MULTÍPLOS OLHARES EM CIÊNCIA DA INFORMAÇÃO
Vol. 9, Nº. 2 Ano 2019

AGRUPAMENTO AUTOMÁTICO DE NOTÍCIAS DE JORNAIS *ON-LINE* USANDO TÉCNICAS DE *MACHINE LEARNING* PARA *CLUSTERING* DE TEXTOS NO IDIOMA PORTUGUÊS

Lúcia Helena de Magalhães
Universidade Federal de Minas Gerais
lhmag@yahoo.com.br

Renato Rocha Souza
Universidade Federal de Minas Gerais
rsouza.fgv@gmail.com

RESUMO

Clusterização é uma técnica de organizar dados em grupos cujos membros apresentam alguma similaridade. Assim, esta pesquisa teve como objetivo utilizar as técnicas de Processamento de Linguagem Natural, *Machine Learning* e *Clustering* para criar aglomerados de notícias a partir de uma amostra coletada dos principais jornais *on-line*. Verificou-se que a etapa de pré-processamento exige um esforço para garantir a qualidade dos resultados. A complexidade da língua portuguesa, a necessidade de atualização da lista de *stopwords*, as dificuldades relacionadas à detecção das características mais importantes e à alta dimensionalidade dos dados foram evidenciadas durante todas as etapas deste estudo. O algoritmo de agrupamento *k-means* obtive os melhores resultados para esse tipo de informação e o *Hierarchical Clustering* teve dificuldades, visto que notícias semelhantes foram alocadas em grupos diferentes. Já o *Affinity Propagation* apresentou divergência quanto ao número ideal de clusters, mas conseguiu um bom desempenho ao agrupar por semelhança.

Palavras-Chave: Agrupamento de notícias, Processamento de Linguagem Natural, Aprendizado de Máquina, Análise de textos.

ABSTRACT

Clustering is a technique of organizing data into groups whose members have some similarity. Thus, this research aimed to use Natural Language Processing, Machine Learning and Clustering techniques to create news clusters from a sample collected from leading online newspapers. It has been found that the preprocessing step requires an effort to ensure the quality of the results. The complexity of the Portuguese language, the need to update the stopwords list, the difficulties related to the detection of the most important characteristics and the high dimensionality of the data were evidenced during all stages of this study. The k-means clustering algorithm obtained the best results for this type of information and Hierarchical Clustering had difficulties, since similar news was allocated in different groups. Affinity Propagation, on the other hand, differed in the optimal number of clusters, but achieved a good performance by grouping by similarity.

Keywords: News Clustering, Natural Language Processing, Machine Learning, Text Analysis.

Universidade Federal de Minas Gerais

Correspondência/Contato
Av. Antônio Carlos, 6627
Pampulha: 31270-901
BELO HORIZONTE - MG

Escola de Ciência da Informação da UFMG

1. INTRODUÇÃO

Com a expansão da grande rede mundial de computadores e o crescimento da popularidade da web, um amplo volume de dados e informação é gerado e publicado em inúmeras páginas da internet nas quais se encontram dados das mais diversas áreas do conhecimento. Deste modo, torna-se imprescindível o desenvolvimento de métodos automatizados de análise de texto para extrair conhecimento de documentos não estruturados da web e criar versões condensadas de informações que possam facilitar a busca por um determinado assunto.

Assim, a proposta é utilizar as técnicas de Processamento de Linguagem Natural, *Machine Learning* e *Clustering* para criar grupos de notícias semelhantes a partir de uma amostra recuperada dos jornais *on-line*. Além do mais, existem poucos estudos relacionados ao tema clusterização de notícias publicadas no idioma português. A lacuna de pesquisas nessa área acaba por reforçar e aprofundar a escassez de informação relacionada a algumas questões: como desenvolver uma solução automatizada capaz de recuperar e comparar as notícias em destaque na mídia, publicadas no idioma português, e agrupá-las por similaridade? As tecnologias existentes apresentam o mesmo desempenho quando alimentadas por um corpus de notícias publicadas nessa língua? Qual técnica apresenta melhor resultado para esse tipo de informação? Os algoritmos apresentam o mesmo desempenho ao ser alimentados com corpora diversificados?

