Comparison of Stylometric Attributes for Writing Authorship Identification: A Case Study of Guimarães Rosa versus Clarice Lispector

Raido Lacorte Galina[1], **Diego do Nascimento Rodrigues Flores**[2], **Karin S. Komati**[3]

[1] Programa de Pós-graduação em Ciência de Dados com Big Data

[2]Diretoria de Ensino

[3]Programa de Pós-graduação em Computação Aplicada (PPComp)
Campus Serra do Instituto Federal do Espírito Santo (Ifes)
Serra – ES – Brasil

raido.l.g@hotmail.com, {diego.flores, kkomati}@ifes.edu.br

Abstract. *When a writer expresses himself, he must decide among a wealth of choices, such as which words/expressions to use or how to punctuate his writing. These choices define the writer's individual characteristics and stylometry is the quantitative study of such writing style. This paper aims to identify the books of writers with well-defined writing styles, Guimarães Rosa and Clarice Lispector, by means of lexical attributes found in their texts: letter frequency, word frequency and TF-IDF. Attributes are compared using the Euclidean distance, cosine similarity and Jaccard similarity index. The results show that by using the set of words with Jaccard similarity index it was possible to separate the books according to authorship.*

Resumo. *Quando um escritor se expressa, deve se decidir entre uma série de escolhas, tais como quais palavras/expressões usar ou como deve ser a pontuação da leitura. Essas escolhas definem as características individuais do escritor e a estilometria é a estudo quantitativo desse estilo de escrita. A proposta deste trabalho é conseguir identificar os livros de dois escritores com estilos bem definidos, Guimarães Rosa e Clarice Lispector, por meio dos atributos léxicos de seus textos: frequência de letras, frequência de palavras e TF-IDF. A comparação dos atributos é feita pela distância euclidiana, similaridade cosseno e similaridade de Jaccard. Os resultados mostram que o uso do conjunto de palavras com similaridade de Jaccard foi possível separar os livros por sua autoria.*

1. Introdução

Estilometria é o campo da linguística que analisa textos a partir de atributos mensuráveis [Juola 2013]. Os atributos mensuráveis podem ser, por exemplo: as frequências de letras, as frequências de palavras, o comprimento das palavras, o comprimento das sentenças, o uso de palavras incomuns e uso de sequência de palavras [Venčkauskas et al. 2015]. O conjunto de características obtido definirá o estilo de cada autor. Assim, a estilometria é frequentemente usada para atribuir autoria a documentos anônimos ou que estão sob contestação de autoria.

A busca pelo melhor e menor conjunto de atributos que definem a assinatura de uma autoria de textos é tema de estudo há décadas [Yule 1939], e continua presente em pesquisas [Abbasi and Chen 2008]. O trabalho de Akimushkin et al. (2017) e o trabalho de Antiqueira et al. (2007) usam teoria complexa de redes para identificação automática de autoria em documentos. Já o trabalho de Amancio et al. (2015) usa um modelo híbrido de classificação baseado em padrões de classificação e características topológicas de redes colaborativas para fazer a desambiguação entre autores. Outras propostas recentes usam modelo de linguagem sintática para estudar o significado das estruturas das sentenças [Sundararajan and Woodard 2018] [Kutuzov and Kuzmenko 2015], e o trabalho de Gamon [Gamon 2004] usa técnicas de aprendizagem de máquina para o problema de autoria.

Diferente dos trabalhos citados, a premissa deste trabalho é que: quando há estilos bem definidos de escrita, então não são necessárias técnicas complexas para identificação entre diferentes escritores. A hipótese deste trabalho é que é possível se fazer a separação de dois escritores de estilos diferentes de escrita em português brasileiro usando medidas de distância/similaridade de atributos léxicos de seus livros.

Este trabalho se propõe a examinar livros de Guimarães Rosa e Clarice Lispector por meio de dois processos. O primeiro analisa as seguintes características de estilometria: contagem de letras, contagem de palavras e importância da palavra, usando duas formas de comparação [López-Escobedo et al. 2016], distância euclidiana e similaridade cosseno. Já o segundo processo se vale da similaridade de Jaccard para fazer a comparação. Como as semelhanças e distâncias são conceitos complementares, essa generalização também se aplica a distâncias de elementos, uma vez que qualquer função de distância pode ser transformada em uma função de similaridade [Jimenez et al. 2016].

Para realizar os experimentos de comprovação da hipótese, foram selecionados dois escritores consagrados pelos seus livros em língua portuguesa: Guimarães Rosa e Clarice Lispector. Guimarães Rosa foi um dos escritores selecionados por ter um estilo bem definido. Muitas de suas obras foram ambientadas no sertão brasileiro, com ênfase nos temas nacionais, marcadas pelo regionalismo e mediadas por uma linguagem inovadora: invenções linguísticas, arcaísmos, palavras populares e neologismos [Rosa et al. 2006].

