RECONHECIMENTO DE AUTORIA DE TEXTOS UTILIZANDO REDES COMPLEXAS

Trabalho de Conclusão de Curso
Engenharia da Computação

Marcello Angelis Coutinho de Medeiros
Orientador: Prof. Dr. Carmelo José Albanez Bastos Filho

Marcello Angelis Coutinho de Medeiros

RECONHECIMENTO DE AUTORIA DE TEXTOS UTILIZANDO REDES COMPLEXAS

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia de Computação pela Escola Politécnica de Pernambuco - Universidade de Pernambuco

Universidade de Pernambuco

Escola Politécnica de Pernambuco

Graduação em Engenharia de Computação

Orientador: Prof. Dr. Carmelo José Albanez Bastos Filho

Recife - PE, Brasil

27 de novembro de 2015

Declaro que revisei o Trabalho de Conclusão de Curso sob o título *"RECONHECIMENTO DE AUTORIA DE TEXTOS UTILIZANDO REDES COMPLEXAS"*, de autoria de *Marcello Angelis Coutinho de Medeiros*, e que estou de acordo com a entrega do mesmo.

Recife, _____ / _____ / _____

_____

Prof. Dr. Carmelo José Albanez Bastos Filho
Orientador

À minha Família.

Resumo

A estilografia e o reconhecimento de autoria vêm sendo frequentemente objeto de estudo de vários pesquisadores ao redor do mundo. Por tratarem os textos majoritariamente por suas propriedades estatísticas, sofreram grande avanço nos últimos anos em conjunto com as técnicas de apredizagem de máquina. Paralelamente, uma nova abordagem para representar e modelar sistemas complexos tem ganhado força: as redes complexas. Elas têm modelado muito satisfatoriamente vários sistemas reais, entre eles, textos. A rede textual mais utilizada para reconhecimento de autoria é a de coocorrência de palavras. Neste trabalho é abordada uma perspectiva não presente na literatura para criar redes a partir de textos. Para que essas redes possam ser criadas, várias etapas de pré-processamento são requeridas. Entre elas, podem ser citadas a segmentação de sentenças e o *part-of-speech tagging*. Pares de classes gramaticais são associados a outros pares vizinhos, criando um grafo de coocorrência desses pares. Várias métricas de redes complexas são extraídas do grafo que representa a rede, e desses valores são extraídas medidas que tentam representar as características estilísticas do autor. O conjunto de características é apresentado a um classificador que reconhecerá quem escreveu um certo texto dado. Os resultados são interessantes e sugerem que palavras comumente não utilizadas para o reconhecimento de autoria, na verdade, podem ser bastante úteis quando aplicadas às redes geradas neste trabalho.

Palavras-chave: Redes Complexas, Classficação textual, Processamento de texto, Estilometria, Reconhecimento de padrões, Perceptron de múltiplas camadas.

Abstract

The stylography and authorship recognition have been frequently studied by many researchers around the world. By treating the texts mainly by its statistical properties, these fields suffered major advance in recent years. Parallelly, a new approach for representing and model complex systems have been obtaining visibility: the complex networks. They have satisfactorily modeled several real systems, among them, texts. The most used text network in authorship recognition is the co-occurrence of words. In this work, a new perspective to create networks from texts is addressed. Many pre-processing steps are required to build these networks. As examples, we can cite sentence segmentation and part-of-speech tagging. We aim to associate part-of-speech pairs with other pairs to build a co-occurrence graph of these pairs. Several complex networks metrics are extracted from the graph that represents the networks. From these values graph characterizing measures are obtained which can represent the author's stylistic properties. The set of features is then presented to a classifier, that recognizes who wrote that text. There are interesting results in this work. They suggest that commonly rejected words (known as stopwords) are useful for the network model proposed by this work.

Keywords: Complex Networks, Text Classification, Text Processing, Stylometry, Pattern Recognition, Multilayer Perceptron.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

MLP	Multilayer Perceptron
EUA	Estados Unidos da América
BA	Barabási-Albert
C	Coeficiente de Aglomeração (*Clustering Coefficient*)
B	Betweenness
ACD	Arthur Conan Doyle
BS	Bram Stoker
CD	Charles Dickens
EAP	Edgar Allan Poe
HHM	Hector Hugh Munro
MK	Mark Twain
PGW	Pelham Grenville Wodehouse
TH	Thomas Hardy

Sumário

1 Introdução

1.1 Qualificação do Problema

Desde o início do século XIX, há o interesse de se determinar características estilísticas de um texto por medidas estatísticas. Muitos trabalhos foram propostos sobre o tema e sua grande maioria continua operando estatisticamente como, por exemplo, calculando a frequência de constituites do texto (palavras, frases, letras, etc). Atualmente, a análise estilística (estilometria) de textos tem aplicações em forense e mineração da web.

Um dos componentes da estilometria é o reconhecimento de autoria. Assim como o primeiro, esse reconhecimento se dá fundamentalmente por explorações nos padrões estatísticos dos textos. Entretanto, recentemente uma nova abordagem tem sido utilizada: modelar os textos como redes complexas e extrair características textuais através de propriedades topológicas das redes que os representa. Esta alternativa tem se mostrado bastante promissora, como pode ser atestado pelos resultados obtidos em (AMANCIO, 2013)(ANTIQUEIRA et al., 2005).

Este trabalho continua a empregar os métodos citados acima, mas, além disso, propõe um método de associação entre palavras que até então não foi concebido pela literatura.

1.2 Objetivos

Este trabalho objetivou reconhecer a autoria de um texto modelando-o como um grafo. Para isso, uma nova abordagem de criação de grafos de textos foi utilizada em conjunto com o uso de um eficaz e reconhecido aproximador de funções.

1.2.1 Objetivos Específicos

1. Propor uma novo padrão de associação entre os elementos que constituem os textos
2. Identificar quais as métricas que melhor caracterizam as propriedades estilísticas dos grafos que modelam textos;
3. Verificar se o modelo encontrado apresenta viabilidade da classificação de autoria.

1.3 Estrutura da Monografia

Esta monografia está dividida em 6 capítulos. No Capítulo 1, são introduzidos os conceitos primodiais que sustentam este trabalho: grafos, redes complexas, processamento de texto (principamente o *part-of-speech tagging*) e redes neurais perceptron de múltiplas camadas. O Capítulo 4 desenvolve acerca da obtenção dos livros, seus pré-processamentos, divisões, criação e caracterização dos grafos, assim como sua classificação por meio das MLPs. O Capítulo 5 é responsável por apresentar os resultados e comentá-los. No 6º e último capítulo, são apresentadas as conclusões e são listados alguns possíveis trabalhos futuros que potencialmente podem melhorar os resultados até agora obtidos.

2 Referencial Teórico

2.1 Grafos

Um grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ é definido por um conjunto $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ de vértices e por um conjunto $\mathcal{E} = \{e_1, e_2, \ldots, e_m\}$ de arestas tais que $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ e $\mathcal{V} \cap \mathcal{E} = \emptyset$. Eles podem ser direcionados (figura 2) ou não (figura 1) No primeiro caso, suas arestas são pares ordenados de vértices, já no segundo, a ordem é indiferente. Além disso, valores numéricos podem ser associados a cada aresta. Se este for o caso, o grafo é denominado ponderado. Um caminho de comprimento L é uma sequência alternada $v_1 e_1 v_2 e_2 v_3 \ldots v_{L-1} e_{L-1} v_L$ de vértices v_i e arestas e_i de forma que o vértice de destino de e_i é v_i, e seu vértice de destino, v_{i+1}. Um cíclo é um caso particular de caminho onde $v_1 = v_L$. Diz-se que um grafo é conectado quando existe ao menos um caminho conectando todos os pares de vértices.

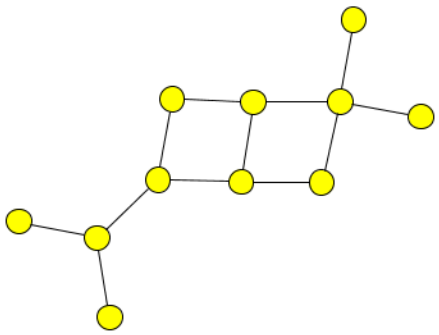

Figura 1 – Exemplo de grafo não direcionado simples.

[Fonte: reproduzido de http://graphml.graphdrawing.org/primer/simple.png]

Um grafo $\mathcal{G}$ com n vértices pode ser representado por uma matriz $\mathcal{A} = [a_{ij}]$ $n \times n$ como segue:

$$a_{ij} = \begin{cases} w_{ij}, & \text{se existe uma aresta de } v_i \text{ para } v_j \\ 0, & \text{caso contrário .} \end{cases} \tag{2.1}$$

Onde:

w_{ij} : é o peso da aresta entre os vértices v_i e v_j.

Vértices podem ser nomeados de diversas maneiras: nós, pontos, agentes, etc. Arestas também: ligações, enlaces, relações, etc.