Para responder essas questões, este estudo objetiva utilizar uma metodologia de aprendizado não supervisionado que seja capaz de agrupar, automaticamente, notícias publicadas no idioma português do Brasil. Para isso, será compilado, mapeado, categorizado e analisado um conjunto de informes da atualidade. Para a captura do corpus, será usado o MediaFrame[1], uma ferramenta de código aberto, desenvolvida na Fundação Getúlio Vargas, que permite coletar um grande número de notícias das mídias *on-line*. Além disso, a pesquisa busca identificar quais são os principais métodos utilizados no processo de *clustering* de textos e aplicar essas técnicas em uma coleção de notícias para verificar o desempenho dos algoritmos de clusterização ao ser alimentados por esse tipo de informação. A metodologia também é aplicada em diferentes corpora para discutir o sucesso da técnica em cada caso, bem como averiguar a possibilidade efetiva de clusterização dos documentos e analisar as dificuldades encontradas para diferentes amostras.

[1] https://mediaframe.io

2. FUNDAMENTOS CONCEITUAIS

Neste tópico serão abordados os principais fundamentos teóricos necessários para uma melhor compreensão deste artigo. A seção foi dividida nos seguintes temas: Processamento de Linguagem Natural, Mineração de Textos, Aprendizado de Máquina e Clusterização.

2.1. Processamento de Linguagem Natural (PLN)

As pesquisas relacionadas ao PLN têm como objetivo fazer com que os computadores possam processar e compreender textos escritos em linguagem natural. Ladeira (2010, p. 43) descreve PLN como sendo a área responsável por manipular automaticamente a linguagem não controlada contida normalmente nos documentos textuais. É um campo de estudos que está preocupado em desenvolver técnicas computacionais para permitir que um computador entenda o significado do texto da linguagem natural (ZHAI; MASSUNG, 2016). É uma subárea da Computação, Inteligência Artificial (IA) e da Linguística que estuda os problemas da geração e compreensão automática de línguas humanas naturais (DAS, 2017, p. 2).

Atualmente, existe uma grande demanda em relação aos aplicativos de PLN e análise de texto, pois a capacidade de extrair informações úteis e *insights* de uma grande quantidade dados não estruturados e brutos é uma tarefa complexa. Segundo Aranha e Passos (2006), o PLN é uma técnica chave para análise de textos que utiliza dos conhecimentos da área da linguística para aproveitar ao máximo o conteúdo do documento, extraindo entidades, seus relacionamentos, detectando sinônimos etc. Participa, normalmente, na fase do pré-processamento dos dados, transformando-os em um formato que seja mais compreensível pelas máquinas.

2.2. Mineração de Textos

A Mineração de Textos, também conhecida como *Text Data Mining* ou *Knowledge Discovery in Texts*, é considerada uma evolução da área de Recuperação da Informação (RI) (SALTON; MCGILL, 1983). É um campo multidisciplinar que se baseia em Mineração de Dados, Aprendizado de Máquina, RI, Linguística Computacional e Estatística. Pode ser vista como uma extensão da área de Mineração de Dados que busca desenvolver técnicas e processos para descoberta automática de conhecimentos valio-

sos a partir de uma coleção de documentos. A Figura 1 mostra os passos necessários para a análise de dados não estruturados.

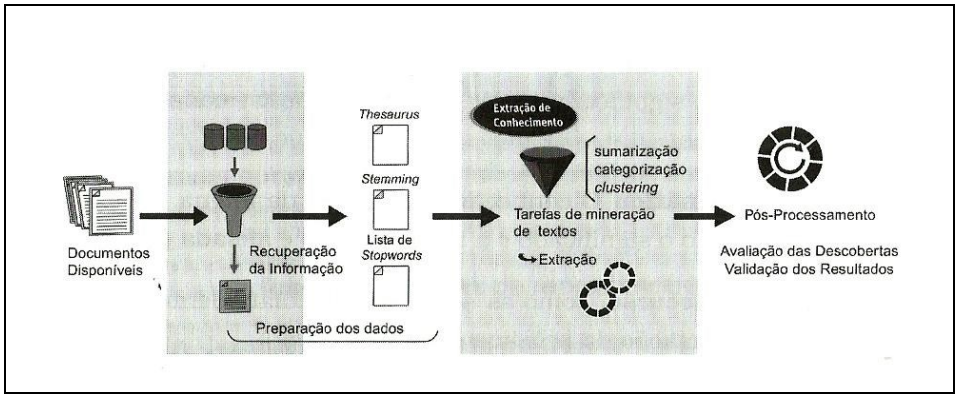

Figura 1. Etapas da Mineração de Textos (EBECKEN; LOPES; COSTA (2003, p. 339).