Clarice Lispector também tem um estilo bem definido. Foi escolhida por ser um contraponto ao estilo de Guimarães Rosa, ela não tratou tanto de regionalismos e preferiu manter seu foco na análise psicológica dos personagens, seus medos, sentimentos e angústias. A marca literária mais forte de Clarice é a literatura intimista, voltada para a natureza humana e para sua consciência. Clarice é considerada a maior expoente da literatura introspectiva do Brasil [Nunes 1989].

A maior contribuição deste trabalho é esta comparação estilométrica destes dois ilustres escritores de língua portuguesa [Bueno 2001]. Embora existam outras pesquisas de autoria em língua portuguesa, [Varela et al. 2011] [Pavelec et al. 2006] [Corso et al. 2005] [Honório et al. 2007], nenhum deles se concentra no estudo de características estilométricas léxicas destes escritores específicos. Este trabalho se limita ao estudo de caso dos livros de dois escritores, assim não é possível avaliar se os resultados se repetiriam em textos curtos, tais como e-mails, *posts* em fóruns ou redes sociais.

O artigo se organiza da seguinte forma: a Seção 2 versa sobre o referencial teórico das técnicas utilizadas, a Seção 3 sobre os materiais e métodos do trabalho, a Seção 4 sobre os resultados dos experimentos e discussão sobre os mesmos, finalizando na Seção 5 com as conclusões e trabalhos futuros.

2. Referencial Teórico

Nesta seção serão detalhadas as características estilométricas e as métricas de distância/similaridade usadas.

2.1. Características Estilométricas

De acordo com Alzahrani, Salim e Abrahama (2011), as características estilométricas podem ser classificadas em 4 grupos: léxicas, sintáticas, semânticas e específica da aplicação. Este trabalho usa apenas as características léxicas.

As características léxicas são baseadas nos caracteres e palavras do texto. A representação baseada em caracteres é a forma mais simples pela qual um documento d é representado, como a frequência de caracteres $d = (c_1, d), (c_2, d), ..., (c_n, d)$, onde (c_i, d) refere-se ao i-ésimo caractere em d, e n é a quantidade de caracteres em d.

A representação baseada em palavras pode ser pelo conjunto de palavras existentes no documento ou pela frequência das palavras. Nesse último, o documento d é representado como a frequência de palavras $d = (w_1, d), (w_2, d), ..., (w_n, d)$, onde (w_i, d) é a quantidade da i-ésima palavra em d, e n é a quantidade total de palavras diferentes em d, sendo que as estruturas das sentenças são ignoradas. Esta contagem de palavras é a técnica *bag-of-words* (BoW).

Uma outra representação baseada em palavras é o TF-IDF (*Term Frequency-Inverse Document Frequency*), que é uma medida estatística que tem o intuito de indicar a importância de uma palavra de um documento em relação a uma coleção de documentos. O valor de uma palavra, nessa representação, aumenta de acordo com o número de ocorrências do termo no documento e a raridade do termo na coleção [Jimenez et al. 2016].

As características sintáticas envolvem o uso de classes gramaticais (substantivo, verbo, adjetivo, pronome, artigo, numeral, preposição, conjunção, interjeição e advérbio) conhecido como POS *tagging* (*Part-Of-Speech tagging*); representação baseada em sentenças avaliando o uso de delimitadores de frases, tais como pontuações e a ordem das palavras. As características semânticas quantificam o uso de sinônimos e antônimos. Assim, é necessária a utilização de dicionários e bases de dados léxico. Muitas vezes usada em conjunto com o POS *tagging*. O último grupo usa recursos específicos da aplicação, que refletem a organização do texto, as palavras-chave específicas do conteúdo e/ou outros recursos específicos do idioma.

2.2. Métricas de Distância

A distância euclidiana é a distância em linha reta entre dois pontos no espaço euclidiano. Para pontos n-dimensionais, a distância euclidiana D_e é computada mostrada na Tabela 1. Quanto menor a distância euclidiana, maior a semelhança entre os vetores.

A similaridade cosseno permite calcular a similaridade entre dois vetores com n dimensões determinando o cosseno do ângulo entre eles. Essa métrica é muito utilizada

em mineração de texto [Ghosh and Strehl 2006] [Lima and Maia 2018]. Sendo dois vetores A e B, o ângulo θ é obtido pelo produto escalar e a norma entre os vetores, conforme Tabela 1.

Como os valores de $\cos\theta$ estão situados no intervalo de $[-1, 1]$, o valor -1 indica vetores opostos, 0 indica vetores independentes (ortogonais) e 1 indica vetores similares (colineares de coeficiente positivo). Valores intermediários permitem avaliar o grau de similaridade. A faixa foi normalizada para $[0, 1]$

O Coeficiente de Jaccard é uma medida da similaridade e diversidade entre dois conjuntos [Leydesdorff 2008]. Ele mede a similaridade entre dois conjuntos finitos e é definido pela divisão do tamanho da interseção pelo tamanho da união dos conjuntos, conforme Tabela 1. Os valores do coeficiente podem variar de 0 a 1. Quanto mais próximo de 1, mais os dois conjuntos são similares, ao passo que quanto mais próximo de 0, mais diferente eles são.

