o valor $-\Delta\lambda_{ij}$, pelo qual λ decresce ao ser removida a aresta (i,j), normalizado por λ:

$$I_{ij} \equiv \frac{-\Delta\lambda_{ij}}{\lambda} \,, \tag{1.13}$$

onde λ é o maior autovalor da matriz de adjacência. Para um vértice k, a importância dinâmica é definida analogamente, pela variação do maior autovalor ao se remover o vértice k da rede, normalizada por λ:

$$I_k \equiv \frac{-\Delta\lambda_k}{\lambda} \tag{1.14}$$

Para uma visão mais detalhada sobre o índice de importância dinâmica consultar a referência (RESTREPO; OTT; HUNT, 2006).

2

Processamento de Linguagem Natural

Processamento de Linguagem Natural (PLN), é a área de estudo que trata dos problemas envolvendo o processamento computacional de textos escritos em linguagem humana. Os avanços recentes da área têm possibilitado a criação de sistemas cada vez mais sofisticados em áreas como:

- Tradução automática.
- Recuperação de informação.
- Geração automática de textos.
- Sumarização e simplificação de textos.
- Categorização textual.
- Reconhecimento de autoria.

e muitas outras. Neste texto temos como foco a categorização textual, mais especificamente o problema de reconhecimento de autoria em discursos escritos em língua portuguesa. Para uma introdução geral ao Processamento de Linguagem Natural, as duas referências básicas são os livros (MANNING; SCHÜTZE, 1999) e (JURAFSKY; MARTIN, 2018).

2.1 Categorização textual

Além do reconhecimento de autoria, muitas outras tarefas em PLN exigem algum tipo de classificação. Dentre elas é possível citar a **detecção de *spam*** em sistemas de correio eletrônico. Outra aplicação importante é a chamada **análise de sentimentos**,

que trata da detecção da polaridade ou orientação de um texto. Por exemplo, sistemas que analisam os comentários a respeito de um determinado produto em um *website* e decidem automaticamente se determinado comentário tem orientação positiva ou negativa. Um exemplo de um classificador probabilistico simples para análise de sentimentos é apresentado a seguir.

2.1.1 Um classificador probabilístico

Um dos métodos mais simples de classificação utilizados em PLN é o classificador ***Naive Bayes***. Este método ignora a ordem com que as palavras aparecem no texto e qualquer relação entre elas, utilizando a representação conhecida como ***bag-of-words***. Nesta representação, o texto é tratado como um conjunto não ordenado de palavras e mantém-se apenas a frequência com que cada palavra aparece no texto. O objetivo do classificador é inferir a classe $\hat{c}$ com a maior probabilidade a *posteriori* dado um documento.

$$\hat{c} = \arg\max_{c \in C} P(c \mid d) \tag{2.1}$$

A ideia deste método é utilizar a regra de Bayes para transformar a Equação 2.1 em outras probabilidades úteis para o processo de classificação. Todo o desenvolvimento matemático do método pode ser encontrado na referência (JURAFSKY; MARTIN, 2018). Aqui apresentaremos apenas a principal equação que decorre destes desenvolvimentos, que nos dá a probabilidade de uma palavra w pertencer a uma das classes $c \in C$, que em nosso exemplo são a classe positiva (+) e a classe negativa (-).

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V}(count(w, c) + 1)} \tag{2.2}$$

Onde V é o vocabulário, que é o conjunto de todas as palavras únicas em todas as classes.

Após calcular a probabilidade de uma palavra conter polaridade positiva ou negativa, basta calcular a verossimilhança para determinar a classe do texto.

$$c_{NB} = \arg\max_{c \in C} P(c) \prod_{i \in positions} P(w_i \mid c) \tag{2.3}$$

Para ilustrar o método, vamos reproduzir o exemplo apresentado na referência. Neste exemplo, é apresentado um conjunto de treinamento, composto por 3 frases com polaridade negativa (-) e duas frases com polaridade positiva (+). Utilizaremos o método para classificar uma nova frase.

A seguir, as frases já rotuladas:

1. (-) ***"just plain boring".***
2. (-) ***"entirely predictable and lacks energy".***
3. (-) ***"no surprises and very few laughs".***
4. (+) ***"very powerful".***
5. (+) ***"the most fun film of the summer".***

Devemos determinar a polaridade da frase a seguir:

1. (?) ***"predictable with no fun".***

O procedimento passo a passo para a classficação é apresentado abaixo.

1. Primeiro é necessário calcular a probabilidade *a priori* de cada uma das classes.

$$P(-) = \frac{3}{5}\ P(+) = \frac{2}{5}$$

2. Depois é necessário calcular a probabilidade de cada uma das palavras pertencer a cada umas das classes utilizando a equação 2.2. Como a palavra *"with"* não aparece no vocabulário, não é necessário fazer os cálculos referentes a ela.

$$P(\text{"}predictable\text{"} \mid -) = \frac{1+1}{14+20}\ P(\text{"}predictable\text{"} \mid +) = \frac{0+1}{9+20}$$
$$P(\text{"}no\text{"} \mid -) = \frac{1+1}{14+20}\ P(\text{"}no\text{"} \mid +) = \frac{0+1}{9+20}$$
$$P(\text{"}fun\text{"} \mid -) = \frac{0+1}{14+20}\ P(\text{"}no\text{"} \mid +) = \frac{0+1}{9+20}$$

3. Então calcula-se a classe a que a sentença S pertence utilizando a Equação 2.3:

$$P(-)P(S \mid -) = \frac{3}{5} \times \frac{2\times2\times1}{34^3} = 6.1 \times 10^{-5}$$

$$P(+)P(S \mid +) = \frac{2}{5} \times \frac{1\times1\times2}{29^3} = 3.2 \times 10^{-5}$$

4. Com base nos resultados do item anterior, concluimos que o modelo classificou a frase com a polaridade *negativa* (-).

Para maiores detalhes a respeito do classificador ***Naive Bayes*** e de outros métodos, como ***Regressão Logística***, consultar a referência (JURAFSKY; MARTIN, 2018).

2.2 Formas de representação textual

Nesta seção, apresentamos algumas formas tradicionais para a representação estatística de textos em tarefas de PLN, como recuperação de informação e classificação.

2.2.1 Modelo de espaço vetorial

Ao considerarmos modelos de representação de textos em língua natural, tipicamente os dados utilizados são representados na forma $(\vec{x}, c)$, onde $\vec{x} \in \mathbb{R}$ é o vetor de atributos para cada exemplo e c é o rótulo da classe.

Uma das formas de representação mais clássicas, muito utilizada na área de recuperação de informação é o *modelo de espaço vetorial* (MANNING; SCHÜTZE, 1999).

Nesse modelo, os textos são representados como vetores em um espaço de alta dimensionalidade, onde cada dimensão corresponde a uma palavra no conjunto de textos. Os valores das coordenadas, usualmente são definidos como a frequência com que a respectiva palavra aparece no documento. A proximidade espacial entre os vetores, determina a similaridade entre os textos, sendo que a medidade de similaridade tipicamente utilizada, principalmente pelos sistemas de recuperação de informação, é o cosseno entre este vetores.

Para uma dicussão mais detalhada a respeito de técnicas tradicionais de classificação, consultar a referência (MANNING; SCHÜTZE, 1999).

2.2.2 TF-IDF

A simples frequência com que uma palavra ocorre em um texto não é uma boa medida de sua importância, pois é possível ter palavras que aparecem muitas vezes em um texto, mas que não possuem valor semântico algum, como é o caso dos artigos, preposições etc. Muitas vezes, palavras que aparecem menos no texto possuem valor semântico maior, o que dificulta o uso da representação descrita anteriormente. Um método que se propõe a resolver isto é o chamado ***tf-idf*** (***term frequency - inverse document frequency***), que é definido como o produto de dois termos, sendo o primeiro a frequência com que a palavra aparece e o segundo um termo que atribui um peso maior a palavras que aparecem em menos documentos.

A seguir, apresentamos a descrição de ambos os termos:

1. ***tf***: Este termo é definido a partir da frequência com que uma palavra aparece no documento, sendo atribuido um valor maior para palavras que ocorrem mais vezes.