Após a coleta dos documentos, realiza-se a preparação do corpus, fase em que as *stopwords* (como artigos e preposições) são removidas e a técnica de *stemming* (redução das palavras distintas a sua raiz gramatical comum) é aplicada. Posteriormente, efetua-se o processamento para extração do conhecimento. Nessa fase, algumas atividades como sumarização, categorização e clusterização podem ser realizadas. Por fim, faz-se a análise e validação dos resultados obtidos (EBECKEN; LOPES; COSTA, 2003).

2.3. Aprendizado de Máquina

Aprendizado de Máquina (AM), ou *Machine Learning*, pode ser definido como uma área que pesquisa métodos computacionais relacionados à aquisição automática de novos conhecimentos, novas habilidades e novas formas de organizar a informação já existente (Mitchell, 1997). Na visão de Muller e Guido (2017), AM é um campo de pesquisa que tem uma interseção com Estatística, IA e Ciência da Computação e que também é conhecida como Análise Preditiva ou Aprendizagem Estatística. Baeza-Yates e Ribeiro-Neto (2013) afirmam que o AM é uma área ampla da IA que está preocupada com o projeto e o desenvolvimento de algoritmos que aprendem padrões presentes nos dados fornecidos como entrada. Nesse contexto, "os padrões aprendidos, que podem ser bem complexos, são então usados para fazer predições relativas a dados ainda não vistos e novos." (BAEZA-YATES; RIBEIRO-NETO, 2013, p. 278).

Com base nos critérios escolhidos para a realização de cada atividade, os métodos de aprendizado podem ser divididos em aprendizado supervisionado e aprendizado não supervisionado. Segundo Matos (*on-line*), o termo aprendizado supervisionado é usado sempre que o programa é "treinado" sobre um conjunto de elementos pré-

estabelecido. "Baseado no treinamento com os dados pré-definidos, o programa pode tomar decisões precisas quando recebe novos dados". O aprendizado supervisionado "refere-se à capacidade que determinados algoritmos têm de aprender a partir de exemplos." (GOLDSCHMIDT; PASSOS; BEZERRA, 2015, p. 73). É uma técnica que utiliza padrões para prever os dados do rótulo em dados adicionais não rotulados. Esse modelo de aprendizagem se divide em duas subcategorias: Classificação e Regressão.

No aprendizado não supervisionado não há classe associada aos exemplos e o indutor analisa os elementos fornecidos e tenta determinar se alguns deles podem ser agrupados de alguma maneira, formando grupos de objetos similares (CHEESEMAN; STUTZ, 1996). Nesse tipo de aprendizado, o conjunto de treinamento consiste apenas de exemplos sem nenhum valor associado. Tipicamente, o problema se resume em particionar a amostra em agrupamentos ou *clusters*, sem receber nenhum dado de treinamento, através de técnicas de clusterização. Assim, para propósito de categorização de textos, *clustering* é o tipo de algoritmo de aprendizado não supervisionado de maior interesse." (GONÇALVES, 2013, p. 278).

2.4. Clustering

Segundo Goldschmidt, Passos e Bezerra (2015, p. 25), a clusterização é uma técnica "utilizada para segmentar os registros de uma base de dados em subconjuntos ou *clusters*, de tal forma que os elementos de um *cluster* compartilhem propriedades comuns que os distingam dos elementos nos demais *clusters*". Segundo Hair *et al*. (2005), a análise de agrupamentos é uma técnica analítica para desenvolver subgrupos significativos de indivíduos ou objetos. Especificamente, o objetivo é categorizar uma amostra em um número menor de grupos mutuamente excludentes, com base nas similaridades entre as entidades.

Clusterização tem sido estudado nas áreas de PLN, Mineração de Dados, RI e Reconhecimento de Padrões. É um modelo de representação espacial, em que um conjunto de objetos é distribuído em subconjuntos menores, chamados de *clusters* (ZURINI; SBORA, 2011).