Tabela 1. Métricas usados nos experimentos.

Métrica	Equação	Faixa
Euclidiano	$D_e(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$	0 a ∞
Cosseno	$\cos\theta = \frac{A.B}{\|\|A\|\|.\|\|B\|\|}$	$[0, 1]$
Jaccard	$J(A,B) = \frac{\|A \cap B\|}{\|A \cup B\|} = \frac{\|A \cap B\|}{\|A\|+\|B\|-\|A \cap B\|}$	$[0, 1]$

3. Materiais e Métodos

Nesta seção serão detalhados todos os materiais e métodos necessários para os experimentos deste trabalho. Todas as técnicas foram desenvolvidas em linguagem Python 3.7.3 com a biblioteca NLTK.

3.1. Materiais

Para montar a base de dados deste estudo, foram utilizados nove livros de Guimarães Rosa e cinco livros de Clarice Lispector. Estes livros foram obtidos por meio de *download* no site português LeLivros[1] criado em 2013 que disponibiliza milhares de títulos gratuitamente. O LeLivros é um site sem fins lucrativos em que as obras são disponibilizadas no formato 'PDF', 'ePUB' e 'Mobi'.

Os livros foram baixados no formato 'ePUB', e posteriormente foram transformados para o formato 'txt' por meio do conversor *online* gratuito Convertio[2]. Na Tabela 2, há a lista dos livros do Guimarães Rosa usados nos experimentos, bem como o número de referência (coluna ID) do livro que é usado nos resultados. Da mesma forma, a Tabela 3 para a Clarice Lispector.

3.2. Métodos

Este trabalho separa os livros de Guimarães Rosa e de Clarice Lispector usando métricas de distância e similaridade. Isto é realizado por meio de dois processos. Em ambos os processos, os arquivos dos livros são pré-processados para gerarem um corpora.

[1] lelivros.love

[2] convertio.co/pt

Tabela 2. Livros de Guimarães Rosa usados nos experimentos.

ID	Título do livro
1	Grande Sertão: Veredas
2	Saragana
3	Manuelzão e Miguilim
4	No Urubuquaquá, no Pinhém
5	Noite do Sertão
6	Primeiras Estórias
7	Tutameia
8	Estas estórias
9	Ave, Palavra

Tabela 3. Livros de Clarice Lispector usados nos experimentos.

ID	Título do livro
10	A Hora da Estrela
11	Água Viva
12	Laços de Família
13	Perto do Coração Selvagem
14	Um Sopro de Vida

Foram usados os mesmos critérios que o trabalho de Pavelec, Justino e Freitas (2006). Assim, no pré-processamento para separação de cada elemento (ou *token*):

- Separador de palavras: espaço em branco, final de linha e caracteres não considerados *tokens*.
- Pontuações não foram consideradas *tokens*;
- Não houve diferenciação entre letras maiúsculas e minúsculas;
- Não foram considerados algarismos como *tokens*;
- Caracteres especiais não foram considerados *tokens*.

No primeiro experimento, de cada livro do corpora, extraem-se os atributos por três características distintas: frequência de letras, contagem de palavras e a técnica TF-IDF. O cálculo da frequência das letras é realizado por livro. Vetores são gerados contendo as informações de frequência para cada texto para posterior avaliação. A contagem de palavras também é feita por livro. Cada livro gera um vetor em que há computadas a quantidade de palavras do livro no vocabulário gerado por todo o corpora. Os valores do TF-IDF também são calculados por palavra e por livro, sendo assim cada obra tem um vetor com suas medidas individuais.

Então, essas características são comparadas entre si utilizando tanto a distância euclidiana quanto a similaridade cosseno. Um diagrama demonstrando este processo está presente na Figura 1.

No segundo experimento, após o pré-processamento, é feita a segmentação das palavras (em inglês *tokenize*) de cada livro. É aplicada a função que calcula a similaridade Jaccard a todos os 14 livros, em pares, e é feita a avaliação dos resultados. Na Figura 2 é possível ver um diagrama do processo.

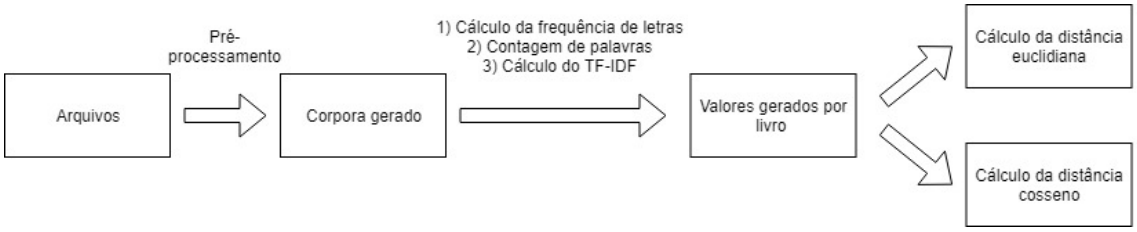


Figura 1. Fluxograma do primeiro experimento.