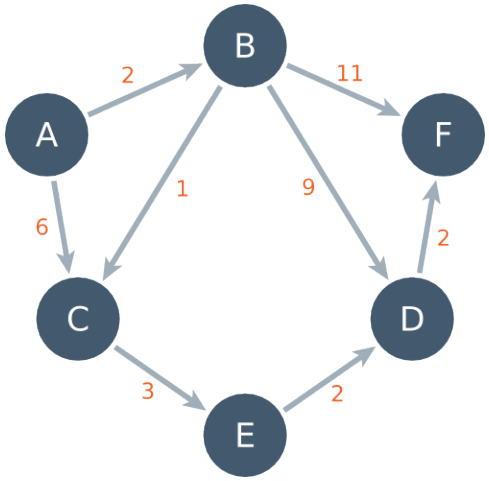

Figura 2 – Exemplo de grafo direcionado ponderado.

[Fonte: reproduzido de
http://www.hipster4j.org/assets/css/custom/img/graph-example-hipster-blue.png]

2.2 Redes Complexas

A pesquisa em redes complexas pode ser definida como a intersecção entre a teoria dos grafos e a mecânica estatística, o que lhe garante uma forte natureza interdisciplinar (COSTA et al., 2005). Por muito, se pensou que os relacionamentos entre vértices eram representados por padrões aleatórios, e um importantíssimo modelo de formação de grafos aleatórios foi proposto (ERDÖS; RÉNYI, 1960). Contudo, foi descoberto que as redes reais não se comportam como redes aleatórias, mas sim como explicado em importantes modelos, como os livres de escala (BARABáSI; ALBERT, 1999) e os de redes *small-world* (WATTS; STROGATZ, 1998).

Podem ser mencionados dois importantes motivos para a popularidade das redes complexas:

1. grande parte dos sistemas complexos são modelados por sistemas de equações diferenciais cuja solução analítica não é viável na prática;

2. as redes complexas apresentam grande flexibilidade e poder de generalização para representar virtualmente qualquer estrutura natural, incluindo aquelas dotadas de mudanças topológicas dinâmicas (COSTA et al., 2005)

Redes complexas têm sido estudas e aplicadas em várias de áreas da ciência (WANG; CHEN, 2003). Dentre as várias redes reais amplamente estudas estão:

1. a internet: rede de computadores e roteadores de proporções colossais (figura 4);

2. o cérebro humano (BULLMORE; SPORNS, 2009) (SPORNS, 2010);

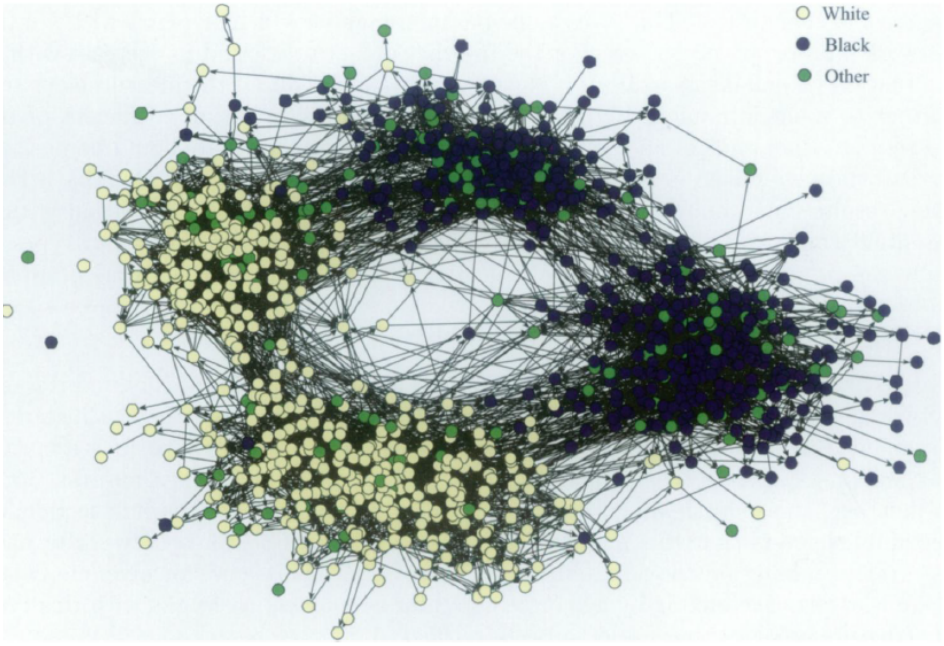

Figura 3 – Rede de amizades de crianças em uma escola dos EUA. Trata-se de um grafo direcionado pois uma criança A pode alegar ser amiga de outra criança B, mas não necessariamente o inverso. Os vértices são coloridos em função da cor da pele. A divisão horizontal da rede reflete o distanciamento entre indivíduos de etnias diferentes, enquanto a vertical remete à faixa etária.

[Fonte: (NEWMAN, 2003)]

3. redes de interação de proteínas;
4. redes sociais (figura 3);
5. redes de transmissão elétrica

2.2.1 Métricas de redes complexas

Medidas estrurais são frequentemente utilizadas para capturar informações topológicas acerca da rede. Existem miríades delas. Esta monografia utiliza algumas métricas que se mostraram potencialmente úteis na tarefa de distinguir traços estilísticos de textos em outros trabalhos (AMANCIO et al., 2011)(AMANCIO; OLIVEIRA; COSTA, 2012)(AMANCIO, 2013). Outras medidas, como o *Pagerank*, também foram utilizadas, pois conceitualmente capturam características de centralidade com eficácia.

2.2.1.1 Grau e força

O grau é uma simples e importante medida de centralidade. Ele pode ser assim classificado, pois vértices que apresentam altos valores de grau são mais importantes

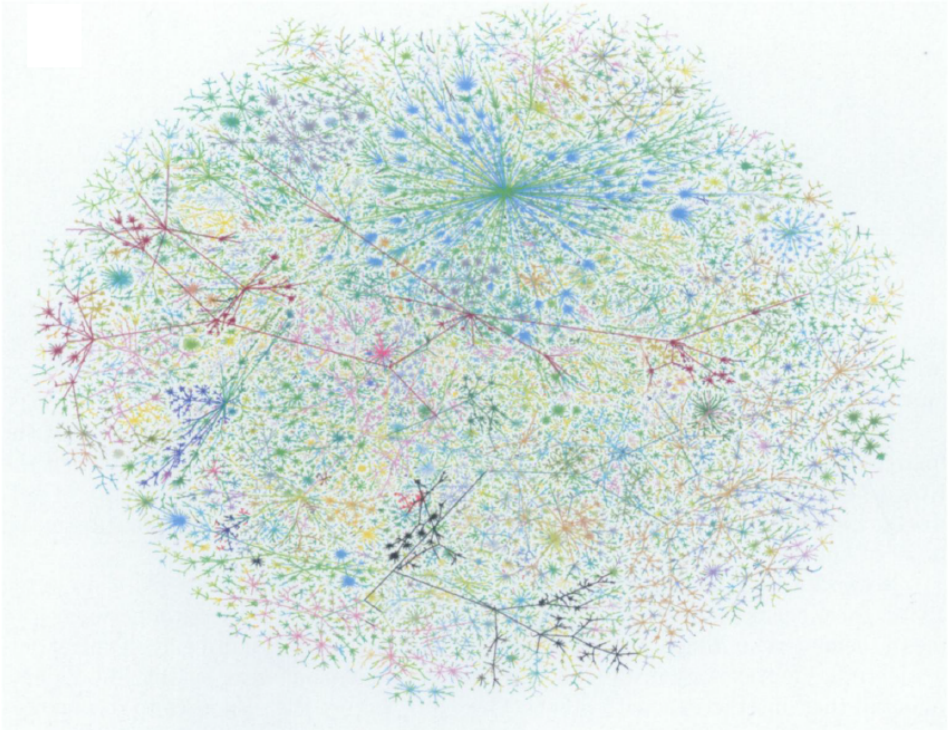

Figura 4 – Visualização da estrutura da rede da internet. Cada nó é um "sistema autônomo"(grupos locais de computadores, cada um representando centenas ou milhares de máquinas.