3. METODOLOGIA

Para as experimentações, primeiramente, coletaram-se as notícias e, em seguida, realizou-se o pré-processamento dos informes, etapa em que as *stopwords* foram removidas e as técnicas de tokenização (quebra do texto em palavras ou *tokens*) e stemming foram aplicadas. Com o corpus preparado, extraíram-se as principais características dos textos e os documentos foram representados em um modelo de espaço vetorial. A semelhança entre as notícias foi encontrada através do cálculo da similaridade. Imediatamente a técnica de *clustering* foi aplicada usando os algoritmos *k-means*, *Hierarchical Clustering* e o *Affinity Propagation* e os grupos foram formados. Para melhor visualização, validação e interpretação dos resultados, apresentaram-se os *clusters* em dendogramas e em diagramas de dispersão.

4. RESULTADOS E DISCUSSÕES

Neste experimento, optou-se por uma amostra contendo cinquenta notícias relacionadas aos temas biologia, futebol, eletricidade e economia. Realizou-se o pré-processamento e, em seguida, aplicou-se a técnica de *clustering* .

4.1. Teste usando o algoritmo k-means

O método *Elbow* é uma das mais tradicionais formas utilizadas para encontrar o número de k para uma determinada amostra. Ao aplicar essa técnica na coleção de notícias, ela apresentou uma inclinação em k=4, indicando o valor ideal para a quantidade de grupos. Posteriormente, extraíram-se as características principais de cada aglomerado e foram agrupadas as notícias que, de certa forma, apresentaram algum padrão ou similaridade.

Em relação à quantidade ideal de características, testes são necessários para descobrir o número que melhor representa os grupos, pois um número muito pequeno pode não ser suficiente para caracterizar o *cluster* e se o valor for muito grande, pode acontecer dos termos não retratarem realmente cada aglomerado. A tabela 1 apresenta as características extraídas.

Tabela 1. Características extraídas para cada *cluster*.

Clusters	Características
Cluster 1	biologia, vivo, trabalho, resultado
Cluster 2	copa, futebol, seleção, time
Cluster 3	carga, energia, carga elétrica, eletrifio
Cluster 4	preside, governo, economia, ministro

Observa-se no Cluster 4 que as principais características extraídas para esse grupo foram as palavras presidente, governo, economia e ministro. Isso quer dizer que as notícias pertencentes a esse aglomerado apresentam essas palavras com maior peso.

Ao analisar cada grupo, observa-se que o Cluster 1, que agrupou as notícias relacionadas ao tema biologia, conseguiu aglomerar um total de 13 textos. Como, inicialmente, foram coletadas somente 12 notícias relacionadas a esse tema, realizou-se uma análise das matérias para verificar se realmente foi pertinente a alocação realizada pelo algoritmo de *clustering*. Enfim, das 13 notícias classificadas no grupo 1, 11 realmente pertenciam ao grupo biologia. As outras duas matérias, que também foram agrupadas no primeiro *cluster*, apesar de ter sido coletadas como pertencente ao tema eletricidade, elas não têm relação com esse assunto e também não pertencem aos temas futebol e economia. Como essas duas notícias tratavam de assuntos relacionadas à antropologia, estudo do corpo humano, relações sexuais, estudo científico com macacos etc, o algoritmo acertou ao classifica-las como pertencentes ao grupo biologia, visto que entre os 4 temas, ele encontrou o *cluster* mais pertinente. Mas por que o coletor recuperou essas duas notícias ao pesquisar pelo tema eletricidade? As matérias tratavam do mesmo assunto, porém, recuperadas de jornais diferentes. Ao analisa-las, percebeu-se que no decorrer do texto apareceram várias vezes os termos "laboratório de carga", local usado para estudar o comportamento humano. Isso justifica o fato delas serem recuperadas como pertencentes ao tema eletricidade.

Em relação ao grupo eletricidade, composto por 12 documentos, o *k-means* agrupou 10 informes como pertencente a esse assunto. Como duas notícias, que foram coletadas como pertencentes ao tema eletricidade, foram aglomeradas no grupo biologia, parece que o algoritmo acertou 100% as demais. Porém, das 10 notícias pertencentes ao grupo eletricidade, uma foi recuperada da amostra de economia. Essa notícia realmente descrevia sobre economia, o algoritmo errou nesse caso. Todavia, como ele não

leva em consideração o contexto, algumas palavras que apareceram nas frases, como por exemplo, 'a China é uma potência', 'tensões comerciais', 'empresas de semicondutores', fizeram com que o *k-means* agrupasse essa notícia no *cluster* sobre eletricidade.