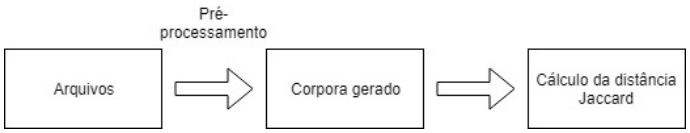


Figura 2. Fluxograma do segundo experimento.

4. Resultados dos Experimentos e Discussão

Para realizar os experimentos, foram usados 14 livros, sendo nove de Guimarães Rosa e cinco de Clarice Lispector. Estes livros formaram um corpora que foi posteriormente examinado por dois processos.

4.1. Primeiro Experimento

O primeiro experimento consistiu em usar a distância euclidiana e a distância cosseno para avaliar alguns parâmetros: frequência de letras, contagem de palavras e o TF-IDF de cada texto.

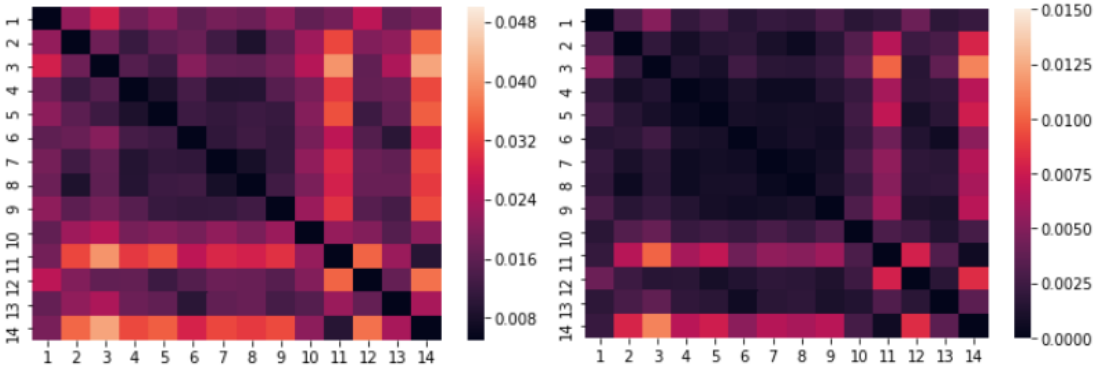
Todos os resultados são tabulares, mas para facilitar a compreensão dos resultados, os mesmos serão apresentados na forma de *heatmap*, que é a representação gráfica de dados, em que os valores individuais contidos em uma matriz são representados como cores.

4.1.1. Frequência de letras

A Figura 3 apresenta um *heatmap* da avaliação da frequência de letras por meio da distância euclidiana(a) e da distância cosseno(b), sendo que a identificação de cada livro pode ser observada nas Tabelas 2 e 3. Examinando o *heatmap* da Figura 3(a) e (b), constata-se que nas diagonais os valores são iguais a zero, pois a comparação é feita entre dois documentos iguais.

Os livros que mais chamam atenção na Figura 3(a) são o 11 e o 14, “Água Viva” e “Um sopro de vida”, ambos de Clarice Lispector, pois possuem valores de distância euclidiana maiores que a média. Entretanto, não há uma divisão clara entre os autores, pois estes livros apresentam valores de distância alto tanto com livros da mesma autora quanto com livros de Guimarães Rosa.

Os livros 12 e o 13, “Laços de Família” e “Perto do Coração Selvagem”, da Clarisse Lispector estão mais próximos do livro do Guimarães Rosa, e assim, reitera-se que não há uma divisão entre os escritores.



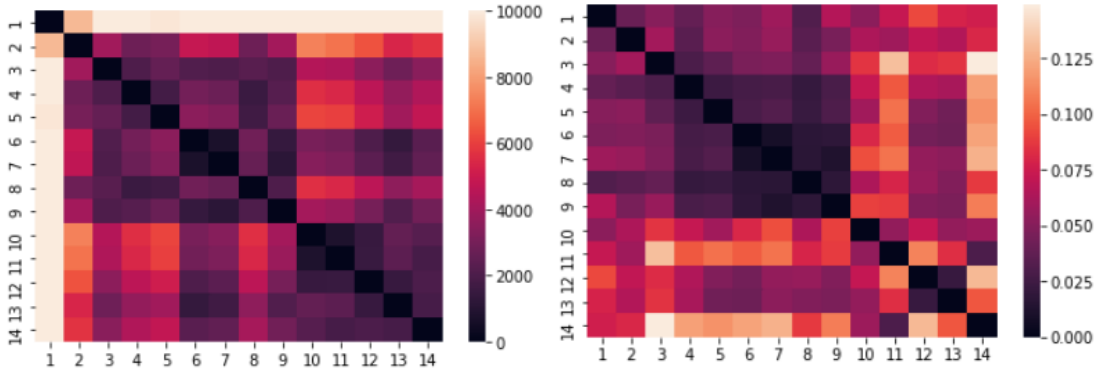
(a) euclidiano (b) cosseno

Figura 3. Heatmap da avaliação da frequência de letras entre os 14 livros do corpora.