O valor de ***tf*** pode ser calculado pela fórmula a seguir:

$$tf_{t,d} = \begin{cases} 1 + \log_{10} count(t, d), & \text{if count(t, d)} > 0 \\ 0, & \text{caso contrário.} \end{cases}$$

2. ***idf***: Este termo atribui um peso maior para palavras que aparecem em um número menor de documentos, pois estes termos são importantes para discriminá-los. A definição de ***idf*** é dada a seguir:

$$idf_t = \log_{10} \frac{N}{df_t}. \tag{2.4}$$

O valor de ***tf-idf*** para uma palavra t em um documento d, $w_{t,d}$ é dado por:

$$w_{t,d} = tf_{t,d} \times idf_t. \tag{2.5}$$

O método ***tf-idf*** é amplamente empregado em PLN, para os mais variados tipos de tarefas, como por exemplo o cálculo da similaridade entre palavras e da similaridade entre documentos. Uma visão mais detalhada é apresentada no livro de Jurafsky e Martin (JURAFSKY; MARTIN, 2018).

2.2.3 Representação por redes complexas

O grafo construído a partir de textos onde os vértices representam palavras e as arestas representam relações de vizinhança entre palavras é chamado de grafo de co-ocorrência. Para o trecho a seguir do poema ***Os Sapos*** de **Manuel Bandeira**, foi construído o grafo da Figura 9.

Enfunando os papos,
Saem da penumbra,
Aos pulos, os sapos.
A luz os deslumbra.

Em ronco que aterra,
Berra o sapo-boi:
- "Meu pai foi à guerra!"
- "Não foi!" - "Foi!" - "Não foi!".

No artigo intitulado *"The Small World of Human Language"* (CANCHO; SOLÉ, 2001), Cancho e Solé demonstram, com experimentos, que as redes formadas a partir de textos possuem propriedades típicas de redes complexas. A primeira propriedade interessante é que essa rede pode ser considerada um *mundo pequeno*, com uma distância

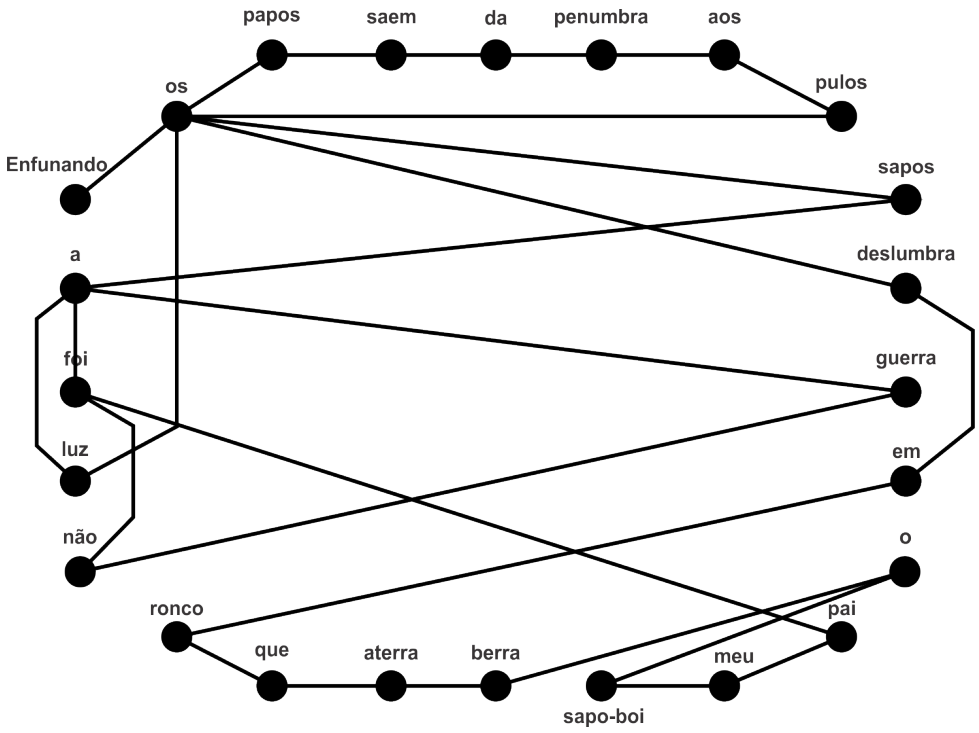


Figura 9 – Rede produzida a partir de trecho da poesia "Os Sapos" de Manuel Bandeira

.

média entre palavras de aproximadamente $d = 2 \sim 3$, para o *British National Corpus - BNC*[1]. Outra propriedade importante é que a rede segue uma distribuição de graus *livre de escala*. Este último pode ser visto como uma consequência da lei de Zipf (ZIPF, 1972), que diz que a distribuição de palavras em um texto segue uma lei de potência.

Estes resultados demonstram que as redes de palavras possuem as mesmas propriedades estatísticas que outras redes do mundo real, logo a utilização de métricas de redes complexas como forma de caracterizar textos é justificável.

Existem outras formas de construção de redes a partir de textos (AMANCIO, 2013), como por exemplo, as **redes de dependência sintática** que associam uma aresta a pares de palavras que possuem alguma dependência sintática entre si. Estas redes se assemelham muito as redes de co-ocorrência, pois a maioria das ligações acontecem entre palavras vizinhas.

Outro forma de modelagem são as **redes semânticas**. Neste tipo de rede, são estabelecidas arestas entre palavras de acordo com regras semânticas. Exemplo de regras para associação são "é um", "oposição" e outras. Neste tipo de rede, os *hubs* usualmente são palavras polissêmicas que possuem vários sentidos associados e assim possuem probabilidade maior de se conectarem a outras palavras.

As abordagens tradicionais de classificação de textos são fortemente baseadas no conteúdo das palavras, ao contrário da abordagem baseada em redes complexas que leva

[1] http://www.natcorp.ox.ac.uk/

em consideração apenas a forma como as palavras estão organizadas.

A técnica desenvolvida por Amancio (AMANCIO, 2013) propõe que os textos sejam representados por vetores, cujas coordenadas são dadas por métricas de redes complexas extraídas a partir do texto modelado como rede co-ocorrência. A princípio, as medidas utilizadas são relativas às palavras (vértices) do texto (grafo), logo é definido um método para que a partir de medidas que caracterizam palavras sejam gerados atributos que representam textos inteiros. Para isto, alguns procedimentos estatísticos são adotados, como o cálculo da média para aquela métrica, o cálculo de médias modificadas, o coeficiente de assimetria da distribuição (*skewness*), o expoente da distribuição da variável quando a mesma é uma lei de potência, dentre outras.

A aplicação desta forma de representação para classificação pode ser resumida nos passos a seguir:

1. **Pré-processamento**: Na primeira etapa do desenvolvimento, os autores realizam uma remoção dos elementos do texto que não serão úteis para a tarefa de classificação, como os sinais de pontuação e dependendo do experimento, as *stopwords*. Além disso o texto passa por processos de *tokenização* e *stemming*.

2. **Redes de palavras**: A segunda etapa é transformar os textos em grafos de co-ocorrência de palavras, conforme descrito anteriormente, onde cada palavra é transformada em um vértice do grafo e as arestas são estabelecidas entre palavras vizinhas no texto. Aqui, cada texto é modelado como uma rede diferente.

3. **Extração de medidas**: Nesta etapa são extraídas diversas métricas de redes para a construção dos vetores de atributos. Tipicamente, as medidas calculadas dizem respeito a vértices, então é necessário um tratamento estatístico para transformar medidas individuais dos vértices em medidas do grafo como um todo. Usualmente, o autor calcula médias, *Skewness*, dentre outras. A partir do resultado desta manipulação são construídos os vetores de atributos. É importante observar que a escolha das métricas a serem utilizadas se dá de forma empírica após diversos testes com métricas conhecidas procurando observar quais possuem mais influência no comportamento geral do experimento.

4. **Classificação**: Nesta etapa, o criador da técnica utiliza os vetores de atributos gerados a partir das métricas de redes e utiliza como entrada para diversos algorítmos de classificação, como *support vector machines*, *k-Nearest Neighbors*, e muitos outros. A partir do resultado, o autor avalia qual classificador gerou o melhor modelo para o dado problema sendo analisado.