Dando continuidade à análise, o algoritmo agrupou muito bem as notícias relacionadas à economia. Das 13 notícias, ele acertou 12. Uma das matérias desse tema ficou no aglomerado sobre eletricidade, conforme comentado anteriormente, e outra notícia, pertencente a categoria futebol, foi alocada no grupo de economia. Entretanto, o algoritmo não errou nesse caso, pois o título da notícia era "Moro é o pai do combate à corrupção". Com base no título, esse documento não poderia ter sido recuperado pelo coletor, cuja palavra-chave da pesquisa foi futebol. Deste modo, para averiguar, foi necessária uma leitura completa da matéria. Assim, observou-se que a notícia realmente falava sobre Moro, porém, o autor faz uma comparação das ações do juiz com o futebol, que é um esporte movido por sentimentos e paixões. Nessa narração, a palavra futebol apareceu várias vezes, isso fez com que o coletor recuperasse essa notícia juntamente com as demais relacionadas ao tema futebol. Assim, não se pode afirmar que o algoritmo de *clustering* errou nesse caso, pois a notícia sobre Moro também continha parágrafos sobre a Reforma da Previdência e outros termos relacionados à economia.

Quanto às notícias cujos assuntos são referentes à futebol, o *k-means* agrupou 6 matérias como pertencentes a essa categoria e uma outra, também relacionada a esse tema, foi para o grupo sobre economia. No total, 8 notícias não foram agrupadas, sendo 6 pertencentes a amostra de notícias sobre futebol. A Tabela 2 resume os acertos e erros do algoritmo *k-means*, conforme já descrito.

Tabela 2. Acerto x erro do *k-means*.

Grupos	Total de notícias relacionadas ao assunto	Notícias agrupadas	Acerto	Erro	Agrupou em outras categorias
Cluster 1 Biologia	12	13	11/13	2/13	2 em eletricidade
Cluster 2 Futebol	13	6	6/6	0	0
Cluster 3 Eletricidade	12	10	9/10	1/10	1 em eletricidade
Cluster 4 Economia	13	13	12/13	1/13	1 em futebol
Total	50	42	38	4	

Por conseguinte, considerando a quantidade de notícias agrupadas, o *k-means* conseguiu uma taxa acima de 90% de acerto e agrupou 84% dos textos. Considerando

as particularidades das notícias, esse algoritmo obteve praticamente 100% de sucesso, ponderando apenas as matérias que foram agrupadas.

Com essas análises quantitativas, realizadas com base na leitura dos informes, identificou-se que nem todas as notícias foram agrupadas corretamente, porém, em alguns casos, o erro foi do coletor e não do algoritmo de *clustering*. Portanto, a análise dos aglomerados auxilia na identificação do conteúdo de documentos e das relações entre eles. Isso é relevante quando se trabalha com uma grande quantidade de textos, porque permite ao leitor identificar os grupos que contém os documentos que mais o interessa. Além disso, por se tratar de uma técnica não supervisionada, pode-se considerar que o algoritmo teve um excelente desempenho.

4.2. Teste usando o algoritmo *Affinity Propagation*

O A*ffinity Propagation* constrói os agrupamentos com base nas propriedades dos dados sem qualquer pressuposto sobre o número de *clusters*. Assim, ao ser alimentado pela coleção de notícias, essa técnica encontrou sete grupos, ou seja, k=7. Um valor diferente do que esperado, visto que a amostra era composta por quatro assuntos distintos, mas ao mesmo tempo não muito discrepante do valor encontrado pelo método *Elbow*. A Tabela 3 mostra as características extraídas usando k=7 e *feature* (n) = 4.

Tabela 3. Características extraídas pelo *Affinity Propagation* (k=7, n = 4).