Os resultados apresentados no *heatmap* da Figura 3(b) ficaram qualitativamente muito parecidos com os obtidos no *heatmap* da Figura 3(a), de forma que as mesmas conclusões aplicadas a um podem ser aplicadas ao outro.

4.1.2. Contagem de palavras

A Figura 4 apresenta um *heatmap* com os resultados da avaliação da contagem de palavras por meio da (a) distância euclidiana e da (b) similaridade cosseno. Da mesma forma, observa-se que as diagonais tem valores iguais a zero.



(a) euclidiano (b) cosseno

Figura 4. Heatmap da avaliação da frequência de palavras entre os 14 livros do corpora.

É possível notar que os valores obtidos com relação ao livro 1, “Grande Sertão: Veredas”, de Guimarães Rosa possui uma distância euclidiana muito superior a todas as outras distâncias calculadas, indiferente do autor.

Apesar disto, há alguns resultados considerados satisfatórios. Ao ignorar o resultado do livro 1, considerando-o como um *outlier*, os livros 2, 4, 5 e 8 (“Sagarana”, “No Urubuquaquá, no Pinhém”, “Noite do Sertão” e “Estas Estórias”), todos de Guimarães Rosa, possuem distâncias euclidianas calculadas condizentes com o autor, ou seja, distâncias menores ao serem comparadas com outros livros de Rosa e distâncias maiores

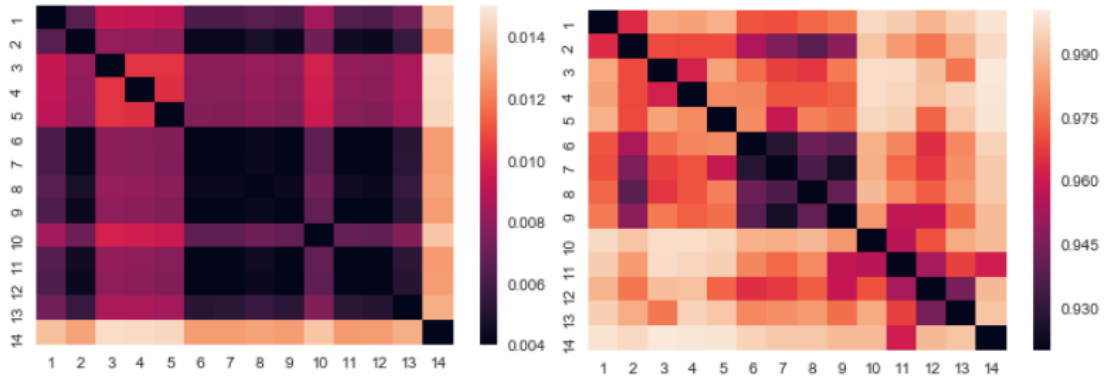
ao serem comparadas com livros de Clarice Lispector. São as células com os valores mais escuros/baixos para Guimarães Rosa e valores mais claros/altos para Clarisse Lispector.

No *heatmap* da Figura 4(b), é possível comparar que o resultado com a similaridade do cosseno foi melhor que com a distância euclidiana. Dado que a região com 9 primeiros livros do Guimarães Rosa apresentam valores escuros/baixos.

No entanto, os livros 12 e o 13, "Laços de Família" e "Perto do Coração Selvagem", ambos de Clarice Lispector, possuem valores que se confundem com os livros de Guimarães Rosa. De novo, não há uma divisão evidente entre os autores.

4.1.3. TF-IDF

A Figura 5 mostra dois *heatmaps* dos resultados das avaliações do TF-IDF entre os livros por meio da (a) distância euclidiana e da (b) similaridade cosseno. Novamente, as diagonais tem valores iguais a zero, conforme esperado.



(a) euclidiano (b) cosseno

Figura 5. Heatmap da Avaliação do TF-IDF entre os 14 livros do corpora.

Para avaliar se o uso da distância euclidiana sobre as medições do TF-IDF é uma boa métrica para identificar livros, é necessário observar se os livros de Guimarães Rosa possuíram valores de distância euclidiana baixo entre eles comparados aos valores que os mesmos obtiveram com os livros de Clarice Lispector, e vice-versa.