Vários artigos foram publicados pelos autores desta técnica, com sua aplicação aos mais diversos problemas de classificação. Para uma visão mais detalhada veja as referências (AMANCIO et al., 2011), (AMANCIO, 2013), (AMANCIO, 2015).

3

Metodologia e Desenvolvimento

Neste capítulo apresentamos a metodologia utilizada no desenvolvimento da pesquisa, descrevendo os recursos fundamentais empregados e a forma como os experimentos foram conduzidos.

Os experimentos realizados neste trabalho buscaram validar a técnica desenvolvida por Amancio (AMANCIO, 2013) em sua tese de doutoramento, para o reconhecimento de autoria de trechos de discursos presidenciais. Além disso foram utilizadas diferentes formas de construção da rede de co-ocorrência buscando verificar se há diferenças na qualidade da classificação com diferente tipos de redes.

Os experimentos realizados tiveram como foco a classificação de autoria em discursos presidenciais, porém é possível adaptar as técnicas empregadas para outros tipos de classificação, como por exemplo, reconhecimento de gêneros literários e classificação de notícias. A implementação de tais experimentos ficarão como sugestão para trabalhos futuros.

3.1 Proposta

Tradicionalmente, os métodos de classificação textual mais utilizados adotam abordagens fortemente relacionadas a conteúdo, considerando principalmente características como a frequência em que as palavras ocorrem, como no exemplo de classificação do capítulo 2. Muitas vezes, o texto é modelado utilizando a estratégia de *bag of words*, não considerando os relacionamentos entre as palavras ou a posição em que uma palavra ocorre no texto. Métodos que utilizam a estratégia de *bag of words* tem dificuldades em capturar a semântica em textos, o que torna necessário o desenvolvimento de técnicas para capturar informações relacionadas a sentido, contexto e relacionamento entre palavras e expressões em um texto.

As técnicas com grafos utilizam métricas de redes complexas para derivar atributos que caracterizem a topologia de textos em língua natural (CANCHO; SOLÉ, 2001). Estas

técnicas exploram o conceito de redes de co-ocorrência, em que as palavras de um texto representam os vértices em um grafo e as arestas são estabelecidas entre palavras vizinhas.

Há uma infinidade de medidas atualmente utilizadas para caracterizar a topologia das redes complexas. As métricas, tradicionalmente, podem ser classificadas de acordo com a quantidade de informação necessária para o cálculo. Enquanto métricas locais só exigem informações sobre os vizinhos de um dado nó, as métricas globais exigem que a conectividade global da rede seja conhecida de antemão. E por último, as métricas quase-locais, exigem informações sobre vizinhos de vizinhos. Considerando a capacidade dos modelos de redes em capturar informações referentes ao relacionamento entre palavras e informações relacionadas a contexto, métodos para a aplicação das métricas de redes complexas para a análise de estilografia têm sido propostos. Estilografia diz respeito a identificação de características de texto como estilo literário e autoria. Este trabalho se concentra na avaliação destes métodos para um problema de classificação de autoria envolvendo pequenos textos em língua portuguesa. Dentre as métricas mais comumente utilizadas em redes complexas, foram selecionadas a média dos caminhos mínimos, o *betweenness* e o *clustering coefficient*. O número de vezes que cada palavra aparece no texto e uma medida de intermitência definida em (AMANCIO et al., 2011) complementam o conjunto de métricas do método proposto.

Para a construção das redes propõe-se a divisão dos discursos originais em trechos com tamanhos variados, chamados de janelas e estes trechos serão os exemplos a serem classificados. Isto é necessário devido ao tamanho limitado do conjunto de dados, então neste estudo optou-se por dividir os discursos de modo a obter mais exemplos para treinamento e classificação.

Para a validação da técnica foram selecionados discursos de cinco ex-presidentes basileiros, **Fernando Henrique Cardoso**, **Getúlio Vargas**, **João Goulart**, **Juscelino Kubitscheck** e **Luís Inácio Lula da Silva**. Os textos foram divididos em janelas de 50, 100 e 150 palavras e os experimentos foram repetidos para cada um destes tamanhos.

Também deriva do trabalho de Amancio (AMANCIO et al., 2011) a escolha das métricas utilizadas, pois os autores demonstram que medidas como a média dos caminhos mínimos, *betweenness* e a intermitência de palavras são medidas importantes para o caso específico do reconhecimento de autoria, sendo que a média dos menores caminhos e a intermitência são fortemente dependentes do estilo do autor.

A proposta desta pesquisa define variações na técnica original citada anteriormente, na tentativa de aprimorar a performance dos classificadores para textos com um número pequeno de palavras, como a já mencionada divisão do texto em trechos menores, as variações na técnica de construção da rede e o uso de autovalores da matriz Laplaciana como atributos.

Devido às propriedades do espectro Laplaciano apresentadas no capítulo 1 e ao bom desempenho de classificadores ao utilizá-las como atributos em tarefas de classificação de imagens (HUMARI, 2016), foram realizados experimentos para verificar se a melhoria também seria verificada em classificação de textos.

3.2 Metodologia

A metodologia de trabalho consiste nos seguintes passos:

- Aquisição de dados e pré-processamento.
- Construção das redes e extração de medidas.
- Geração de atributos.
- Classificação e validação.

A seguir descrevemos cada um deles em detalhes.

3.2.1 Aquisição de dados e pré-processamento

Para a realização dos experimentos, foram coletados discursos de cinco ex-presidentes brasileiros disponibilizados pela Fundação Alexandre de Gusmão[1]. Os textos são disponibilizados em formato pdf, portanto precisaram ser convertidos para um formato que facilitasse a leitura por computador. Este processo de conversão gerou caracteres espúrios que precisaram ser removidos manualmente.

Após isso os textos foram pré-processados. Os passos envolvidos são descritos a seguir:

- Todos os sinais de pontuação foram removidos.
- Todas as letras maiúsculas foram convertidas para minúsculas.
- Os textos passaram por um processo de *stemming* para a redução de todas as flexões das palavras a uma raíz comum.

No trabalho aqui proposto, optou-se por não realizar a remoção das palavras auxiliares (*stopwords*), pois é citado em (AMANCIO et al., 2011) que o uso das palavras auxiliares (artigos, preposições etc.) é fortemente dependente do estilo do autor.

[1] http:funag.gov.br/loja/

Após o pré-processamento, foram gerados três conjuntos de dados a partir dos textos originais, sendo eles:

1. Um conjunto onde todos os discursos foram divididos em trechos de 50 palavras.
2. Um conjunto onde todos os discursos foram divididos em trechos de 100 palavras.
3. Um conjunto onde todos os discursos foram divididos em trechos de 150 palavras.

A partir desses conjuntos, foram construídas redes de co-ocorrência conforme descrito na seção a seguir.

3.2.2 Construção das redes e extração de medidas

Para cada um dos três conjuntos de textos citados anteriormente, foram construídas redes de co-ocorrência de palavras. Para cada texto é criado um grafo onde as palavras são representadas por vértices e arestas são estabelecidas entre palavras vizinhas. É importante observar que para cada conjunto de dados, é gerado um grafo para cada texto e neste trabalho cada trecho do discurso é tratado como um texto diferente.

Com o objetivo de testar variações construtivas da rede, foram construídos os seguintes tipos de grafos para cada um dos conjunto de textos:

- Grafo não direcionado e não ponderado.
- Grafo não direcionado e ponderado.
- Grafo direcionado e ponderado.