Grupos	Características
Cluster 1	Americana, atividade, capitão, controle
Cluster 2	Energia, digit, eletrifio, encontro
Cluster 3	Mundo, partida, jogo, possibilidades
Cluster 4	Operação, possibilidade, política, Brasil
Cluster 5	País, governo, presidente, Brasil
Cluster 5	Período, ministro, estado, 'diária opção'
Cluster 7	Vivo, time, sistema, trabalho

Observando apenas as características, sem analisar as notícias, o *Affinitty Propagation* misturou os termos e não apresentou palavras altamente relacionadas, diferentemente do que aconteceu com o *k-means* que extraiu termos fortemente correlacionados para cada grupo. Outro ponto importante que se percebe ao analisar as características é que nem todas as palavras-chave descrevem o assunto real do referido *cluster*. Assim, seria interessante usar outras técnicas de análise de textos, como por exemplo,

nuvens de palavras ou modelos de tópicos, para extrair informações de cada grupo para assim, melhor representá-lo.

A tabela 4 apresenta um resumo contendo as principais características extraídas para cada *cluster*, a quantidade de notícias que foi alocada em cada grupo, além de mostrar a quantidade de informes por categoria que foi alocada em cada aglomeração.

Tabela 4. Notícias agrupadas (k=4, n = 4).

Grupo	Características extraídas	Quantidade de notícias no grupo	Notícias agrupadas
Cluster 1	Americana, atividade, capitão, controle	13	12 notícias relacionadas à economia 1 notícia relacionada à futebol
Cluster 2	Energia, digit, eletrifio, encontro	10	10 notícias relacionadas à Biologia
Cluster 3	Mundo, partida, jogo, possibilidades	3	3 notícias relacionadas à eletricidade
Cluster 4	Opção, possibilidade, política, Brasil	3	3 notícias relacionadas à eletricidade
Cluster 5	País, governo, presidente, Brasil	4	1 notícia relacionada à economia 2 relacionada à eletricidade
Cluster 6	Período, ministro, estado, 'diária opção'	4	1 notícias relacionadas à biologia 3 notícias relacionadas à eletricidade
Cluster 7	Vivo, time, sistema, trabalho	13	13 notícias relacionadas à futebol
Total		50	

Ao analisar os valores dispostos na Tabela 4, observa-se que o *Affinity Propagation* teve como vantagens o fato de ter agrupado todas as notícias, além de ter formato grupos com documentos semelhantes. Porém, não conseguiu acertar o número de grupos, visto que o corpus era formado por 4 classes e o algoritmo encontrou k=7, subdividindo o grupo eletricidade em 4 subgrupos. Além disso, o algoritmo não apresentou um bom desempenho em relação às características extraídas, visto que muitos dos termos não têm valor significativo em relação às notícias do grupo relacionado e, em contrapartida, outras palavras de peso pertencente a uma determinada classe foram alocadas em outras categorias. Entretanto, acredita-se que, se realizar uma análise das notícias após a coleta e retirar as que foram recuperadas erroneamente, essa técnica de agrupamento teria uma melhor taxa de acerto.

4.3. Teste usando os algoritmos de *Clustering* Hierárquico

Fizeram-se nesta experiência alguns testes usando os algoritmos de agrupamento hierárquico. Ao analisar os grupos formados pelo método *Ward*, nota-se que o primeiro *cluster* foi composto por 4 documentos, sendo cada notícia pertencente a uma classe diferente. Isso significa que esse grupo não foi composto por assuntos semelhantes. Já o segundo *cluster* foi formado por 13 notícias, sendo que quatro delas foram sobre economia, quatro sobre biologia, três sobre eletricidade e duas sobre futebol. O grupo 3 aglomerou 17 notícias pertencentes aos 4 temas. Já o quarto *cluste*r conseguiu agrupar 16 notícias, com maior concentração de matérias relacionadas à economia, mas também com informes pertencentes às demais categorias. Essa dificuldade de agrupamento também aconteceu com os demais algoritmos de *clustering* hierárquico, portanto, essa técnica não é indicada para agrupar esse tipo de documento.

4.4. Validação dos resultados

As análises realizadas nos experimentos anteriores mostram que o algoritmo *k-means* apresentou melhor desempenho para o agrupamento de notícias no idioma português. Assim, realizou-se a avaliação do melhor número de *cluste*r usando esse algoritmo através do coeficiente da silhueta, técnica que mede a qualidade dos agrupamentos. Fez-se o teste com k variando de 2 a 8. O resultado é exibido na Tabela 5.

