Agrupamento de Tweets

Augusto Krejci Bem-Haja
RA 227017
augustohaja@gmail.com

Lucas Vieira de Miranda
RA 211499
lucasvieirademiranda@gmail.com

I. Introdução

Este trabalho tem como objetivo descobrir estruturas organizacionais do conjunto de dados chamado *Health News in Twitter* através de técnicas de aprendizado não supervisionado.

O *Health News in Twitter* é um conjunto de dados sobre saúde fornecido pela *University of California Irwine* (UCI), e contém dados de tweets coletados no ano de 2015. Neste trabalho utiliza-se apenas uma parte do conjunto que contém 13299 tweets referentes as agências de notícias *BBC*, *CNN*, *Fox News* e *Everyday Health*.

Os dados são encontrados em três arquivos health.txt, bags.csv e word2vec.csv. O primeiro arquivo contém as sentenças encontradas nos tweets, o segundo arquivo contém o *bag of words* das sentenças e o terceiro arquivo contém *word2vec* das sentenças.

II. Análise dos Dados

Inicialmente, buscou-se entender alguns padrões presentes nos dados de *tweets* para facilitar a análise dos agrupamentos, para isso, optou-se por ler os tweets armazenados no arquivo health.txt e realizar um pré-processamento sobre o texto.

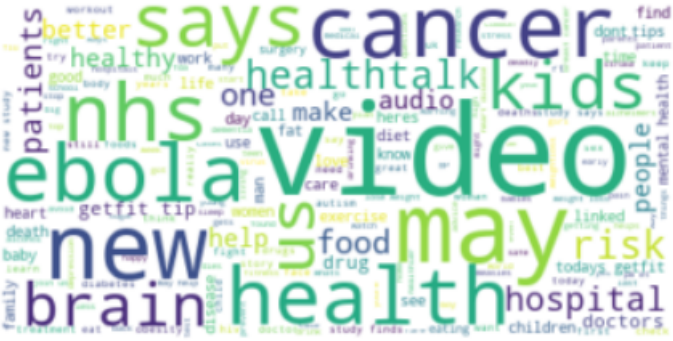
Durante o pré-processamento, foram removidos as ocorrências das palavras iniciadas com @ (referências), as palavras RT (retweet) e caracteres especiais. Em seguida, removeu-se as palavras com menos significado (is, isn't, etc) e extraiu-se a raiz das palavras para agrupar palavras parecidas.

Por fim, calculou-se o histograma de frequência das palavras e plotou-se uma nuvem de palavras contendo as 200 mais frequentes, como pode ser visto na figura 1. As 10 palavras mais frequêntes encontradas foram: video, healthtalk, health, ebola, may, cancer, new, us, says e study.

III. Representação dos Dados

Os algoritmos de aprendizado não supervisionado necessitam de dados numéricos para funcionar [9], dessa forma, é necessário converter os textos em números, para isso existem duas abordagens muito conhecidas que são o *bag of words* e *word2vec*.

Figure 1. Núvem de Palavras.



Fonte: produzida pelos autores

O *bag of words* é um vetor com tamanho igual ao número de palavras existentes no corpo de texto e o valor de cada elemento indica quantas vezes a palavra ocorre naquela sentença [13].

O *word2vec* é um vetor com tamanho variável proveniente da camada oculta de uma rede neural treinada para encontrar a relação entre palavras. Uma variação do *word2vec* utilizada para codificar sentenças se chama *doc2vec*, e é utilizada para encontrar a relação entre sentenças [14].

IV. Métodos de Avaliação para Agrupamentos

Existem diversas métricas de avaliação para agrupamentos e elas podem ser divididas em dois grupos, que são: métricas de avaliação para dados anotados e métricas de avaliação para dados não anotados [8].

Devido ao fato da nossa base de dados não possuir dados anotados, utilizou-se o segundo grupo de métricas do qual fazem parte o Coeficiente de Silhueta, Índice de Calinski-Harabaz e o Índice de Davies-Bouldin.

O Coeficiente de Silhueta é uma métrica que varia entre -1 e 1, no qual os valores próximos de -1 representam agrupamentos ruins e valores próximos de 1 representam bons agrupamentos, já os valores próximos de 0 representam agrupamentos sobrepostos. A principal desvantagem do Coeficiente de Silhueta é o tempo necessário para calculá-lo [1] [10].