Para cada um dos tipos de grafos gerados, foram extraídas, além das medidas de *betweenness* e *clustering coefficient* descritas no capítulo 1, as seguintes medidas:

1. **Média dos caminhos mais curtos:** Esta métrica é extraída calculando-se os menores caminhos a partir de um nó de referência para todos os outros nós da rede. Após isto é feita a média de todos estes valores.
2. **Número de ocorrências:** Quantidade de vezes em que uma palavra aparece no texto.
3. **Intermitência:** A intermitência é calculada considerando-se a quantidade de palavras entre duas ocorrências da palavra que esta sendo usada como referência. Por exemplo, na expressão *"A universidade não oferece curso de graduação em ciência da*

computação, porém oferece o curso de informática biomédica.", entre as duas ocorrências da palavra "oferece"há 9 palavras. Para o caso da palavra "curso"há 10 palavras entre as ocorrências. São calculados todos os intervalos entre todas as ocorrências da palavra. Este valores são organizados em um vetor $T_i = \{t_1, t_2, t_3, ..., t_{fi-1}\}$. A partir deste vetor, a intermitência é dada por

$$I \equiv \frac{\sigma_T}{\overline{T}} \tag{3.1}$$

onde

$$\sigma_T = \sqrt{\overline{T^2} - \overline{T}^2}. \tag{3.2}$$

Para os dois casos em que a rede é não direcionada, também extraímos o maior e o segundo menor autovalores da matriz Laplaciana. Para uma visão geral a respeito do espectro de grafos veja o capítulo 1.

3.2.3 Geração de atributos

Dadas as 5 medidas extraídas de cada palavra em cada texto (*betweenness*, média dos caminhos mais curtos, coeficiente de agrupamento, número de ocorrências de palavras e intermitência), e 2 autovalores (o maior e o segundo menor), os vetores característicos de cada texto foram gerados da seguinte maneira:

Foi calculada a média das métricas N, B, L, C, I gerando 5 atributos para o vetor característico.

Uma média modificada para as métricas N, B, definida como $\langle \log(X) \rangle$.
Uma segunda média modificada para as métricas L, C, I definida como $\langle \frac{X \log(N)}{\log(N)} \rangle$. Estas duas médias modificadas geram mais 5 atributos para o vetor característico.

Outro 5 atributos são gerados da seguinte forma:

Calcula-se o expoente α *em* $X^{-\alpha}$ para N e B. calcula-se o skewness para L, C, I.

Onde N é o número de ocorrências da palavra, B é o *betweenness*, L é a média dos menores caminhos, C o *clustering coefficient* e I a intermitência.

A partir destas medidas podemos construir vetores com 15 atributos. A análise detalhada de cada uma dessas medidas vem da referência (AMANCIO et al., 2011).

Alternativamente foram gerados vetores característicos com 17 atributos, incluindo também o maior e o segundo menor autovalor da matriz Laplaciana, para os casos onde a rede é não direcionada. Estes vetores foram utilizados em experimentos separados, como pode ser observado no capítulo 4.

A extração destas medidas gera 5 *datasets* para cada um dos conjuntos de textos, por exemplo, o conjunto composto por textos de 50 palavras gerará os seguintes *datasets*:

- Um dataset composto por medidas extraídas a partir de redes não direcionadas e não ponderadas, sem o uso de medidas espectrais, com vetores característicos de 15 atributos.
- Um dataset composto por medidas extraídas a partir de redes não direcionadas e ponderadas, sem o uso de medidas espectrais, com vetores característicos de 15 atributos.
- Um dataset composto por medidas extraídas a partir de redes direcionadas e ponderadas, sem o uso de medidas espectrais, com vetores característicos de 15 atributos.
- Um dataset composto por medidas extraídas a partir de redes não direcionadas e não ponderadas, utilizando medidas espectrais, com vetores característicos de 17 atributos.
- Um dataset composto por medidas extraídas a partir de redes não direcionadas e ponderadas, utilizando medidas espectrais, com vetores característicos de 17 atributos.

Os mesmos conjuntos de dados são gerados para os textos com 100 palavras e com 150 palavras, totalizando 15 *datasets*. Neste estudo aplicamos os métodos de classificação para cada um destes *datasets* separadamente afim de verificar se há diferença de performance para as diferentes formas de construção da rede.

3.3 Visualização dos dados

Muitas vezes antes de aplicar métodos de aprendizado de máquina é útil explorar o conjunto de dados através de visualizações, buscando ter um entendimento melhor de como os dados se distribuem em um dado espaço. Para isto, é necessário utilizar técnicas de redução de dimensionalidade, afim de tornar possível a construção de gráficos e imagens. Aqui aplicamos a técnica de análise de componentes principais para reduzir a dimensão dos conjuntos de dados gerados e plotar gráficos representativos destes dados. Na próxima seção descrevemos a técnica utilizada.

3.3.1 Análise de componentes principais

A Análise de Componentes Principais (PCA) é um método muito utilizado para gerar um conjunto de dados com dimensão menor a partir de transformações realizadas no conjunto

original. Uma das definições de PCA encontradas no livro de Bishop (BISHOP, 2006) diz que PCA pode ser definida como uma projeção ortogonal dos dados em um espaço linear de dimensionalidade menor.

De forma bastante simplificada, a PCA consiste no cálculo da matriz de covariância dos dados, da qual são extraídos os autovalores e autovetores, utilizados para o cálculo das componentes principais. Os autovalores são ordenados em ordem decrescente, o que nos indica quais sao as componentes principais em ordem de importância, assim podemos descartar as componentes que carregam menor informação a respeito do conjunto de dados. A descrição do método em detalhes pode ser encontrada em (BISHOP, 2006).

3.4 Classificação e validação

Após a criação das redes e a geração dos atributos derivados das medidadas de redes complexas, a fase de classificação consistiu na utilização dos seguintes algorítmos de classificação com seus parâmetros configurados com os valores padrão:

Random Forests: Este algoritmo é baseado na combinação de diversos estimadores do tipo árvore de decisão de forma a conseguir alta capacidade de generalização e imunidade a ruídos. Uma árvore de decisão é um algoritmo que considera um determinado número de possíveis *caminhos de decisão* (GRUS, 2015), onde cada caminho leva a um resultado de classificação. As *random forests* são muito utilizadas, pois costumam apresentar excelentes performance mesmo sem o ajuste fino dos parâmetros do classificador. Este método cria vários modelos de árvores de decisão e os combina com o objetivo de melhorar a acurácia da classificação. Neste trabalhos foram utilizados 128 estimadores (BREIMAN, 2001).

k-Nearest Neighbor: Este classificador é baseado no conceito de vizinhos mais próximos. Este método classificará um novo exemplo baseado na classe de seus k vizinhos mais próximos (FACELI et al., 2011).

Naive-Bayes: Este classificador utiliza como princípio o teorema de Bayes. Trata-se de um classificador probabilístico simples (FACELI et al., 2011).

Os experimentos deste trabalho foram realizados utilizando a linguagem Python[2] que possui diversos pacotes para uso em aprendizado de máquina e processamento de linguagem natural. Os pacotes utilizados durante o desenvolvimento deste projeto estão listados no apêndice A. Muitos detalhes práticos a respeito de tecnologia e de formas de implementação deste tipo de experimento podem ser encontrados no livro de Grus (GRUS, 2015).

[2] https://www.python.org/

Os resultados obtidos a partir deste experimentos, serão analisados no capítulo 4.

3.4.1 Validação cruzada

A técnica mais simples de validação em aprendizado de máquina é chamada de ***holdout*** e consiste em dividir o conjunto de dados em dois subconjuntos, um para treinamento, tipicamente contendo 70% das amostras e um conjunto para testes, tipicamente contendo 30% das amostras. Caso o conjunto de dados seja muito pequeno, o erro de predição pode sofrer grandes variações dependendo de quais amostras são selecionadas para treinamento e testes. Assim, a técnica de ***houldout*** é indicada apenas em situações em que o conjunto de dados contém uma grande quantidade de exemplos.

Os conjuntos de dados utilizados nos experimentos deste trabalho possuem uma quantidade pequena de exemplos, portanto utilizaremos a técnica conhecida como ***10-fold cross-validation***. Esta técnica consistem em dividir o conjunto de dados em 10 subconjuntos, por amostragem, e utilizar 9 conjuntos para treinamento e 1 para testes. São feitas 10 rodadas de treinamento e testes, sendo que em cada rodada o conjunto de testes muda, de acordo com a Figura 10. Em cada rodada calculamos o F_1 *Score*, definido na seção a seguir, e fazemos o cálculo da média e do desvio padrão.