[Fonte: (NEWMAN, 2003)]

(centrais) para a rede. O grau de um vértice i, denotado por k_i é o número de arestas associadas a esse vértice. Para grafos não direcionados, seu valor pode ser obtido por

$$k_i = \sum_j a_{ij}. \tag{2.2}$$

O grau médio de uma rede é a média dos valores de k_i para todos os seus nós

$$\langle k \rangle = \frac{1}{N} \sum_i k_i. \tag{2.3}$$

Para redes representadas por grafos direcionados, há dois tipos de grau: o grau de saída, que representa a quantidade de arestas que partem de um vértice, e o grau de entrada, que corresponde ao número de arestas que tem o vértice em questão como destino. O grau total é definido como a soma dos dois anteriores

$$k_i^{saída} = \sum_j a_{ij}, \tag{2.4}$$

$$k_i^{entrada} = \sum_j a_{ji}, \tag{2.5}$$

$$k_i = k_i^{saída} + k_i^{entrada}. \tag{2.6}$$

As médias de graus de entrada e saídas são as mesmas para redes conectadas.

$$\langle k^{saída} \rangle = \langle k^{entrada} \rangle = \frac{1}{N} \sum_i k_i. \tag{2.7}$$

As definições acima mencionadas podem ser extendidas para grafos ponderados, mas frequentemente uma outra medida, chamada *strength* (força), é empregada. Ela é definida pelas seguintes expressões:

$$s_i^{saída} = \sum_j w_{ij}, \tag{2.8}$$

$$s_i^{entrada} = \sum_j w_{ji}. \tag{2.9}$$

Essa métrica pode ser empregada em redes de citações, para verificar o número de citações recebidas por um artigo científico, ou em redes sociais, onde representa o grau de influência do indivíduo (AMANCIO, 2013).

2.2.1.2 Caminho mínimo médio

Sendo dist(i, j) o comprimento do menor caminho que liga o vértice v_i ao vértice v_j, o comprimento médio dos caminhos iniciados por v_i pode ser expressado como

$$L_i = \frac{1}{M-1} \sum_{j=1}^{M} \text{dist}(i, j). \tag{2.10}$$

Deve-se atentar para que dist$(i, i) = 0$, logo o denominador não deve ser o número total de vértices presentes na rede, mas sim esse número subtraído em uma unidade. Outra definição importante é o diâmetro da rede

$$d = \max \text{dist}(i, j). \tag{2.11}$$

De acordo com (AMANCIO, 2013), mesmo o menor caminho médio não sendo correlacionado com a frequência com que uma palavra aparece no texto, vocábulos com altos valores de L aparecem pouco frequentemente.

Os grafos tratados neste trabalho são direcionados e ponderados, logo a distância entre dois vértices é a soma dos pesos das arestas que compõem o caminho entre eles.

Tradicionalmente, para grafos sem arestas múltiplas, o peso de uma aresta representa o número de vezes seus vértices aparecem conectados. No presente trabalho, contudo, os pesos das arestas foram invertidos (Figura 5):

$$w_{ij}^{novo} = \frac{1}{w_{ij}^{antigo}}. \tag{2.12}$$

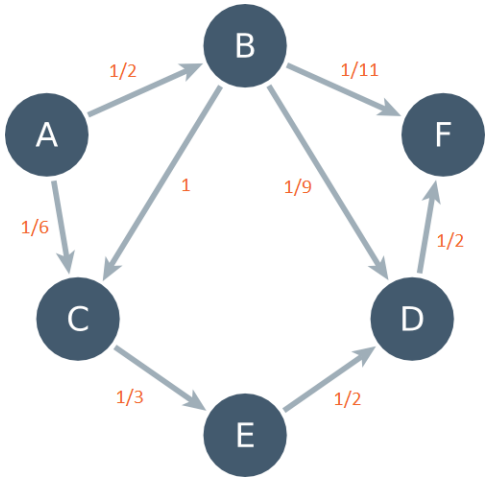

Figura 5 – Grafo da Figura 2 com pesos redefinidos pela Equação 2.12.

[Fonte: modificado de
http://www.hipster4j.org/assets/css/custom/img/graph-example-hipster-blue.png]

Assim foi decido, pois no modelo construído para este trabalho, dois nós que frequentemente aparecem conectados devem ser considerados próximos.

2.2.1.3 Coeficiente de Aglomeração (*Clustering Coefficient*)

O coeficiente de clusterização (*clustering coefficient*) C quantifica o quão próximo de um clique (grafo totalmente conectado) é o subgrafo composto pelos vizinhos de um nó. Esta é uma medida frequentemente utilizada no estudo de redes sociais e, nesse cenário, pode ter seu significado exemplificado pela frase: "quantos amigos meus são amigos entre si?". Vértices que apresentam máximos valores de C satisfazem a transitividade de vizinhos, ou seja, se dois vértices v_i e v_j são vizinhos de v_k, estão eles também estão conectados entre si. Percebe-se que esta métrica não pode ser definida para vértices que não possuem pelo menos dois vizinhos. Então, se Ψ_i é a quantidade de arestas entre os vizinhos de v_i, esse vértice tem seu coeficiente de aglomeração C_i definido por

$$C_i = \begin{cases} 2\Psi_i/(k_i^2 - k_i) & \text{para } k_i > 1, \\ 0 & \text{para } k_i \leq 1. \end{cases} \tag{2.13}$$

De acordo com (AMANCIO et al., 2011), palavras com altos valores de coeficiente de aglomeração tem maior probabilidade de aparecer em contextos mais restritos. Ou seja, valores relativamente baixos de C caracterizam palavras que aparecem em uma grande gama de contextos, explicando assim porque seus vizinhos são relativamente pouco conectados. Deve-se atentar para que C é uma medida de centralidade local, isto é, para o seu cálculo são necessários apenas os nós vizinhos àquele que se tem interesse.

2.2.1.4 *Betweenness*

O *betweenness* (B) é uma medida de de centralidade que foi proposta (FREEMAN, 1977). Esta métrica considera que um vértice é importante se esse é acessado por um grande número de caminhos mínimos. Vértices com altos valores de *betweenness* têm grande influência na rede, pois distribuem a informação pelo grafo. Por isso, em redes de sistemas de comunicação, esses vértices são os que cuja remoção mais impacta na comunicação entre outros vértices. (NEWMAN, 2010). Caso haja mais de um caminho entre dois vértices, o *betweenness* será divido igualmente para todos. Desta forma, se existem n_L caminhos mínimos entre dois nós, cada um desses terá seu *betweenness* ponderado por n_L^{-1}. Assim como feito em (AMANCIO, 2013), neste trabalho o *betweenness* será calculado de modo a evitar correlações com outras métricas. Sendo η_{sit} o número de caminhos mínimos de v_s a v_t que passam por v_i, e η_{st} o número de caminhos mínimos que partem de v_s para v_t, o *betweenness* pode ser definido como

$$B_i = \frac{1}{M^2} \sum_{s=1}^{M} \sum_{t=1}^{M} \frac{\eta_{sit}}{\eta_{st}}. \tag{2.14}$$

Majoritariamente, no contexto de análise textual, as palavras que apresentam altos valores de *betweenness* são aquelas com alta frequência, e também algumas que conectam comunidades de conceitos relacionados.

(AMANCIO et al., 2011) sugere que vocábulos com altos valores de B ligam conceitos de comunidades semânticas distintas porque têm alta probabilidade de aparecerem em vários contextos. Similarmente como faz o coeficiente de agloremeração C, o *betweenness* também representa a variedade de contextos que uma palavra pode aparecer, embora esse se baseie em padrões de conectividade globais, enquanto o primeiro, em padrões locais.

Os valores de *betweenness* calculados nos grafos gerados nesta monografia diferem daqueles gerados por (AMANCIO, 2013) pois, como dito na seção 2.2.1.2, os valores dos pesos das arestas foram redefinidos.

2.2.1.5 Assortatividade de Grau

Muito frequentemente é desejável saber se as ligações de uma rede são estabelecidas preferencialmente entre vértices pertencentes a uma mesma classe ou entre aqueles de classes distintas. Pensando nisso, a assortatividade foi proposta em (NEWMAN, 2002). Nete trabalho, foi utilizada a assortatividade de grau, ou seja, aquela que tem objetivo determinar se as ligações de um dado vértice são correlacionadas com o seu grau e de seus vizinhos. Um dos modos de definir matematimente esta correlação é por meio de probabilidades condicionais. Sendo $P(k, k')$ a probabilidade de uma aresta conectar vértices com graus k e k', a probilidade condicional de que um vértice com grau k esteja conectado

com um de grau k, é dada pela expressão:

$$P(k^{'}|k) = \frac{\langle k \rangle P(k, k^{'})}{kP(k)}. \tag{2.15}$$

Contudo, a equação mais utilizada é aquela proposta no artigo que definiu a assortatividade:

$$r = \frac{M^{-1}\sum_{j>1} k_i k_j a_{ij} - [M^{-1}\sum_{j>i} \frac{1}{2}(k_i + k_j)a_{ij}]^2}{M^{-1}\sum_{j>i} \frac{1}{2}(k_i^2 + k_j^2)a_{ij} - [M^{-1}\sum_{j>i} \frac{1}{2}(k_i + k_j)a_{ij}]^2}. \tag{2.16}$$

A rede é chamada não assortativa quando $r < 0$, isto é, vértices de graus baixos tendem a se conectar a outros nós de graus baixos. Quando $r > 0$, nós com elevados graus têm maiores chances de se conectarem a outros nós com altos valores de grau. Neste último caso, a rede é dita assortativa.