O Índice de Calinski-Harabaz é uma métrica que não possui um intervalo definido, entretanto, sabe-se que quanto maior o valor melhores são os agrupamentos. A principal vantagem do Índice de Calinski-Harabaz é o fato de ser muito rápido para se calcular [2].

O Índice de Davies-Bouldin é uma métrica que não possui um intervalo definido, entretanto, sabe-se que quanto mais próximo de zero o valor, melhores são os agrupamentos. A principal vantagem do Índice de Davies-Bouldin é o fato de ser muito rápido para se calcular e a principal desvantagem é que nem sempre bons valores do índice representam bons agrupamentos [3].

V. Experimentos com Mean Shift

O algoritmo *Mean Shift* consiste em definir uma janela deslizante (ou *kernel*) ao redor de cada amostra, calcular a média na região da janela, mover a janela em direção a média e continuar até que a variação da média torne-se infima. Dessa forma, no final das iterações as janelas terão se movido para as regiões mais densas do conjunto de dados, em seguida, executa-se um pós processamento sobre as médias eliminando os valores muito próximos para encontrar os grupos [4].

Relizaram-se então, dois experimentos utilizando o algoritmo *Mean Shift* um utilizando o arquivo bags.csv (*bag of words*) e o outro utilizando word2vec.csv (*word2vec*).

O primeiro experimento consistiu em utilizar o arquivo bags.csv e o algoritmo *Mean Shift* e obteve-se apenas um agrupamento contendo todos os tweets.

O segundo experimento consistiu em utilizar o arquivo word2vec.csv e o algoritmo *Mean Shift* e obteve-se, também, apenas um agrupamento contendo todos os tweets.

Analisando-se os resultados imagina-se que o fato de apenas um agrupamento ser formado se explique pelo fato dos dados estarem agrupados de forma muito densa, uma vez que todos são referentes a saúde.

VI. Experimentos com Affinity Propagation

O algoritmo *Affinity Propagation* consiste no cálculo de duas métricas chamadas de responsabilidade e disponibilidade para determinar como os dados devem ser agrupados. A responsabilidade indica o quanto uma amostra *k* é um exemplo para uma amostra *i*, já a disponibilidade indica o quanto uma amostra *i* deve eleger *k* como exemplo. Dessa forma, as amostras elegem como exemplo a amostra que é mais similar a elas, e essa amostra se torna a representante do grupo e as demais são agrupadas ao redor dela [5].

Realizaram-se dois experimentos utilizando o algoritmo *Affinity Propagation* um utilizando o arquivo bags.csv (*bag of words*) e outro utilizando o arquivo word2vec.csv (*word2vec*).

O primeiro experimento consistiu em utilizar o arquivo bags.csv e o algoritmo *Affinity Propagation* com os parâmetros *damping* com valor 0.5 (default do *sklearn*) e *max_iteration* com um valor de 300. Obtendo-se 1293 grupos, e os valores 0.08, 7.34 e 2.49 para o Coeficiente de Silhueta, Índice de Calinski-Harabaz e Índice de Davies-Bouldin.

O segundo experimento consistiu em utilizar o arquivo word2vec.csv e o algoritmo *Affinity Propagation* com os parâmetros *damping* com valor 0.5 (default do *sklearn*) e *max_iteration* com um valor de 300. Obtendo-se 1139 grupos, e os valores 0.02, 6.19 e 2.49 para o Coeficiente de Silhueta, Índice de Calinski-Harabaz e Índice de Davies-Bouldin.

Analisando-se apenas os valores das métricas não foi possível chegar a uma boa conclusão, pois os valores obtidos pelo Coeficiente de Silhueta foram baixos e não há uma referência para o Índice de Calinski-Harabaz e o Índice de Davies-Bouldin.

Optou-se então por analisar os agrupamentos formados e observou-se que ao utilizar o bag of words as frases são agrupadas pelas palavras chaves mais frequentes nos tweets, já utilizando word2vec as frases são agrupadas considerando palavras mais frequentes e as palavras relacionadas, mas em alguns casos, a relação não é tão evidente.

De modo geral, os agrupamentos encontrados possuem sentido, mas a quantidade é grande demais.

VII. Experimentos com K-Means

O algoritmo *K-Means* consiste em definir um número de pontos (*K*), inicializá-los de forma aleatória (existem outros meios, como *K-Means++*), calcular a distância entre as amostras e cada ponto, agrupar as amostras mais próximas de cada ponto, calcular a média entre as amostras e transformá-las nos novos pontos, e assim, sucessivamente calculando o erro para cada iteração até que o erro pare de variar. O maior problema encontrado ao utilizar o *K-Means* é descobrir o melhor valor de *K*, com o objetivo de se obter o melhor agrupamento possível [6] [9].