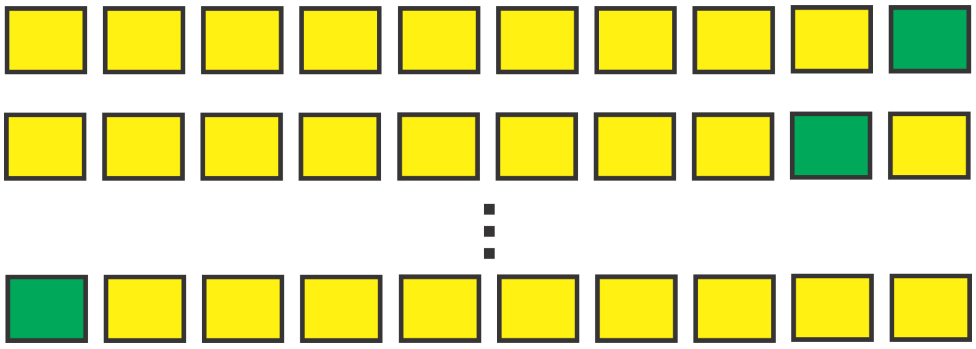


Figura 10 – 10-*fold cross-validation*.

3.5 Avaliação de modelos preditivos

Nesta seção apresentaremos as principais métricas utilizadas em Aprendizado de Máquina para a avaliação de modelos preditivos.

- **Acurácia**: A acurácia é definida como a divisão do número de predições corretas realizadas pelo número total de predições:

$$acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.3}$$

A acurácia nem sempre é uma boa medida para quantificarmos a qualidade de um modelo de aprendizado de máquina, principalmente quando os dados são desbalanceados, pois a mesma ignora a existência de falsos positivos.

Onde VP = Verdadeiros Positivos, VN = Verdadeiros Negativos, FP = Falsos Positivo, FN = Falsos Negativos.

- **Precisão**: Esta medida busca respoder à pergunta *"Qual a proporção dos exemplos classificados como positivos que está verdadeiramente correta?"*. A precisão é dada pela expressão:

$$precisão = \frac{VP}{VP + FP} \tag{3.4}$$

- ***Recall***: Esta medida busca responder à pergunta *"Quando um exemplo realmente pertence a uma classe, com que frequência ele é classificado como tal?"*. A medida de ***recall*** é dada pela expressão:

$$recall = \frac{VP}{VP + FN} \tag{3.5}$$

- F_1 ***Score***: A medida F_1 é definida como a média harmônica da ***precisão*** e do ***recall*** de modo a obter uma única métrica da qualidade geral do modelo. Esta medida é utilizada quando o conjunto de dados contém classes desbalanceadas e é dada pela expressão:

$$F_1 = 2 \times \frac{precisão \times recall}{precisão + recall} \tag{3.6}$$

  Esta é a medida que será utilizada na apresentação dos resultados deste estudo.

- **Matriz de confusão**: A matriz de confusão é uma forma de resumir o desempenho de um classificador apresentando em formato de tabela os erros e os acertos para cada uma das classe.

  Nesta tabela, as colunas representam a classe encontrada pelo algoritmo e as linhas representam a classe real. Desta forma podemos determinar facilmente quantos exemplos foram classificados corretamente. A seguir, um exemplo de matriz de confusão:

$$\begin{array}{c} \\ FHC \\ Lula \\ Jango \\ Getulio \\ JK \end{array} \begin{array}{c} \begin{array}{ccccc} FHC & Lula & Jango & Getulio & JK \end{array} \\ \begin{pmatrix} 446 & 0 & 0 & 24 & 25 \\ 0 & 244 & 13 & 14 & 0 \\ 0 & 17 & 333 & 0 & 0 \\ 22 & 23 & 0 & 405 & 0 \\ 20 & 0 & 0 & 0 & 377 \end{pmatrix} \end{array}$$

  Analisando a matriz, observe que dos exemplos que pertencem à classe *"FHC"*, 446 foram classificados corretamente, 24 foram classificados como pertencentes à

classe *"Getulio"*, 25 pertencentes à classe *"JK"* e nenhum exemplo classificado como pertencente às classes restantes.

É possível encontrar uma revisão geral sobre avaliação de modelos preditivos na referência (FACELI et al., 2011) e no *website* ***Machine Learning Crash Course***[3].

[3] https://developers.google.com/machine-learning/crash-course/

4

Resultados

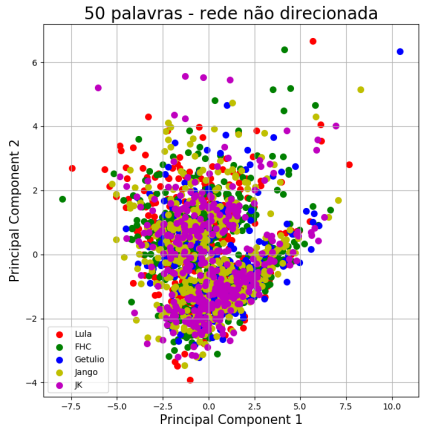
A seguir fazemos uma análise dos conjuntos de dados gerados para diferentes construções de rede e apresentamos uma visualização dos dados através da técnica de *Análise de Componentes Principais* apresentada no capítulo 3.

4.1 Importância dos diferentes atributos para a classificação

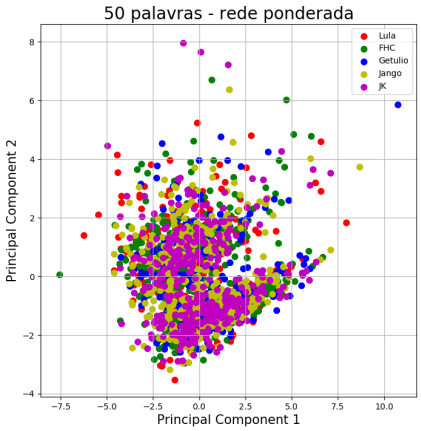
No artigo (AMANCIO et al., 2011), os autores analisam a importância relativa dos atributos gerados para os textos através do estudo da dependência mútua entre os atributos. Na referência citada é possível encontrar um ranking dos atributos para cada um dos classificadores utilizados. Os três atributos apresentados como mais influentes são a **média no número de ocorrências (frequência) das palavras**, o ***skewness* da intermitência** e **a média logarítmica do número de ocorrências das palavras**. Aqui fazemos uma análise visual de cada um dos conjuntos de dados gerados através de diferentes técnicas de construção da rede, incluindo medidas espectrais não analisadas no artigo utilizado como referência.

4.2 Visualização dos dados por análise de componentes principais (PCA)

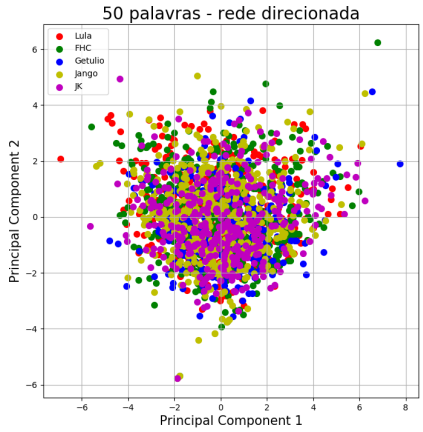
Com o objetivo de fazer uma análise preliminar dos dados gerados, aplicamos a técnica de PCA, para redução de dimensionalidade, de modo a obter um conjunto de dados com apenas duas dimensões, possibilitando a representação gráfica dos dados.



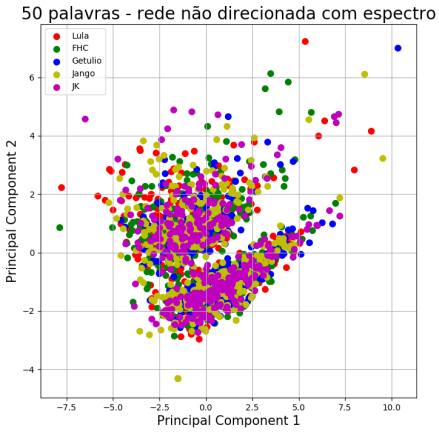
(a) Rede não direcionada e não ponderada sem valores espectrais.