Dos valores obtidos não é possível fazer uma divisão entre os livros de Guimarães Rosa e Clarice Lispector. É possível, entretanto, destacar alguns livros. O livro 14, "Um sopro de vida", de Clarice Lispector, é o mais significativamente diferente dos outros, já que obteve os maiores valores de distância euclidiana com todos os outros livros, inclusive os da mesma autora, chegando a valores superiores a 0,012.

Outros livros que merecem atenção são os livros 3, 4 e 5, "Manuelzão e Miguilim", "No Urubuquaquá, no Pinhém" e "Noites do Sertão", todos de Guimarães Rosa. Estes livros também obtiveram valores mais altos de distância euclidiana comparados com outros livros. É interessante notar que originalmente estes três romances foram publicados juntos com o nome de "Corpo de Baile" e só a partir da segunda edição que foram publicados separadamente. No entanto, mesmo possuindo essa ligação, eles também apresentaram valores de distância euclidiana altos entre si.

Os resultados mostrados no *heatmap* da Figura 5(b), na qual os livros são avaliados pela distância cosseno aos pares. Assim como na análise anterior, para avaliar se o uso da distância cosseno sobre as medições do TF-IDF é uma boa métrica para identificar livros, também é necessário observar se os livros de Guimarães Rosa possuíram valores de distância cosseno baixo entre eles comparados aos valores que os mesmos obtiveram com os livros de Clarice Lispector, e vice-versa.

Examinando os valores obtidos por esta métrica, constata-se que todos eles são altos e próximos a 1. Na verdade, nenhum destes valores é inferior a 0,93. Esta análise indica que todos os documentos são muito diferentes entre si, pois valores altos de distância cosseno significam baixos valores de similaridade cosseno, que por sua vez determinam alto ângulo entre os vetores.

Há, contudo, um subconjunto de livros do Guimarães Rosa que estão agrupados, os livros de 6 à 9, que formam um grupo similar entre eles e separável de todos os outros.

4.2. Segundo Experimento

O segundo experimento usou a similaridade Jaccard para analisar a possibilidade de identificação dos livros. A Tabela 4 mostra os resultados das comparações entre os livros utilizando a similaridade Jaccard, sendo que a identificação de cada livro pode ser observada nas Tabelas 2 e 3.

Tabela 4. Avaliação por meio da similaridade Jaccard dos 14 livros do corpora

Livro	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	0,963981	0,98651	0,984734	0,98812	0,97154	0,970541	0,974164	0,977653	0,99607	0,99349	0,98856	0,993793	0,998192
2	0,963981	0	0,969963	0,96948	0,969974	0,954186	0,945381	0,938447	0,948365	0,99173	0,983717	0,977035	0,987009	0,996069
3	0,98651	0,969963	0	0,961874	0,985135	0,975374	0,968282	0,966496	0,977507	0,996933	0,996655	0,990326	0,977413	0,999202
4	0,984734	0,96948	0,961874	0	0,980707	0,979929	0,971419	0,971375	0,973339	0,996968	0,995551	0,991133	0,994831	0,998977
5	0,98812	0,969974	0,985135	0,980707	0	0,980886	0,959361	0,978544	0,975796	0,995571	0,993824	0,973781	0,992236	0,998491
6	0,97154	0,954186	0,975374	0,979929	0,980886	0	0,928319	0,942061	0,938765	0,987904	0,979771	0,964817	0,980023	0,994758
7	0,970541	0,945381	0,968282	0,971419	0,959361	0,928319	0	0,936116	0,924401	0,987239	0,974893	0,96679	0,981291	0,992508
8	0,974164	0,938447	0,966496	0,971375	0,978544	0,942061	0,936116	0	0,940372	0,989471	0,980517	0,972986	0,983665	0,992618
9	0,977653	0,948365	0,977507	0,973339	0,975796	0,938765	0,924401	0,940372	0	0,982857	0,95806	0,958693	0,975903	0,990017
10	0,99607	0,99173	0,996933	0,996968	0,995571	0,987904	0,987239	0,989471	0,982857	0	0,956611	0,97095	0,98684	0,990127
11	0,99349	0,983717	0,996655	0,995551	0,993824	0,979771	0,974893	0,980517	0,95806	0,956611	0	0,952988	0,968189	0,960828
12	0,98856	0,977035	0,990326	0,991133	0,973781	0,964817	0,96679	0,972986	0,958693	0,97095	0,952988	0	0,944127	0,989453
13	0,993793	0,987009	0,977413	0,994831	0,992236	0,980023	0,981291	0,983665	0,975903	0,98684	0,968189	0,944127	0	0,991701
14	0,998192	0,996069	0,999202	0,998977	0,998491	0,994758	0,992508	0,992618	0,990017	0,990127	0,960828	0,989453	0,991701	0

A Figura 6 mostra um *heatmap* dos resultados das comparações entre os livros utilizando a similaridade Jaccard.