2.2.1.6 *PageRank*

O *Pagerank* também é uma medida de centralidade e foi proposta em (PAGE et al., 1998). Ela é usada como uma das principais técnicas usadas no Google para mensuração da relevância de páginas *web* em buscas feitas por usuários cotidianamente. Está técnica pode ser vista como um avanço à centralidade de Katz (KATZ, 1953) (NEWMAN, 2010). A centralidade de Katz, como outras medidas dessa natureza, mede a influência de um nó na rede. Nessa métrica, o seu valor numérico não é de todo relevante, entretanto a informação útil está na identificação dos vértices que apresentam de centralidade altos ou baixos. Vértices com alta centralidade de Katz são acessíveis por vários outros nós. Todavia, um dos pontos negativos da centralidade de Katz é que se um nó com com alta centralidade têm um grande número de vizinhos, esses nós também apresentarão alta centralidade. O *Pagerank* pondera essa redistribição de relevância na rede e matematicamente pode ser definido como

$$\mathbf{x} = \mathbf{D}(\mathbf{D} - \alpha\mathbf{A})^{-1}\mathbf{1}, \tag{2.17}$$

em que $\mathbf{1}$ é o vetor $(1, 1, 1, \dots)$, $\mathbf{D}$ é a matriz diagonal com elementos $D_{ij} = \max(k_i^{out}, 1)$, $\mathbf{x}$ é o vetor de centralidades, $\mathbf{A}$ é a matriz de adjacência do grafo e α é um fator de ponderação definido empiricamente. Com a reponderação de importância, o *Pagerank* permite que apenas os reais nós centrais tenham valores elevados de centralidade.

2.3 *Part-of-Speech tagging*

Part-of-Speech tagging é a tarefa de rotular palavras em um texto de acordo com sua classe gramatical (substantivo, adjetivo, verbo, artigo, etc). Como muitas palavras, nas mais diversas línguas, são passíveis de serem classificadas em mais de uma categoria, o processo de rotulação torna-se ambíguo. Nesta monografia, foi utilizado o *part-of-Speech tagger*

desenvolvido pela Universidade de Stanford (MANNING et al., 2014) (TOUTANOVA; MANNING, 2000)(TOUTANOVA et al., 2003).

2.4 Tipos de Redes Textuais

A seguir, os três tipos de redes mais utilizados para modelar textos são expostas.

2.4.1 Redes de coocorrência

A forma mais natural de associar palavras é pela conexão de palavras vizinhas, pois a linguagem escrita é composta por cadeiras lineares de palavras. Isso faz deste tipo de rede textual um dos mais empregados na literatura (CANCHO; SOLé, 2001). Os grafos gerados podem ser direcionados ou não. A ordem das palavras pode refletir parcialmente nas relações sintáticas e semânticas entre elas (SOLé et al., 2010). Se houver *stopwords* (palavras de baixo valor semântico, contudo de grande importância gramatical) no texto, essas palavras serão *hubs* na rede. Outras duas importantes características das redes de coocorrência de palavras é que elas são livres de escala e apresentam o comportamento *small-world* (AMANCIO, 2013), ou seja, poucos vértices são vizinhos uns dos outros, mas ainda assim a distância média entre eles é relativamente curta ($L \propto \ln N$) (NEWMAN, 2010).

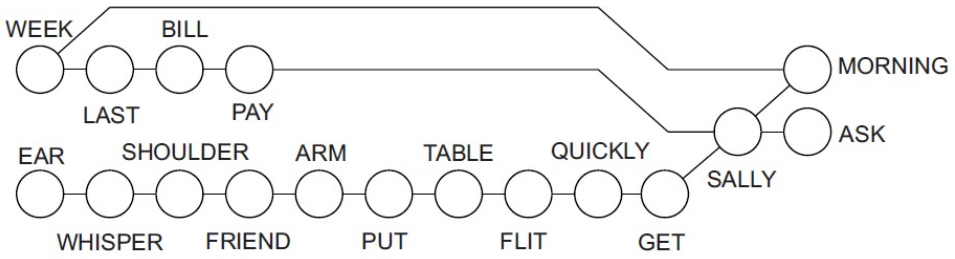

Figura 6 – Grafo não orientado obtido a partir da sentença "What's that? Asked Sally. Pay my bill for last week, due this morning. Sally got up quickly, and flitting down the table, put her arm round her friend's shoulder and whispered in her ear."do livro "The Adventures of Sally"de P.G. Wodehouse.

[Fonte: (AMANCIO, 2013)]

2.4.2 Redes Sintáticas

As redes sintáticas associam as palavras que apresentam dependência sintática entre si. Essas dependências podem ser trazidas à tona por meio das gramáticas de dependências. Esse formalismo é capaz de definir a estrutura sintática de uma sentença como uma árvore.

Os *hubs* neste tipo de rede são palavras funcionais, mas seus graus de entrada e saída são diferentes da rede anterior (SOLé et al., 2010). Entretanto, 90% das conexões das redes sintáticas também ocorrem nas redes de coocorrência (AMANCIO, 2013). Pode-se interpretar então que as redes de coocorrência são aproximações das redes sintáticas. Um outro fato que corrobora esta ideia é que a rede em questão também apresenta, como a anterior, propriedades livre de escala e *small-world*.

2.4.3 Redes Semânticas

Redes semânticas podem ser construídas a partir de palavras que representam conceitos e associando-as se se possuírem alguma relação semântica básica como, por exemplo, "é-um", "parte-todo"ou "oposição binária"(SIGMAN; CECCHI, 2002)(SOLé et al., 2010). Também foi verificado que essas redes apresentam uma organização altamente eficiente e que os *hubs* representam palavras polissêmicas (SOLé et al., 2010). Comumente, as redes semânticas apresentam um alto coeficiente de clusterização. Essa característica permite que buscas por associação sejam feitas rapidamente (MOTTER et al., 2003).

2.5 Classificador

2.5.1 Perceptron de Múltiplas Camadas

As redes neurais Perceptron de Múltiplas Camadas são aproximadores de funções conexionistas. São formadas por neurônios (unidades de processamento inspirados nas células homônimas) interconectados e divididos em camadas. Para cada um desses componentes, há uma função de ativação (em muitos casos, não linear) que serão responsáveis pela resposta do neurônio para dados estímulos (entradas). MLPs apresentam um bom poder de generalização e seu conhecimento é armazenado como pesos de cada ligação entre neurônios de camadas distintas (VALENçA, 2007).

Seu modelo de aprendizado é supervisionado. Isso significa que durante a fase de treinamento, devem ser apresentadas à rede as entradas e suas respectivas saídas. Para cada exemplo apresentado, os parâmetros internos (pesos de cada conexão) são ajustados em função do erro encontrado e do estado atual desses mesmos parâmetros. Após o treinamento, os sinais são propagados apenas no sentido das entradas para as saídas. Por este motivo, as MLPs são classificadas como redes *feedforward*.

A arquitetura das MLPs (figura 7) deve apresentar os seguintes constituintes:

1. Uma camada de entrada: nesta camada estão presentes os neurônios que sofreram os estímulos externos (entradas);

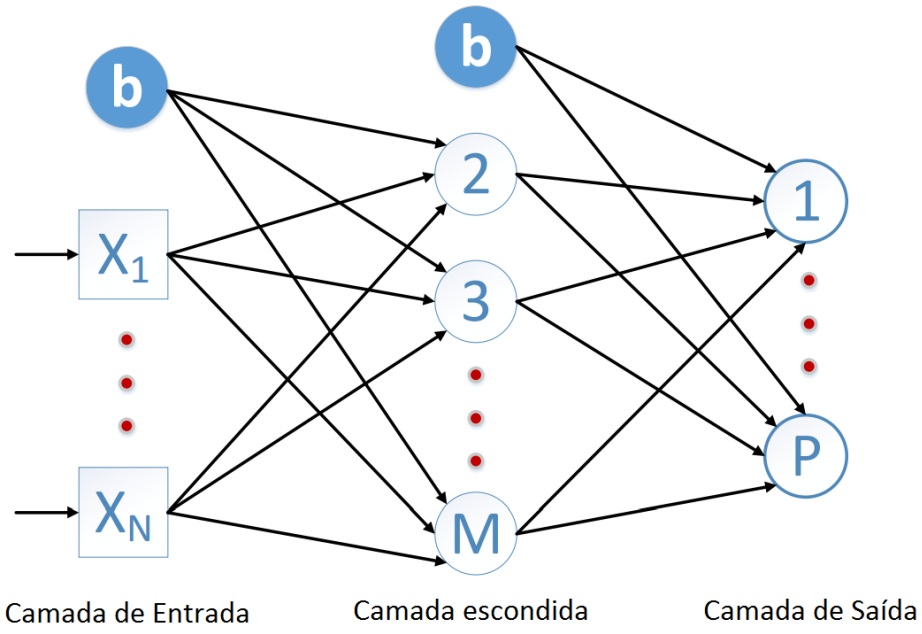

Figura 7 – MLP com 1 camada escondida.