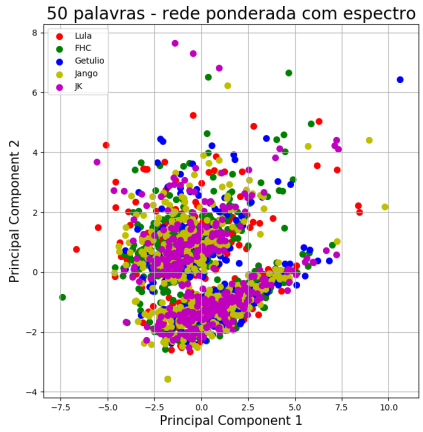
(b) Rede não direcionada e ponderada. Sem valores espectrais.



(c) Rede direcionada e ponderada. Sem valores espectrais.

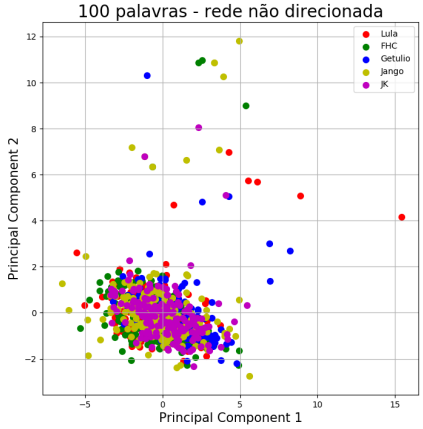


(d) Rede não direcionada e não ponderada. Com valores espectrais.

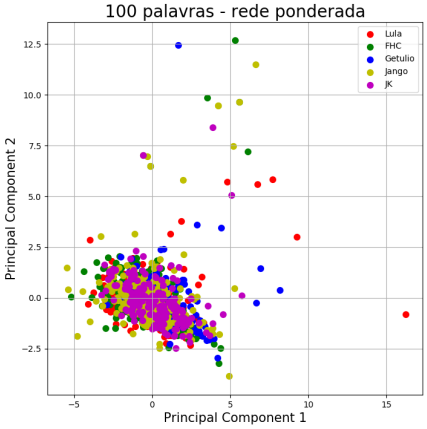


(e) Rede não direcionada e ponderada. Com valores espectrais.

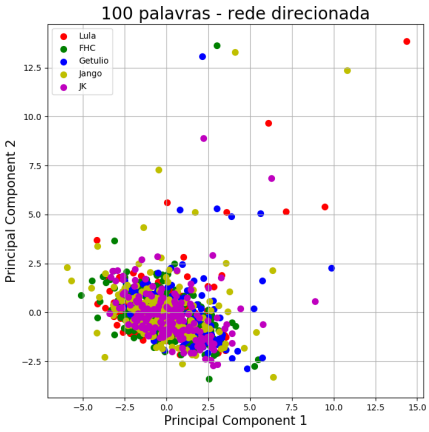
Figura 11 – Análise dos conjuntos de dados em duas dimensões para textos divididos em janelas de 50 palavras.



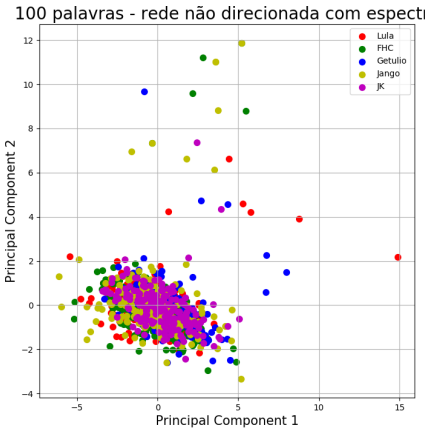
(a) Rede não direcionada e não ponderada sem valores espectrais.



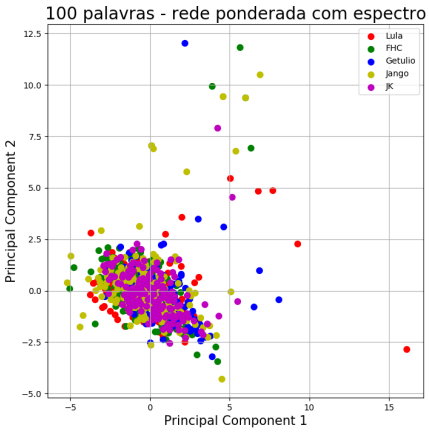
(b) Rede não direcionada e ponderada. Sem valores espectrais.



(c) Rede direcionada e ponderada. Sem valores espectrais.

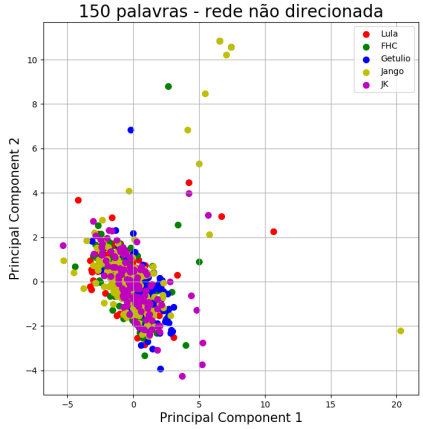


(d) Rede não direcionada e não ponderada. Com valores espectrais.

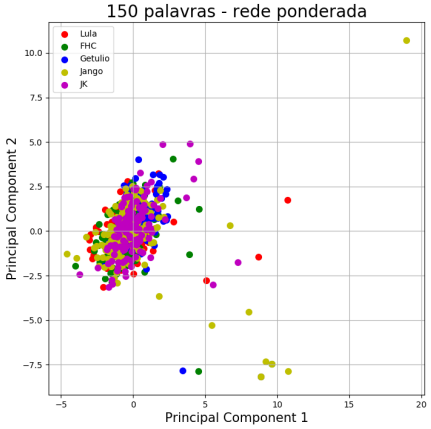


(e) Rede não direcionada e ponderada. Com valores espectrais.

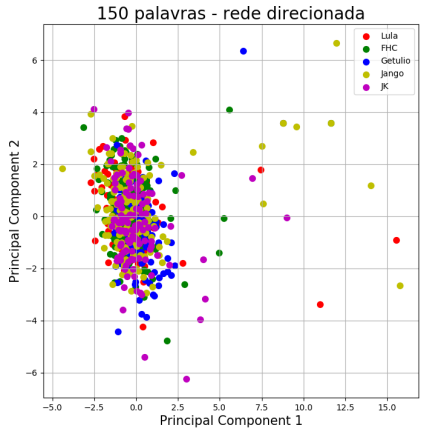
Figura 12 – Análise dos conjuntos de dados em duas dimensões para textos divididos em janelas de 100 palavras.



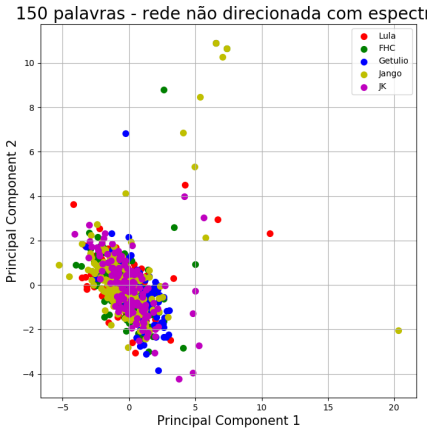
(a) Rede não direcionada e não ponderada sem valores espectrais.



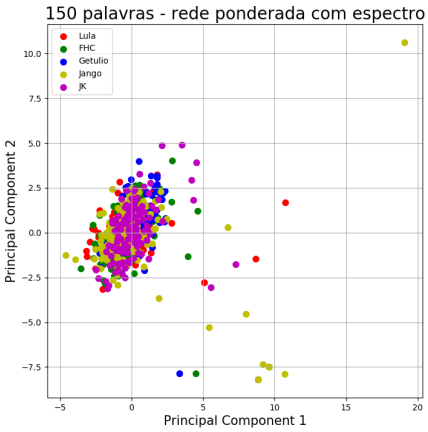
(b) Rede não direcionada e ponderada. Sem valores espectrais.



(c) Rede direcionada e ponderada. Sem valores espectrais.



(d) Rede não direcionada e não ponderada. Com valores espectrais.



(e) Rede não direcionada e ponderada. Com valores espectrais.

Figura 13 – Análise dos conjuntos de dados em duas dimensões para textos divididos em janelas de 150 palavras.