Examinando os valores obtidos na Tabela 4 e na Figura 6, percebe-se que há uma diagonal de números 1. Estes valores já eram aguardados, uma vez que a diagonal representa a comparação de um livro com ele mesmo, e na similaridade Jaccard, quando os conjuntos são iguais, a similaridade é 1.

É possível constatar que os livros de 1 a 9 possuem maior similaridade Jaccard entre eles do que quando comparados estes livros com os livros de 10 a 14. Da mesma forma, os livros de 10 a 14 também possuem maior similaridade Jaccard entre eles do que quando comparados estes livros com os livros de 1 a 9. Assim, é possível se separar os autores pela similaridade Jaccard.

Quando comparados livros do mesmo autor, o valor da similaridade Jaccard foi superior a 0,200; e quando comparados livros de autores diferentes, o valor da similari-

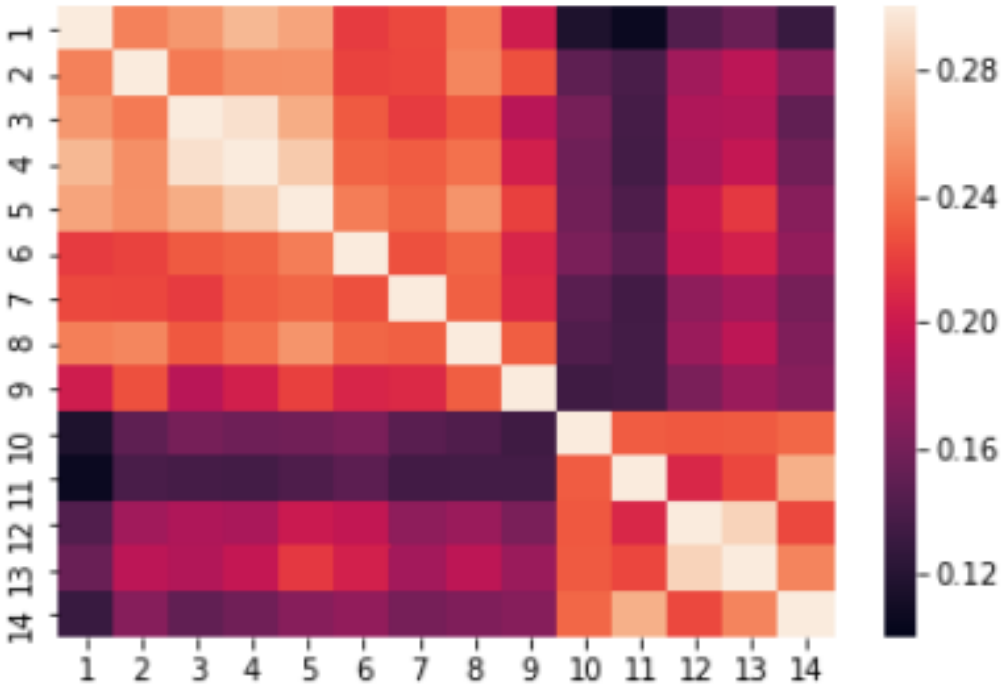


Figura 6. *Heatmap* da avaliação da similaridade Jaccard entre os 14 livros do corpora.

dade Jaccard ficou abaixo de 0,200. A única exceção foram os pares 5-13 e 6-13, que embora sejam livros cada um de um autor, apresentaram valores de similaridade Jaccard um pouco acima de 0,200 (0,217 e 0,204 respectivamente); além do par 3-9, que apesar de serem do mesmo autor, Guimarães Rosa, apresentou um valor de similaridade Jaccard um pouco abaixo de 0,200 (0,191).

5. Conclusão e Trabalhos Futuros

A hipótese estudada é se é possível, por meio de atributos estilométricos léxicos e um método de comparação, avaliar se os livros de Guimarães Rosa são mais parecidos entre si e distinguíveis dos livros de Clarice Lispector.

A técnica da similaridade Jaccard foi suficiente para separar a autoria dos livros, mesmo sendo uma técnica mais simples que as anteriores.

Embora simples, a similaridade de Jaccard avalia o conjunto de palavras que se encontram em ambos os documentos comparados. Este é o métodos que melhor captura as idiossincrasias do estilo de Guimarães Rosa e corrobora a ideia de Koppel e Schler (2003), que propõe explorar as características discriminatórias para atribuição de autoria.

Neste trabalho, ressalta-se que os escritores analisados tem estilos bem marcantes. E portanto, não se pode concluir que os mesmos métodos teriam os mesmos resultados com outros escritores. Todo autor tem estilo. Hemingway (norte-americano, ganhador do Nobel de Literatura), por exemplo, é quase o oposto de Guimarães Rosa, pois escreve numa linguagem quase jornalística (simples, direta, transparente), e nem por isso menos interessante. Mas para avaliação de métodos quantitativos, se um autor não tem um estilo tão bem definido, seria possível avaliar a autoria de seu texto?