[Fonte: modificada de (FARIAS, 2014)]

2. n camadas escondidas: está é a origem da não linearidade das MLPs. Tipicamente possuem funções de ativação não lineares, sendo exemplos clássicos a sigmóide logística e a tangente hiperbólica. Em (HAYKIN, 2007) é mostrado que, com uma camada escondida, a rede é capaz de aproximar qualquer função contínua, e, que com duas, qualquer função. Entretanto o número de camadas não deve ser arbritariamente grande, pois para muitos algoritmos de treinamento, as taxas de erro podem aumentar (PRECHELT, 1994);

3. Uma camada de saída: representam a saída da rede. Geralmente são utilizadas funções *softmax* (função que generaliza as sigmóides) para problemas de classificação, mas também podem ser simples funções lineares.

Existem vários algoritmos de treinamento para MLPs. Um dos mais conhecidos e utilizados é o *backpropagation*. Nessa técnica de treinamento, geralmente é utilizado o método de otimização gradiente descendente. De acordo com (VALENçA, 2007), o *backpropagation* é constituído por duas fases:

1. Os sinais de entrada são propagados em direção à saída da rede. Ao alcançar a última camada, a saída é calculada e o erro a ela atribuído;

2. O erro calculado na etapa anterior é propagado de volta para a entrada. Nesta travessia, é calculada a contribuição de cada neurônio para o erro, e seu peso é

ajustado em função disso. A equação de atualização dos pesos também de acordo com (HAYKIN, 2007) é:

$$w_{ij}^{m}(t+1) = w_{ij}^{m}(t) + \alpha\delta_{i}^{m} f^{m-1}(net_{j}^{m-1}) + \beta\Delta w_{ij}^{m}(t-1), \tag{2.18}$$

em que α é a taxa de aprendizagem, β é a taxa de momento, δ é a sensibilidade que pode ser calculada de acordo com as Equações (2.19) e (2.20) para a camadas de entrada e as demais respectivamente:

$$\delta_{i}^{m} = (d_i - yi) f^{'}(net_i), \tag{2.19}$$

$$\delta_{i}^{m-1} = f^{'(m-1)}(net_{j}^{m-1}) \sum_{i=l}^{N} w_{ij}^{m}\delta_{i}^{m}. \tag{2.20}$$

3 Modelo de Pesquisa

Diferentemente do método mais usual empregado na literatura, que modela o texto como uma rede de coocorrência de palavras (AMANCIO, 2013), a metodologia aqui empregada constrói uma rede de coocorrência de pares de funções gramaticais. Até o presente conhecimento, esta alternativa é inédita. Ela foi pensada com o propósito de capturar ainda mais relações sintáticas e, até certo ponto, também semânticas dos textos em questão.

Tendo em vista que foram utilizadas 37 *tags* para classificar os vocábulos, existem $\frac{37!}{(37-2)!} = 1332$ vértices possíveis na rede e $1332(1332-1) = 1772892$ possíveis arestas. Entretanto, apenas uma pequena fração desses vértices aparecem em cada grafo. As redes obtidas são esparsas e, neste aspecto, condizem com as características esperadas para redes reais.

Outra importante característica dos grafos gerados está nos pesos de suas arestas. Comumente o peso w_{ij} de uma aresta entre os vértices v_i e v_j representa o número de vezes em que há uma ligação unitária partindo do vértice v_i para o vértice v_j. Os pesos são de fundamental importância na rede e em sua caracterização, pois influenciam os valores de menores caminhos entre nós. Para evitar que nós frequentemente conectados sejam considerados distantes (altos valores dos pesos das arestas que os conectam), os pesos foram redefinidos. Após o cálculo de todos os pesos, esses terão seus valores invertidos, como definido na Equação 2.12.

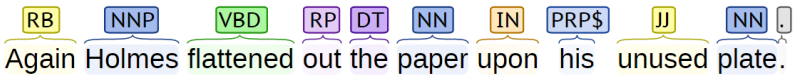

Figura 8 – Frase do livro *The Valley of Fear*, de Arthur Conan Doyle, classificado em *part-of-speech tags*

[Fonte: reproduzido de http://nlp.stanford.edu:8080/corenlp/process]

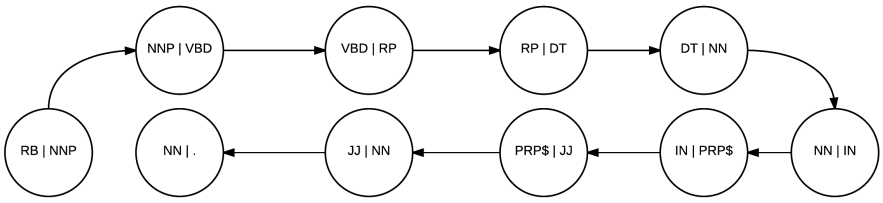

Figura 9 – Grafo da frase da imagem 8 gerado com o novo modelo proposto

4 Método de Pesquisa

Este capítulo elucidará os passos (Figura 10), processos e métodos utilizados para classificar textos, por autoria, usando redes complexas. Esta tarefa já foi realizada antes (AMANCIO et al., 2011)(AMANCIO, 2013), porém o presente trabalho modifica abordagem no tocante à criação da rede a partir dos textos. Nos trabalhos supracitados, os textos são modelados como redes de coocorrência, enquanto neste, pares de classificações gramaticais (*part-of-speech tagging*) são conectados ao primeiro par consecutivo como foi descrito no Capítulo 3.

Tabela 1 – Lista de livros utilizados no reconhecimento de autoria.

[Fonte: elaboração própria]

Título	**Autor**
The Tragedy of the Korosko	Arthur Conan Doyle
The Valley of Fear	Arthur Conan Doyle
The War in South Africa: Its Cause and Conduct	Arthur Conan Doyle
Through the Magic Door	Arthur Conan Doyle
Uncle Bernac - A Memory of the Empire	Arthur Conan Doyle
Dracula's Guest	Bram Stoker
The Jewel of Seven Stars	Bram Stoker
The Lady of the Shroud	Bram Stoker
The Lair of the White Worm	Bram Stoker
The Man	Bram Stoker
A Tale of Two Cities	Charles Dickens
American Notes	Charles Dickens
Barnaby Rudge: A Tale of the Riots of Eighty	Charles Dickens
Great Expectations	Charles Dickens
Hard Times	Charles Dickens
The Works of Allan Poe (5 volumes)	Edgar Allan Poe
Beasts and Super Beasts	Hector Hugh Munro
The Chronicles of Clovis	Hector Hugh Munro
The Toys of Peace and Other Papers	Hector Hugh Munro
The Unbearable Bassington	Hector Hugh Munro
When William Came	Hector Hugh Munro
A Connecticut Yankee in King Arthur's Court	Mark Twain
Adventures of Huckberry Finn	Mark Twain
The Adventures of Tom Sawyer	Mark Twain
The Mysterious Stranger	Mark Twain
The Prince and the Pauper	Mark Twain
My Man Jeeves	Pelham Grenville Wodehouse
Tales of St. Austin's	Pelham Grenville Wodehouse
The Adventures of Sally	Pelham Grenville Wodehouse
The Clicking of Cuthbert	Pelham Grenville Wodehouse
The Man with Two Left Feet	Pelham Grenville Wodehouse
A Changed Man and Other Tales	Thomas Hardy
A Pair of Blue Eyes	Thomas Hardy
Far from the Madding Crowd	Thomas Hardy
Jude the Obscure	Thomas Hardy
The Hand of Ethelberta	Thomas Hardy

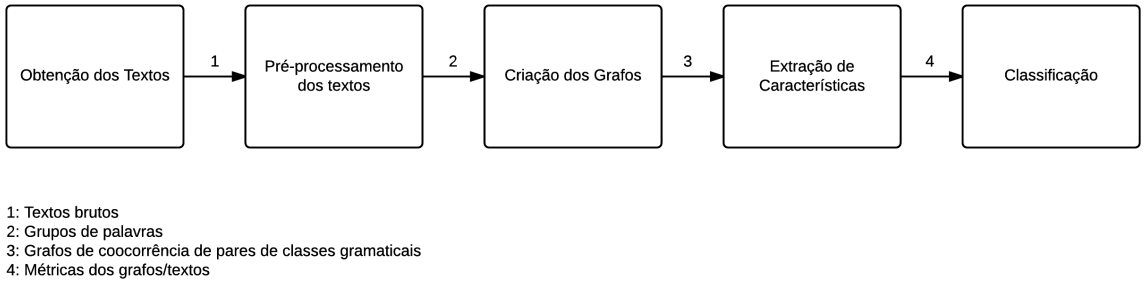

Figura 10 – Fluxograma de classificação de autoria de textos.