Tabela 5. Notícias agrupadas (k=4, n = 4)

Nº de clusters	Método de avaliação: coeficiente da silhueta	
	Inicialização com k-Means++	Inicialização randômica
2	0.580	0.580
3	0.625	0.625
4	0.668	0.668
5	0.693	0.693
6	0.701	0.701
7	0.675	0.675
8	0.666	0.682

Como os índices variaram entre 0,5 e 0,7, significa que uma estrutura razoável foi encontrada. O valor maior do coeficiente da silhueta foi para k=6, indicando que esse corpus seria melhor organizado se as notícias fossem subdivididas em seis grupos e não em quatro, conforme era esperado.

Por fim, apesar de cada grupo ter predominância de notícias de um determinado assunto, todos os *clusters* apresentaram algumas matérias que não deveriam estar em tal aglomerado. Por conseguinte, para melhorar a qualidade do agrupamento é necessária, também, uma análise das notícias recuperadas, visto que muitos informes relatam mais de um assunto em seu conteúdo, dificultando, assim, a tarefa de *clustering*.

5. CONSIDERAÇÕES FINAIS

Observou-se durante os experimentos que a etapa de pré-processamento exige um esforço especial para garantir a qualidade dos dados. A complexidade da língua portuguesa, a necessidade de atualização da lista de *stopwords*, a detecção de quais características são mais importantes e, em geral, a complexidade dos problemas relacionados à alta dimensionalidade dos dados foram evidenciados durante todo o processo desta pesquisa.

Os resultados mostraram que a precisão obtida pela técnica de *clustering* está relacionada à qualidade dos dados, ou seja, a falta de características em comum nos textos dificulta a identificação de semelhanças entre as notícias. Assim, é importante que o sistema de recuperação das notícias também seja eficaz na coleta, pois a qualidade da amostra influência diretamente nos resultados.

Nesta pesquisa, realizaram-se alguns experimentos utilizando três métodos de clustering, particionado, propagação por afinidade e hierárquico, para verificar a separabilidade das notícias publicadas no idioma português e coletadas dos principais jornais on-line. Verificou-se que, apesar de alguns *clusters* possuírem homogeneidade entre seus elementos, observou-se, também, que em outros grupos ocorreu uma mistura entre diferentes assuntos. Fato que pode ter ocorrido devido a necessidade de mais amostras para compor as métricas ou por uma seleção de atributos mais abrangente, que possa agregar mais informação aos dados.

O algoritmo *k-means* obtive os melhores resultados, pois considerando a quantidade de notícias agrupadas, ele conseguiu uma taxa acima de 90% de acerto e agrupou 84% dos textos. Se considerar as particularidades das notícias, esse algoritmo obteve praticamente 100% de acerto, ponderando apenas as matérias que foram congregadas. O *Hierarchical Clustering* apresentou dificuldades, visto que notícias semelhantes foram alocadas em grupos diferentes, portanto, essa técnica não é indicada para agrupar esse tipo de documento. Já o algoritmo *Affinity Propagation* teve dificuldades na ex-

tração das características e apresentou divergência quanto ao número ideal de *clusters*, sendo um resultado que deve ser levado em consideração em trabalhos futuros para verificar a separabilidade entre outras amostras de textos, pois, essa técnica apresentou bom desempenho ao agrupar por semelhança, porém subdividiu a coleção em vários grupos, ou seja, um número maior do que o valor de k encontrado pelo método *Elbow* e pelo Coeficiente da Silhueta.

O tamanho da amostra também influência nos resultados e quanto maior for o número de grupos, mais forte é a estrutura encontrada intragrupo. Além disso, os algoritmos conseguem melhor desempenho quanto mais diversificado for o corpus e quanto mais bem definidas forem as características dos textos. Portanto, para melhorar a qualidade do agrupamento, é necessária uma análise das notícias recuperadas, visto que muitos informes relatam mais de um assunto em seu conteúdo, dificultando, assim, a tarefa de *clustering*. Mas, ao mesmo tempo, a possibilidade de descobrir agrupamentos não óbvio é a principal vantagem desta técnica.

REFERÊNCIAS

ARANHA, C.; PASSOS, E. A Tecnologia de Mineração de Textos. **RESI-Revista Eletrônica de Sistemas de Informação**, v. 2, p. 1–8, 2006. Disponível em: http://www.periodicosibepes.org.br/index.php/reinfo/article/download/171/66. Acesso em: 30 set. 2019.