Analisando as figuras 11, 12 e 13 podemos observar a maneira como os dados se comportam quando as redes são construídas de diferentes formas, porém em termos aparentes não é possível observar nenhuma melhora em relação a separação dos dados por categorias específicas. Na próxima seção apresentamos os resultados obtidos pelos diferentes algoritmos de classificação aplicados aos diferentes conjuntos de dados.

4.3 Resultados da classificação

Nesta seção apresentamos os resultados obtidos ao aplicarmos a técnica desenvolvida por Amancio e colaboradores (AMANCIO et al., 2011) (AMANCIO, 2013) a conjuntos de dados gerados a partir de discursos presidenciais, com variações na técnica de construção das redes e com a inclusão de dois atributos gerados a partir do espectro laplaciano destas redes. Para a validação dos modelos preditivos utilizamos a técnica de *cross-validation* com 10 *folds* e calculamos a métrica F_1 *Score* com intervalo de confiança de 95% para cada um dos exemplos de classificação. Também apresentamos a matriz de confusão para o classificador que apresentou o maior F_1 *Score* em cada uma das subseções a seguir. Detalhes sobre cada uma das técnicas e métricas utilizadas podem ser encontrados no capítulo 3 sobre metodologia.

4.3.1 Janelas de 50 palavras com e sem espectro

Tabela 1 – Classificação utilizando rede não direcionada e não ponderada com janelas de 50 palavras. Sem atributos do espectro laplaciano.

Classificador	**F_1 *Score***
Random Forest (128 estimadores)	0.92(±0.32)
KNN (k = 1)	0.65(±0.16)
KNN (k = 2)	0.60(±0.14)
KNN (k = 3)	0.70(±0.19)
Naive Bayes Gaussiano	0.91(±0.31)

Para janelas de 50 palavras é possível observar que não há diferenças significativas na performance dos classificadores para as diversas formas de se construir a rede. A seguir também apresentamos, como exemplo, a matriz de confusão para o classificador com o maior F_1 *Score*, *Random Forest*, quando aplicado ao conjunto de dados formado a partir de uma rede não direcionada e ponderada utilizando atributos do espectro laplaciano.

Tabela 2 – Classificação utilizando rede não direcionada e ponderada com janelas de 50 palavras. Sem atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.32)
KNN (k = 1)	0.65(±0.15)
KNN (k = 2)	0.60(±0.14)
KNN (k = 3)	0.70(±0.20)
Naive Bayes Gaussiano	0.92(±0.31)

Tabela 3 – Classificação utilizando rede direcionada e ponderada com janelas de 50 palavras. Sem atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.32)
KNN (k = 1)	0.66(±0.16)
KNN (k = 2)	0.60(±0.16)
KNN (k = 3)	0.70(±0.16)
Naive Bayes Gaussiano	0.92(±0.32)

Tabela 4 – Classificação utilizando rede não direcionada e não ponderada com janelas de 50 palavras utilizando atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.32)
KNN (k = 1)	0.64(±0.15)
KNN (k = 2)	0.60(±0.12)
KNN (k = 3)	0.68(±0.18)
Naive Bayes Gaussiano	0.90(±0.34)

Tabela 5 – Classificação utilizando rede não direcionada e ponderada com janelas de 50 palavras utilizando atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.32)
KNN (k = 1)	0.64(±0.15)
KNN (k = 2)	0.59(±0.16)
KNN (k = 3)	0.68(±0.19)
Naive Bayes Gaussiano	0.90(±0.34)

	FHC	*Lula*	*Jango*	*Getulio*	*JK*
FHC	446	0	0	24	25
Lula	0	244	13	14	0
Jango	0	17	333	0	0
Getulio	22	23	0	405	0
JK	20	0	0	0	377

4.3.2 Janelas de 100 palavras com e sem espectro

Tabela 6 – Classificação utilizando rede não direcionada e não ponderada com janelas de 100 palavras. Sem atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.32)
KNN (k = 1)	0.64(±0.15)
KNN (k = 2)	0.60(±0.17)
KNN (k = 3)	0.70(±0.15)
Naive Bayes Gaussiano	0.92(±0.28)

Tabela 7 – Classificação utilizando rede não direcionada e ponderada com janelas de 100 palavras. Sem atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.32)
KNN (k = 1)	0.66(±0.14)
KNN (k = 2)	0.61(±0.14)
KNN (k = 3)	0.72(±0.17)
Naive Bayes Gaussiano	0.92(±0.30)

Tabela 8 – Classificação utilizando rede direcionada e ponderada com janelas de 100 palavras. Sem atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.32)
KNN (k = 1)	0.68(±0.13)
KNN (k = 2)	0.60(±0.15)
KNN (k = 3)	0.70(±0.16)
Naive Bayes Gaussiano	0.92(±0.27)

Tabela 9 – Classificação utilizando rede não direcionada e não ponderada com janelas de 100 palavras utilizando atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.32)
KNN (k = 1)	0.61(±0.15)
KNN (k = 2)	0.58(±0.16)
KNN (k = 3)	0.65(±0.15)
Naive Bayes Gaussiano	0.92(±0.27)

Assim como na seção anterior, com 100 palavras é possível observar que não há diferenças significativas. A seguir também apresentamos matriz de confusão para o

Tabela 10 – Classificação utilizando rede não direcionada e ponderada com janelas de 100 palavras utilizando atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.32)
KNN (k = 1)	0.64(±0.13)
KNN (k = 2)	0.60(±0.16)
KNN (k = 3)	0.67(±0.12)
Naive Bayes Gaussiano	0.92(±0.27)

classificador com o maior F_1 *Score*, *Random Forest*, quando aplicado ao conjunto de dados formado a partir de uma rede não direcionada e ponderada utilizando atributos do espectro laplaciano.

	FHC	*Lula*	*Jango*	*Getulio*	*JK*
FHC	223	0	0	12	13
Lula	0	127	6	7	0
Jango	0	9	168	0	0
Getulio	11	11	0	207	0
JK	10	0	0	0	192

4.3.3 Janelas de 150 palavras com e sem espectro

Tabela 11 – Classificação utilizando rede não direcionada e não ponderada com janelas de 150 palavras. Sem atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.30)
KNN (k = 1)	0.68(±0.08)
KNN (k = 2)	0.63(±0.11)
KNN (k = 3)	0.69(±0.14)
Naive Bayes Gaussiano	0.91(±0.23)

Com 150 palavras o comportamento se mantém o mesmo em relação aos experimentos com número menor de palavras, ou seja, sem diferenças significativas em relação a forma de construção da rede. Nesta seção também apresentamos matriz de confusão para o classificador com o maior F_1 *Score*, *Random Forest*, quando aplicado ao conjunto de dados formado a partir de uma rede não direcionada e ponderada utilizando atributos do espectro laplaciano. No capítulo 5 fazemos um resumo e discutimos os resultados aqui apresentados.

Tabela 12 – Classificação utilizando rede não direcionada e ponderada com janelas de 150 palavras. Sem atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.31)
KNN (k = 1)	0.66(±0.10)
KNN (k = 2)	0.62(±0.15)
KNN (k = 3)	0.70(±0.19)
Naive Bayes Gaussiano	0.91(±0.23)

Tabela 13 – Classificação utilizando rede direcionada e ponderada com janelas de 150 palavras. Sem atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.32)
KNN (k = 1)	0.70(±0.14)
KNN (k = 2)	0.65(±0.17)
KNN (k = 3)	0.71(±0.17)
Naive Bayes Gaussiano	0.92(±0.26)

Tabela 14 – Classificação utilizando rede não direcionada e não ponderada com janelas de 150 palavras utilizando atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.30)
KNN (k = 1)	0.63(±0.10)
KNN (k = 2)	0.57(±0.13)
KNN (k = 3)	0.66(±0.15)
Naive Bayes Gaussiano	0.92(±0.23)

Tabela 15 – Classificação utilizando rede não direcionada e ponderada com janelas de 150 palavras utilizando atributos do espectro laplaciano.