4.1 Criação da base de textos

Para montagem da base de textos utilizada nesta monografia, foram utilizados 40 textos. Esses são, em sua maioria, os mesmos utilizados em (AMANCIO, 2013). São todos de autores da língua inglesa que viveram em meados do século XIX. Além da motivação de usar textos já submetidos ao reconhecimento de autoria, também se pode adicionar uma outra: dado que os autores compartilham o mesmo idioma e época, é de se esperar que seus estilos sejam relativamente parecidos, tornando assim a tarefa de reconhecimento de autoria mais complexa. A lista completa de títulos utilizados está presente na Tabela 1. No total, são 8 autores, 5 livros cada, totalizando 40 livros.

Como a quantidade de textos é insuficiente para aplicação em um classificador, e seus tamanhos são grandes e bastante desiguais, como demonstrado pela Tabela 2, eles foram divididos em grupos de 5000 palavras. A princípio, outros tamanhos foram testados, contudo esse foi o que apresentou melhores resultados. Trechos muito curtos foram evitados, pois não conseguiriam criar um grafo de expressividade significativa. Por outro lado, segmentos demasiadamente longos também foram rejeitados, pois diminuiam substancialmente o número de exemplos capazes de serem expostos ao classificador. Durante a divisão dos fragmentos, houve o cuidado de que se evitasse o rompimento de parágrafos. Isto porque, pela própria natureza e definição desse componente textual, das sentenças neles contidas é esperada uma grande similaridade semântica. Por isso, os blocos resultantes de palavras têm aproximadamente 5000 palavras.

No decorrer deste escrito, quando conveniente, as menções aos autores serão feitas por uso das abreviaturas de seus nomes como mostrado na Tabela 3. Todos os textos foram obtidos gratuitamente a partir do portal www.gutemberg.org. Fora feito um único pré-processamento antes dos textos serem divididos: a remoção dos cabeçalhos e rodapés inseridos pelo *site* supracitado.

Tabela 2 – Número, por autor, de blocos textuais com 5000 palavras.

[Fonte: elaboração própria]

Autor	**Número de blocos textuais**
Arthur Conan Doyle	51
Bram Stoker	88
Charles Dickens	103
Edgar Allan Poe	85
Hector Hugh Munro	51
Mark Twain	78
Pelham Grenville Wodehouse	59
Thomas Hardy	119

Tabela 3 – Abreviatura dos nomes dos autores.

[Fonte: elaboração própria]

Autor	**Abreviatura**
Arthur Conan Doyle	ACD
Bram Stoker	BS
Charles Dickens	CD
Edgar Allan Poe	EAP
Hector Hugh Munro	HHM
Mark Twain	MT
Pelham Grenville Wodehouse	PGW
Thomas Hardy	TH

4.2 Pré-processamento

Para cada bloco de texto resultante da etapa anterior, foram executados os seguintes passos:

1. remoção dos indicadores de capítulos, dedicatórias, notas e quaisquer trechos que não compusessem o corpo efetivo da obra em questão. Esta decisão foi tomada pois o principal objetivo é a caracterização da autoria na obra. Fragmentos avulsos, ou até mesmo de autoria diferente, poderiam representar ruídos no processo de classificação;

2. segmentação de linhas: a fragmentação do texto em frases viabiliza algumas das operações citadas a baixo. Além disso, inicialmente teve-se a ideia de modelar os vértices do grafo como sendo as frases do texto (compostas não pelas palavras propriamente ditas, mas por suas classes gramaticais (*part-of-speech tagging*), todavia essa abordagem não obteve sucesso devido a pouca repetição da estrutura completa das frases;

3. remoção de *stopwords* (quando aplicável): *stopwords* são palavras que carregam pouco valor semântico. Geralmente são artigos, alguns pronomes, e palavras do gênero. Embora uma grande quantidade de trabalhos opte por removê-las, em uma das abordagens executadas neste trabalho, elas não foram removidas;

4. *part-of-speech tagging*: é o processo onde a cada palavra do texto será associada a sua respectiva classe gramatical *part-of-speech tagging*. Foi utilizado o *Stanford part-of-speech tagger* (MANNING et al., 2014)(TOUTANOVA; MANNING, 2000)(TOUTANOVA et al., 2003). Esta etapa tem uma relevância especial para a presente monografia, pois a abordagem nela utilizada cria grafos cujos nós são pares de *tags* encontradas pelo classificador acima. Todas as *tags* aplicáveis pelo *tagger* estão disponíveis na Tabela 7 do Apêndice A. Ao todo, as palavras podem ser classificadas em 45 categorias, mas os 9 primeiros itens da tabela referenciada acima (exceto os delimitadores de sentenças) não foram utilizados, restando assim 37 *tags*. A Figura 11 mostra um exemplo da classificação de palavras em *tags*.

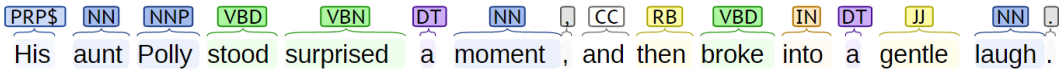

Figura 11 – Frase do livro *The Adventures of Tom Sawyer*, de Mark Twain, classificado em *part-of-speech tags*

[Fonte: reproduzido de http://nlp.stanford.edu:8080/corenlp/process]

4.3 Criação dos Grafos

Cada um dos blocos textuais dá origem a um grafo como descrito no capítulo 3.

4.4 Caracterização dos Grafos

A grande maioria das métricas aplicadas nos grafos gerados neste trabalho são de natureza individual para cada vértice do grafo, ou seja, mesmo que algumas utilizem conceitos globais, essas métricas são individuais para cada nó. Entretanto, são necessárias medidas que caracterizem o grafo como todo. Existe um grande número de maneiras de se extrair características numéricas de grafos a partir das métricas se seus vértices. Em (AMANCIO, 2013), por exemplo, as medidas dos vértices são ponderadas de acordo com sua distribuição de probabilidade. Para o escopo deste trabalho, o método empregado será o mais direto e convencional: médias aritméticas e desvios padrão. Desta forma, cada uma das medidas: coeficiente de aglomeração C, *betweenness* B, *Pagerank* e caminho mínimo médio L apresentará dois valores para cada grafo.

Além das métricas de redes complexas citadas acima, foram adicionadas algumas características a elas:

1. média e desvio padrão dos comprimentos das sentenças;
2. média e desvio padrão das frequências de cada *tag*.

4.5 Classificação

Para classificar os textos, foi utilizada uma rede neural MLP treinada com o algoritmo *backpropagation*. Esta técnica foi escolhida pois também fora empregada na literatura (AMANCIO, 2013) e mostrou bons resultados. Todas as configurações de MLPs para o problema tratado neste trabalho possuem 13 neurônios de entrada (um para cada característica citada acima) e 8 neurônios de saída, sendo cada um ativado para um autor diferente.

Como o número de blocos de textos é consideravelmente diferente para cada autor, conforme observado na Tabela 2, n blocos são selecionados pseudoaleatoriamente para cada autor, onde n é a menor quantidade total de blocos possuídos por um deles. Essa seleção é importantíssima pois evita um grande desbalanceamento das classes, desbalanceamento este que poderia comprometer ou, até mesmo, mascarar os resultados obtidos pelo classificador.

4.5.1 Configuração do Classificador

Nos experimentos realizados, o MLP teve as seguintes configurações:

1. função de ativação: sigmóide logística para todos os nós;
2. algoritmo de treinamento: *backpropagation*
3. critério de parada: 20 épocas consecutivas sem melhora ou 5000 iterações
4. momentum: 0.3
5. tamanho do lote: 100
6. arquitetura:

   1 ou 2 camadas escondidas

   5, 10 ou 20 nós em cada camada escondida

Para evitar uma explosão combinatória de experimentos, deve-se perceber que alguns parâmetros da rede MLP foram fixados. Apenas aqueles que demonstraram maior impacto no desempenho do classificador são variáveis.

As configurações acima foram aplicadas sobre grafos gerados a partir de textos com e sem *stopwords*.

4.5.2 Avaliação da Classificação

Para validação das classificações obtidas foi utilizado o *k-fold*. Esse é um método de validação cruzada no qual os dados são divididos em k subconjuntos (*folds*) disjuntos

de cardinalidade aproximadamente iguais. $k-1$ conjuntos são utilizados para treinamento, enquanto o $fold$ restante é usado para teste. Este processo é realizado k vezes até que todos os conjuntos tenham sido utilizados para teste uma vez. O resultado é a média e o desvio padrão obtidos para todos os experimentos.

Para cada conjunto de blocos selecionados, são efetuadas 30 simulações utilizando o *k-fold* com $k = 10$.