O questionamento existe e mais ainda, é necessário mais estudos sobre escritores e textos em português brasileiro. Assim, os resultados deste trabalho são motivadores para continuar o estudo, pois deve-se aumentar a quantidade de escritores e seus textos para serem analisados. Com isso, provavelmente mais atributos devem ser estudados, como o conjunto de palavras mais frequentes por autor, bem como empregar categorias de palavras mais específicas(adjetivos, advérbios, tempos verbais) além de outras técnicas de comparação.

É interessante também o aprofundamento na pesquisa sobre nuances mais sutis de diferenças textuais e o ritmo do texto, como o autor usa a pontuação, por exemplo. Neste trabalho, não foi feita a diferenciação entre maiúsculas e minúsculas, mas grafar com maiúscula ou minúscula, em literatura, pode fazer diferença, como por exemplo, textos fabulares e o teatro de moralidades medieval.

Referências

Abbasi, A. and Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):7.

Akimushkin, C., Amancio, D. R., and Oliveira Jr, O. N. (2017). Text authorship identified using the dynamics of word co-occurrence networks. *PloS one*, 12(1):e0170527.

Alzahrani, S. M., Salim, N., and Abraham, A. (2011). Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(2):133–149.

Amancio, D. R., Oliveira Jr, O. N., and Costa, L. d. F. (2015). Topological-collaborative approach for disambiguating authors' names in collaborative networks. *Scientometrics*, 102(1):465–485.

Antiqueira, L., Pardo, T. A. S., Nunes, M. d. G. V., and Oliveira Jr, O. N. (2007). Some issues on complex networks for author characterization. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 11(36):51–58.

Bueno, L. (2001). Guimarães, Clarice e antes. *Teresa*, (2):249–261.

Corso, G., Fossa, C. R., and de Oliveira, G. B. (2005). Uma aplicação da teoria de redes a estilometria: Comparando machado de assis e tribuna do norte. *Revista Brasileira de Ensino de Física*, 27(2):389–393.

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics*, page 611. Association for Computational Linguistics.

Ghosh, J. and Strehl, A. (2006). Similarity-based text clustering: a comparative study. In *Grouping Multidimensional Data*, pages 73–97. Springer.

Honório, T. C. S., Nobre Neto, F. D., Almeida, T. P., Duarte, R. C. M., Barbosa, Y. A. M., Rocha, V. M., and Batista, L. V. (2007). Atribuição de autoria com WEKA. In *Anais do IX Encontro de Extensão e X Encontro de Iniciação*, pages 42–42. Editora Universitária/UFPB.

Jimenez, S., Gonzalez, F. A., and Gelbukh, A. (2016). Mathematical properties of soft cardinality: Enhancing jaccard, dice and cosine similarity measures with element-wise distance. *Information Sciences*, 367:373–389.

Juola, P. (2013). Stylometry and immigration: A case study. *Journal of Law and Policy*, 21(2):287–298.

Koppel, M. and Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, volume 69, pages 72–80.

Kutuzov, A. and Kuzmenko, E. (2015). Comparing neural lexical models of a classic national corpus and a web corpus: the case for russian. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 47–58. Springer.

Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data: Salton's cosine versus the Jaccard index. *Journal of the American Society for Information Science and Technology*, 59(1):77–85.

Lima, J. M. C. and Maia, J. E. B. (2018). A topical word embeddings for text classification. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*, pages 25–35. SBC.

López-Escobedo, F., Solorzano-Soto, J., and Sierra Martínez, G. (2016). Analysis of intertextual distances using multidimensional scaling in the context of authorship attribution. *Journal of Quantitative Linguistics*, 23(2):154–176.

Nunes, B. (1989). *O drama da linguagem: uma leitura de Clarice Lispector*, volume 12. ŧica.

Pavelec, D., Justino, E., and Freitas, C. (2006). Identificação da autoria de documentos digitais com base em atributos estilométricos da língua portuguesa. In *TIL-06, 4o workshop em Tecnologia da Informaçao e da Linguagem Humana*, pages 1659–1668.

Rosa, J. G., de Athayde Sandroni, L. C. A., and de Aguiar, F. W. (2006). *João Guimarães Rosa*. Editora Nova Fronteira.

Sundararajan, K. and Woodard, D. (2018). What represents "style" in authorship attribution? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2814–2822.

Varela, P. J., Justino, E. J., and Oliveira, L. E. (2011). Identificação de autoria de textos através do uso de classes linguísticas da língua portuguesa (authorship identification using linguistic classes for portuguese)[in portuguese]. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology*.

Venčkauskas, A., Damaševičius, R., Marcinkevičius, R., and Karpavičius, A. (2015). Problems of authorship identification of the national language electronic discourse. In *International Conference on Information and Software Technologies*, pages 415–432. Springer.

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrica*, 30:363–390.