Classificador	$\mathbf{F_1}$ ***Score***
Random Forest (128 estimadores)	0.92(±0.30)
KNN (k = 1)	0.61(±0.10)
KNN (k = 2)	0.57(±0.10)
KNN (k = 3)	0.66(±0.18)
Naive Bayes Gaussiano	0.90(±0.25)

	FHC	*Lula*	*Jango*	*Getulio*	*JK*
FHC	152	0	0	8	8
Lula	0	84	4	5	0
Jango	0	6	115	0	0
Getulio	8	6	0	142	0
JK	7	0	0	0	128

5

Conclusão

Neste trabalho aplicamos a técnica desenvolvida por Amâncio (AMANCIO et al., 2011) (AMANCIO, 2013) para o problema de reconhecimento de autoria em textos pequenos, tendo como objeto de estudo discursos presidenciais. É de conhecimento geral que tais discursos tipicamente não são redigidos pelos presidentes em questão, nem ao menos são redigidos por uma única pessoa. Apesar destas particularidades, os classificadores apresentaram um desempenho aceitável, indicando que tais discursos guardam elementos de estilo que são típicos de cada um dos presidentes analisados.

Os experimentos realizados buscaram fazer modificações na técnica original com o objetivo de melhorar o desempenho dos classificadores, porém tais modificações não mostraram grande efetividade, mantendo um desempenho parecido com o dos experimentos realizados pelo autor dos trabalhos utilizados como base para esta dissertação. As principais modificações realizadas dizem respeito a construção da rede, sendo utilizadas as seguintes variações:

1. Rede não direcionada e não ponderada
2. Rede não direcionada e ponderada
3. Rede direcionada e ponderada

Sendo que para os casos onde a rede era **"não direcionada e não ponderada"** e onde era **"não direcionada e ponderada"** também foram realizados experimentos utilizando medidas espectrais da rede.

A análise do capítulo 4 mostra que alguns casos, os classificadores apresentaram F_1 *Score* de 0.92, porém com desvio padrão muito alto entre as rodadas de classificação utilizando ***10-fold cross-validation***. Os algoritmos que apresentaram menor desvio padrão, e portanto mais consistência entre os experimentos, apresentaram F_1 *Score* entre 0.6 e 0.7, o que é desempenho aceitável para a técnica proposta.

Para a realização dos experimentos foi necessário dividir os textos em subconjuntos ou janelas, de modo a obter um conjunto de dados maior para classificação. Então os experimentos tratam de atribuir autoria a estes subconjuntos. Foram utilizados subconjuntos de tamanhos variados, com 50, 100 e 150 palavras, sendo que não houve diferença significativa no desempenho dos classificadores em tais variações. Este resultado também demonstra que com a técnica utilizada, 50 palavras são suficientes para obter um desempenho considerável.

Apesar do surgimento de novas formas de abordar o problema da categorização textual, como as descritas aqui e também técnicas baseadas em *Deep Learning* (JOHNSON; ZHANG, 2014), as técnicas tradicionais de categorização textual (SEBASTIANI, 2002) continuam sendo relevantes, pois apresentam boa acurácia e já estão amplamente testadas e documentadas na literatura.

A seguir listamos algumas ideias de melhorias ou expansão deste trabalho para serem utilizadas no futuro.

5.0.1 Sugestões para trabalhos futuros

- Otimizar o tamanho das janelas para obter o número mínimo de palavras necessárias para o reconhecimento de autoria.
- Aplicar a mesma metodologia desta dissertação para classificação de gênero literário ou outros problemas de estilografia.
- Analisar em profundidade a influência dos autovalores do espectro laplaciano para grafos gerados a partir de textos.

Referências

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics*, APS, v. 74, n. 1, p. 47, 2002.

AMANCIO, D. R. *Classificação de textos com redes complexas*. Tese (Doutorado) — Universidade de São Paulo, 2013.

AMANCIO, D. R. Probing the topological properties of complex networks modeling short written texts. *PloS one*, Public Library of Science, v. 10, n. 2, p. e0118394, 2015.

AMANCIO, D. R. et al. Comparing intermittency and network measurements of words and their dependence on authorship. *New Journal of Physics*, IOP Publishing, v. 13, n. 12, p. 123024, 2011.

ANTIQUEIRA, L. et al. Modelando textos como redes complexas. In: *Anais do III Workshop em Tecnologia da Informação e da Linguagem Humana*. [S.l.: s.n.], 2005. p. 22–26.

BARÁBASI, A.-l. Linked: a nova ciência dos networks. *São Paulo: Leopardo Editora*, 2009.

BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. *science*, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999.

BISHOP, C. M. *Pattern recognition and machine learning*. [S.l.]: springer, 2006.

BOCCALETTI, S. et al. Complex networks: Structure and dynamics. *Physics reports*, Elsevier, v. 424, n. 4, p. 175–308, 2006.

BRANDES, U.; ERLEBACH, T. *Network analysis: methodological foundations*. [S.l.]: Springer Science & Business Media, 2005. v. 3418.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

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

CORMEN, T. H. et al. *Algoritmos: teoria e prática*. [S.l.: s.n.], 2012. v. 3.

COSTA, L. d. F. et al. Characterization of complex networks: A survey of measurements. *Advances in physics*, Taylor & Francis, v. 56, n. 1, p. 167–242, 2007.

ERDÖS, P.; RÉNYI, A. On random graphs, i. *Publicationes Mathematicae (Debrecen)*, v. 6, p. 290–297, 1959.

EULER, L. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, v. 8, p. 128–140, 1741.

FACELI, K. et al. *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. [S.l.: s.n.], 2011.

FEOFILOFF, P.; KOHAYAKAWA, Y.; WAKABAYASHI, Y. Uma introdução sucinta à teoria dos grafos. 2011.

FRITSCHER, E. *Propriedades espectrais de um grafo*. 2011.

GRUS, J. *Data science from scratch: first principles with python*. [S.l.]: "O'Reilly Media, Inc.", 2015.

HUMARI, J. H. C. *Estudo do espectro Laplaciano na categorização de imagens*. Tese (Doutorado), 2016.

JOHNSON, R.; ZHANG, T. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*, 2014.

JURAFSKY, D.; MARTIN, J. *Speech & language processing (3rd ed. draft)*. https://web.stanford.edu/ jurafsky/slp3/: [s.n.], 2018.

MANNING, C. D.; SCHÜTZE, H. *Foundations of statistical natural language processing*. [S.l.]: MIT Press, 1999. v. 999.

METZ, J. et al. Redes complexas: conceitos e aplicações. *Relatórios Técnicos do ICMC-USP São Carlos*, 2007.

MILGRAM, S. The small world problem. *Psychology today*, New York, v. 2, n. 1, p. 60–67, 1967.

MOHAR, B. et al. The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications*, v. 2, n. 871-898, p. 12, 1991.

NETTO, P. O. B. *Grafos: teoria, modelos, algoritmos*. [S.l.]: Edgard Blücher, 2003.

RESTREPO, J. G.; OTT, E.; HUNT, B. R. Characterizing the dynamical importance of network nodes and links. *Physical review letters*, APS, v. 97, n. 9, p. 094102, 2006.

ROSEN, K. H. *Discrete Mathematics and Its Applications*. 5th. ed. [S.l.]: McGraw-Hill Higher Education, 2002. ISBN 0072424346.

SANTOS, P.; RANGEL, M.; BOERES, M. *Teoria espectral de grafos aplicada ao problema de isomorfismo de grafos*. 2010.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, ACM, v. 34, n. 1, p. 1–47, 2002.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. *nature*, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998.

ZIPF, G. K. *Human behavior and the principle of least effort: An introduction to human ecology*. [S.l.]: Ravenio Books, 1972.

Apêndices

A

Links úteis

Lista de links para acessar os datasets e pacotes computacionais utilizados.

1. Coleção de discursos presidenciais (www.funag.gov.br)
    - Getúlio Vargas (https://goo.gl/1CvSJz)
    - João Goulart (https://goo.gl/5Sykfr)
    - Juscelino Kubitschek (https://goo.gl/Xv8smr)
    - Fernando Henrique Cardoso (https://goo.gl/RqT1uZ)
    - Lula (https://goo.gl/YD2TGQ)
2. Pacotes computacionais
    - scikit-learn (https://scikit-learn.org/stable/)
    - pandas (https://pandas.pydata.org/)
    - NetworkX (https://networkx.github.io/)