5 Resultados

Na Tabela 4 são mostrados a média e desvio padrão (entre parênteses) dos resultados de acurácia dos seis experimentos descritos na Seção 4.5.1. As configurações da rede podem ser observadas na Tabela 5. Todos os resultados são expostos mais claramente na Figura 12, onde estão representados os *boxplots* das taxas de classificação em função da configuração da MLP. Pode-se perceber que os melhores resultados foram obtidos para uma única camada escondida e que a taxa de classificação obtida foi superior a 50%. Deve-se perceber que se o processo de escolha fosse aleatório, o valor esperado da taxa de classificação seria de 12,5%.

Tabela 4 – Resultados MLP

Configuração	Grafo com *Stopwords*	Grafo sem *Stopwords*
1	52,70(10,03)	47,96(9,86)
2	51,56(9,13)	47,61(7,85)
3	49,03(7,67)	46,55(9,10)
4	40,77(12,08)	34,87(11,56)
5	38,36(13,61)	34,05(14,79)
6	36,87(13,68)	29,57(12,16)

Tabela 5 – Configurações da MLP

Configuração	Número de Camadas	Número de Neurônios
1	1	[10]
2	1	[20]
3	1	[5]
4	2	[10, 10]
5	2	[20, 20]
6	2	[5, 5]

Os resultados sugerem que a remoção das *stopwords* compromete a capacidade da rede de representar características estilísticas. A matriz de confusão mostrada na Tabela 6 pode indicar que naturalmente existem autores cujos traços de estilo não são bem capturados pela rede. É o caso de Arthur Conan Doyle e Thomas Hardy, que foram muito confundidos com Hector Hugh Munro e Mark Twain, respectivamente.

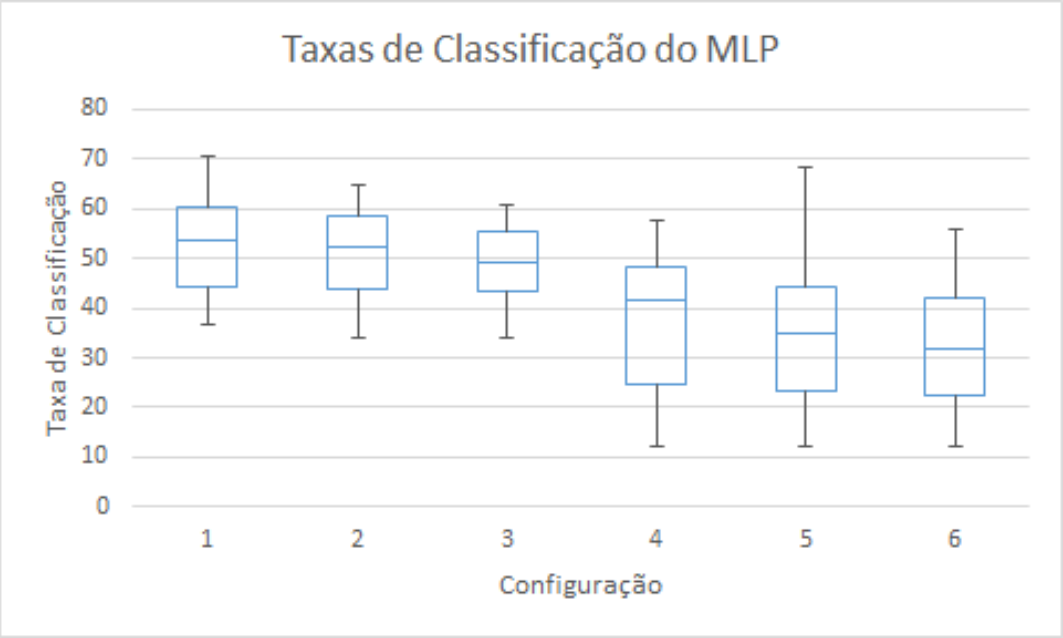

Figura 12 – *Boxplots* das configurações da MLP

Tabela 6 – Matriz de confusão para uma execução da MLP em configuração 1 com escore 52,94%

	HHM	ACD	EAP	BS	CD	TH	MT	PGW
HHM	36	4	3	5	2	0	1	0
ACD	16	9	5	7	3	2	1	8
EAP	4	3	38	2	2	0	1	1
BS	4	6	0	35	3	2	0	1
CD	2	0	7	2	21	3	16	0
TH	2	4	4	2	7	16	5	11
MT	1	2	1	1	15	3	24	4
PGW	0	2	0	2	0	6	4	37

6 Conclusões e Considerações Finais

6.1 Conclusões

O estudo de redes complexas é um campo muito ativo e desafiador na ciência. Elas apresentam ótima capacidade de representar sistemas dinâmicos e têm sido empregadas interdisciplinarmente nas mais diversas áreas do conhecimento.

Mesmo com promissoras aplicações de redes complexas no processamento de textos e de linguagem natural, muito ainda pode ser avançado. Este trabalho propôs uma nova metodologia de composição de grafos a partir de textos a partir das classes gramaticais. No dito modelo, os nós são pares de *part-of-speech tags* e as arestas representam que sua coocorrência. Além disso, foi feito uso de palavras comumente desconsideradas na construção de redes de coocorrência: as *stopwords*, e os resultados dos classificadores apontaram que esses vocábulos são importantes para o tipo de rede proposta.

Os resultados sugerem que as redes mais simples (menor número de camadas) têm uma melhor capacidade de generalização do problema independentemente da presença de *stopwords*. Eles ainda indicam que redes de classes gramaticais tem uma um bom potencial para capturar traços estilísticos em textos.

6.2 Trabalhos Futuros

Entre os trabalhos futuros, estão:

1. Construir novos modelos de construção de grafos de textos
2. Usar novas métricas mais específicas para grafos ponderados e direcionados como a Reciprocidade e as novas definições de coeficiente de aglomeração;
3. Utilizar algumas métricas de séries temporais como feito em (AMANCIO, 2013);
4. Modificar a ponderação de métricas na caracterização do grafo como um todo;
5. Aumentar o número de parâmetros variáveis na MLP;
6. Avaliar outros classificadores com novas configurações, como o SVM e Random Forests;
7. Modelar textos de autores de épocas diferentes;
8. Modelar textos de outros idiomas.

Referências

AMANCIO, D. R. *Classificação de textos com redes complexas*. Tese (Doutorado) — Instituto de Física de São Carlos - USP, 2013.

AMANCIO, D. R. et al. Comparing intermittency and network measurements of words and their dependency on authorship. *New Journal of Physics*, dez. 2011. ISSN 1367-2630.

AMANCIO, D. R.; OLIVEIRA, O. N.; COSTA, L. da F. Identification of literary movements using complex networks to represent texts. *New Journal of Physics*, v. 14, n. 4, p. 043029+, abr. 2012. ISSN 1367-2630.

ANTIQUEIRA, L. et al. Modelando textos como redes complexas. In: *Anais do XXV Congresso da Sociedade Brasileira de Computação (III Workshop em Tecnologia da Informação e da Linguagem Humana - TIL)*. São Leopoldo-RS, Brasil: [s.n.], 2005. p. 2089–2098.

BARABáSI, A.-L.; ALBERT, R. Emergence of Scaling in Random Networks. *Science*, American Association for the Advancement of Science, Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA., v. 286, n. 5439, p. 509–512, out. 1999. ISSN 1095-9203. Disponível em: <http://dx.doi.org/10.1126/science.286.5439.509>.

BULLMORE, E.; SPORNS, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, Nature Publishing Group, v. 10, n. 3, p. 186–198, mar. 2009. ISSN 1471-0048.

CANCHO, R. F. i; SOLé, R. V. The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, v. 268, p. 2261–2266, 2001.

COSTA, L. da F. et al. Characterization of complex networks: A survey of measurements. In: *ADVANCES IN PHYSICS*. [S.l.: s.n.], 2005.

ERDÖS, P.; RÉNYI, A. On the evolution of random graphs. In: *PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES*. [S.l.: s.n.], 1960. p. 17–61.

FARIAS, F. C. *Interface Cérebro-Máquina: Reconhecimento de sinais cerebrais através de resevoir computing*. Dissertação (Mestrado) — Escola Politécnica da Universidade de Pernambuco, junho 2014.

FREEMAN, L. C. A set of measures of centrality based on betweenness. *Sociometry*, American Sociological Association, v. 40, n. 1, p. 35–41, mar. 1977.

HAYKIN, S. *Neural Networks: A Comprehensive Foundation (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2007. ISBN 0131471392.

KATZ, L. A new status index derived from sociometric analysis. *Psychometrika*, v. 18, n. 1, p. 39–43, March 1953.

MANNING, C. D. et al. The Stanford CoreNLP natural language processing toolkit. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, 2014. p. 55–60. Disponível em: <http://www.aclweb.org/anthology/P/P14/P14-5010>.

MOTTER, A. E. et al. Topology of the conceptual network of language. *Physical Review E*, v. 65, 2003.

NEWMAN, M. *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010. ISBN 0199206651, 9780199206650.

NEWMAN, M. E. Assortative mixing in networks. *Phys. Rev. Lett.*, v. 89, n. 20, p. 208701, 2002.