BAEZA-YATES, R. RIBEIRO-NETO, B. **Recuperação de Informação**: Conceitos e tecnologia das máquinas de busca. Tradução técnica: Leandro Krug Wives, Viviane Pereira Moreira. 2. ed. Porto Alegre: Bookman, 2013.

DAS. Formação Cientista de Dados. **Curso de Machine Learning**. Data Science Academy. E-book. 2017.

EBECKEN, N. F. F.; LOPES, M. C. S.; COSTA, M. C. de A. Mineração de Textos. In: REZENDE, S. O. **Sistemas Inteligentes**: Fundamentos e Aplicações, 1. ed. São Paulo: Manole, 2003, cap.13, p. 337-370.

FAYYAD, U. M. *et al*. From data mining to knowledge discovery: an overview. *In*: **Advances in knowledge discovery and data mining**. California: AAAI/The MIT, 1996. p.1-34

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data mining**: conceitos, técnicas, algoritmos, orientações e aplicações. Rio de Janeiro: Elsevier, 2015.

GONÇALVES, M. **Classificação de Textos**. *In*: BAEZA-YATES, R. RIBEIRO-NETO, B. Recuperação de Informação: Conceitos e tecnologia das máquinas de busca. Tradução técnica: Leandro Krug Wives, Viviane Pereira Moreira. 2. ed. Porto Alegre: Bookman, 2013. p. 277-338.

HAIR, J. F. et al. Análise multivariada de dados. Trad. Adonai S. Sant'Anna e Anselmo C. Neto. 5 ed. Porto Alegre: Bookman, 2005.

LADEIRA, A. P. **Processamento de Linguagem Natural: Caracterização da produção científica dos pesquisadores brasileiros**. 2010. 159 f. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais. Disponível em: https://repositorio.ufmg.br/bitstream/1843/ECID-8B3Q6C/1/tese_anapaulaladeira_cd.pdf. Acesso em 25 nov. 2019.

MATOS, D. **Conceitos Fundamentais de Machine Learning**. Disponível em: http://www.cienciaedados.com/conceitos-fundamentais-de-machine-learning/. Acesso em: 20 nov. 2019. http://www.cienciaedados.com/conceitos-fundamentais-de-machine-learning

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997.

MÜLLER, A. C. e GUIDO, S. **Introduction to Machine Learning with Python**. O'Reilly Media, 2017.

CHEESEMAN, P. & J. STUTZ. Bayesian Classification (Auto Class): Theory and Results. In: FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (Eds.), **Advances in Knowledge Discovery and Data Mining.** American Association for Artificial Intelligence. Menlo Park, CA. 1996, p. 153-180

SALTON, G; MCGILL, M. J. **Introduction to Modern Information Retrieval**. John Wiley and Sons, New York, 1983.

ZHAI, C. X.; MASSUNG, S. **Text Data Management and Analysis**: A Practical Introduction to Information Retrieval and Text Mining. Association for Computing Machinery and Morgan, Claypool, New York, NY, USA, 2016.

ZURINI, M.; SBORA, C. Clustering Analysis within Text Classification Techniques. **Informática Economica**. v. 15, n. 4, p. 178-189, 2011. Disponível em: http://revistaie.ase.ro/content/60/14%20-%20Zurini,%20Sbora.pdf. Acesso em: 25 nov. 2019.

Lúcia Helena de Magalhães

Mestre em Ciências em Engenharia Civil com área de concentração em Sistemas Computacionais pela Universidade Federal do Rio de Janeiro. Doutorado em andamento em Ciência da Informação pela UFMG. Professora no Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais.

Renato Rocha Souza

Doutor em Ciência da Informação pela Universidade Federal de Minas Gerais (2005) e pós-doutorado (2010) em Tecnologias Semânticas para Recuperação de Informação - University of South Wales, UK. É atualmente professor e pesquisador da Escola de Matemática Aplicada (EMAp) da Fundação Getúlio Vargas e professor colaborador da Escola de Ciência da Informação da Universidade Federal de Minas Ge-

rais. É Visiting Fellow da University of South Wales (2009-2019) e Pesquisador Adjunto Sênior da Universidade de Colúmbia.