NEWMAN, M. E. J. The structure and function of complex networks. *SIAM REVIEW*, v. 45, p. 167–256, 2003.

PAGE, L. et al. *The PageRank Citation Ranking: Bringing Order to the Web*. [S.l.], 1998.

PRECHELT, L. *PROBEN1 - a set of neural network benchmark problems and benchmarking rules*. [S.l.], 1994.

SIGMAN, M.; CECCHI, G. Global organization of the wordnet lexicon. *Proceedings of the National Academy of Science*, v. 99, p. 1742–1747, 2002.

SOLé, R. V. et al. Language networks: Their structure, function, and evolution. *Complexity*, Wiley Subscription Services, Inc., A Wiley Company, v. 15, n. 6, p. 20–26, 2010. ISSN 1099-0526.

SPORNS, O. *Networks of the Brain*. 1st. ed. [S.l.]: The MIT Press, 2010. ISBN 0262014696, 9780262014694.

TOUTANOVA, K. et al. Feature-rich part-of-speech tagging with a cyclic dependency network. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (NAACL '03), p. 173–180. Disponível em: <http://dx.doi.org/10.3115/1073445.1073478>.

TOUTANOVA, K.; MANNING, C. D. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: *In EMNLP/VLC 2000*. [S.l.: s.n.], 2000. p. 63–70.

VALENçA, M. *Fundamentos das Redes Neurais*. 1st. ed. [S.l.]: The MIT Press, 2007. ISBN 8577163423.

WANG, X. F.; CHEN, G. Complex networks: small-world, scale-free and beyond. *Circuits and Systems Magazine, IEEE*, IEEE, v. 3, n. 1, p. 6–20, 2003. ISSN 1531-636X.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of /'small-world/' networks. *Nature*, Nature Publishing Group, Department of Theoretical and Applied Mechanics, Cornell University, Ithaca, New York 14853, USA. djw24@columbia.edu, v. 393, n. 6684, p. 440–442, jun. 1998. ISSN 0028-0836.

APÊNDICE A – Tabela de *part-of-speech tags*

A tabela a seguir contém a lista de *tags* utilizadas pelo *Stanford part-of-speech tagger*. Vale ressaltar que as classes gramaticais abaixo pertencem à gramática da língua inglesa e portanto, nem sempre, possuem um equivalente no português.

Tabela 7 – Lista de *tags* utilizadas pelo *Stanford part-of-speech tagger*

Tag	**Descrição**	**Exemplos**
$	dólar	$ -$ –$ A$ C$ HK$ M$ NZ$ S$ U.S.$ US$
“	início de citação	‘ “
”	fim de citação	’ ”
(	parênteses à esquerda	([{
)	parênteses à direita	)] }
,	vírgula	,
–	trevessão	–
.	delimitador de frase	. ! ?
;	ponto e vírgula ou elipse	: ; ...
CC	conjunção coordenada	& ’n and both but either et for less minus neither nor or plus so therefore times v. versus vs. whether yet
CD	numeral cardinal	mid-1890 nine-thirty forty-two one-tenth ten million 0.5 one forty-seven 1987 twenty ’79 zero two 78-degrees eighty-four IX ’60s .025 fifteen 271,124 dozen quintillion DM2,000
DT	determinador	all an another any both del each either every half la many much nary neither no some such that the them these this those
		Continuação na próxima página

Tabela 7 – continuação da página anterior

Tag	**Descrição**	**Exemplos**
EX	*there* existencial	there
FW	palavra estrangeira	gemeinschaft hund ich jeux habeas Haementeria Herr K'ang-si vous lutihaw alai je jour objets salutaris fille quibusdam pas trop Monte terram fiche oui corporis
IN	preposição ou conjunção subordinada	astride among uppon whether out inside pro despite on by throughout below within for towards near behind atop around if like until below next into if beside
JJ	adjetivo ou numeral ordinal	third ill-mannered pre-war regrettable oiled calamitous first separable ectoplasmic battery-powered participatory fourth still-to-be-named multilingual multi-disciplinary
JJR	adjetivo comparativo	thirdbleaker braver breezier briefer brighter brisker broader bumper busier calmer cheaper choosier cleaner clearer closer colder commoner costlier cozier creamier crunchier cuter
JJS	adjetivo superlativo	calmest cheapest choicest classiest cleanest clearest closest commonest corniest costliest crassest creepiest crudest cutest darkest deadliest dearest deepest densest dinkiest
LS	marcador de itens em listas	A A. B B. C C. D E F First G H I J K One SP-44001 SP-44002 SP-44005 SP-44007 Second Third Three Two a b c d first five four one six three two

Continuação na próxima página

Tabela 7 – continuação da página anterior

Tag	**Descrição**	**Exemplos**
MD	verbo modal auxiliar	can cannot could couldn't dare may might must need ought shall should shouldn't will would
NN	nome comum, singular ou*mass* (incontável)	common-carrier cabbage knuckle-duster Casino afghan shed thermostat investment slide humour falloff slick wind hyena override subhumanity machinist blood
NNP	nome próprio, singular	Motown Venneboerger Czestochwa Ranzer Conchita Trumplane Christos Oceanside Escobar Kreisler Sawyer Cougar Yvette Ervin ODI Darryl CTCA Shannon A.K.C. Meltex Liverpool
NNPS	nome próprio, plural	Americans Americas Amharas Amityvilles Amusements Anarcho-Syndicalists Andalusians Andes Andruses Angels Animals Anthony Antilles Antiques Apache Apaches Apocrypha
NNS	nome comum, plural	undergraduates scotches bric-a-brac products bodyguards facets coasts divestitures storehouses designs clubs fragrances averages subjectivists apprehensions muses factory-jobs
PDT	pré-delimitador	all both half many quite such sure this
POS	indicador genitivo	' 's
PRP	pronome pessoal	hers herself him himself hisself it itself me myself one oneself ours ourselves ownself self she thee theirs them themselves they thou thy us
PRP$	pronome possessivo	her his mine my our ours their thy your

Continuação na próxima página

Tabela 7 – continuação da página anterior

Tag	**Descrição**	**Exemplos**
RB$	advérbio	occasionally unabatingly maddeningly adventurously professedly stirringly prominently technologically magisterially predominately swiftly fiscally pitilessly
RBR$	advérbio comparativo	further gloomier grander graver greater grimmer harder harsher healthier heavier higher however larger later leaner lengthier lessperfectly lesser lonelier longer louder lower more
RBS$	advérbio superlativo	best biggest bluntest earliest farthest first furthest hardest heartiest highest largest least less most nearest second tightest worst
RP$	partícula	aboard about across along apart around aside at away back before behind by crop down ever fast for forth from go high i.e. in into just later low more off on open out over per pie raising start teeth that through under unto up up-pp upon whole with you
SYM$	símbolo	% & ’ ” ”.)). * + ,. < = > @ A[fj] U.S U.S.S.R
TO$	"to"como preposição ou indicador de infinitivo	to
UH$	interjeição	Goodbye Goody Gosh Wow Jeepers Jee-sus Hubba Hey Kee-reist Oops amen huh howdy uh dammit whammo shucks heck anyways whodunnit honey golly man baby diddle hush sonuvabitch

Continuação na próxima página

Tabela 7 – continuação da página anterior

Tag	**Descrição**	**Exemplos**
VB$	verbo, forma básica	ask assemble assess assign assume atone attention avoid bake balkanize bank begin behold believe bend benefit bevel beware bless boil bomb boost brace break bring broil brush build
VBD$	verbo, passado	dipped pleaded swiped regummed soaked tidied convened halted registered cushioned exacted snubbed strode aimed adopted belied figgered speculated wore appreciated contemplated
VBG$	verbo, particípio presente ou gerúndio	telegraphing stirring focusing angering judging stalling lactating hankerin' alleging veering capping approaching traveling besieging encrypting interrupting erasing wincing
VBN$	verbo, particípio passado	multihulled dilapidated aerosolized chaired languished panelized used experimented flourished imitated reunifed factored condensed sheared unsettled primed dubbed desired
VBP$	verbo, presente, não 3ª pessoa do singular	predominate wrap resort sue twist spill cure lengthen brush terminate appear tend stray glisten obtain comprise detest tease attract emphasize mold postpone sever return wag
VBZ$	verbo, presente, 3ª pessoa do singular	bases reconstructs marks mixes displeases seals carps weaves snatches slumps stretches authorizes smolders pictures emerges stockpiles seduces fizzes uses bolsters slaps speaks pleads

Continuação na próxima página

Tabela 7 – continuação da página anterior

Tag	**Descrição**	**Exemplos**
WDT$	WH-determinante	that what whatever which whichever
WP$	WH-pronome	that what whatever whatsoever which who whom whosoever
WP$	WH-pronome, possessivo	whose
WRB$	WH-advérbio	how however whence whenever where whereby whereever wherein whereof why