A primeira tentativa para descobrir um bom intervalo de valores para *K* consistiu na utilização da técnica conhecida como curva do cotovelo. A curva do cotovelo pode ser obtida através da execução de apenas uma iteração do *K-Means* para cada valor de *K*, e então para cada *K*, calcula-se o erro e plota-se um gráfico do erro x *K*, entretanto, nesse caso as curvas obtida não possuiam a forma de um cotovelo e o método se mostrou ineficiente.

A segunda tentativa para descobrir um bom intervalo de valores para *K* consistiu em dois experimentos nos quais um utilizou o arquivo bags.csv e outro utilizou word2vec.csv, também, utilizou-se os índices de Calinski-Harabaz e Davies-Bouldin devido a velocidade de execução em relação ao Coeficiente de Silhueta.

O primeiro experimento consistiu em executar o *K-Means* utilizando os dados do bags.csv, a inicialização *K-Means++*, variando o *K* entre 2 e 13000, incrementar *K* de 250 em 250, calcular os índices de Calinski-Harabaz e Davies-Bouldin.

O segundo experimento consistiu em executar o *K-Means* utilizando os dados do word2vec.csv, a inicialização *K-Means++* , variando o *K* entre 2 e 13000, incrementar *K* de 250 em 250, calcular os índices de Calinski-Harabaz e Davies-Bouldin.

Analisando-se os resultados obtidos e considerando a natureza das métricas, optou-se por utilizar o índice de Calinski-Harabaz como referência e observou-se que no intervalo de *K* entre 2 e 1500 os valores obtidos estavam entre 86.56 e 8.68 para bags.csv e 403.86 e 6.27 para word2vec.csv, ou seja, como essa faixa apresenta os maiores valores para os índices, deve também, possuir os melhores agrupamentos.

A terceira tentativa para descobrir um bom intervalo de valores para *K* consistiu em utilizar a informação descoberta no experimento anterior e realizar dois novos experimentos nos quais um utilizou o arquivo bags.csv e o outro utilizou o arquivo word2vec.csv.

O primeiro experimento consistiu em executar o *K-Means* utilizando os dados do bags.csv, a inicialização *K-Means++*, utilizar os valores 250, 500, 750, 1000, 1250 e 1500 para *K*, verificar a quantidade de itens em cada agrupamento e a relação entre os *tweets* nos mesmos.

O segundo experimento consistiu em executar o *K-Means* utilizandos os dados do word2vec.csv, a inicialização *K-Means++*, utilizar os valores 250, 500, 750, 1000, 1250 e 1500 para *K*, verificar a quantidade de itens em cada agrupamento e a relação entre os *tweets* nos mesmos.

Novamente, ao analisar os agrupamentos formados observou-se que ao utilizar *bag of words* as frases são agrupadas pelas palavras chaves mais frequentes nos *tweets*, já utilizando o *word2vec* as frases são agrupadas considerando as palavras mais frequentes e as palavras relacionadas a elas, mas em alguns casos, a relação não é evidente.

Além disso, observou-se que conforme o valor de *K* aumenta entre 2 a 1500 os agrupamentos se tornavam cada vez menores e mais específicos. Por fim, observou-se que era possível obter bons agrupamentos com valores de *K* variando entre 250 e 1500, como esperado através da análise do Índice de Calinski-Harabaz.

VIII. Experimento PCA + K-Means

Após analisar os experimentos realizados, optou-se por utilizar o arquivo bags.csv, ou seja, a representação de dados bag of words, e também, o algoritmo *K-Means* com $K = 500$ e a inicialização *K-Means++*.

Continuando, aplicou-se o algoritmo *Principal Component Analysis* (*PCA*) e testaram-se algumas variâncias, como por exemplo: 0.85, 0.9 e 0.95.

Analisando os resultados obtidos para as métricas (Coeficiente de Silhueta, Índice de Calinski-Harabaz e Davies-Bouldin), sabendo que o intervalo da variância está definido entre 0 e 1 e que quanto mais próximo de 1 menos informação é perdida. Optou-se por utilizar uma variância de 0.9 e obteve-se uma redução de 1203 para 904 features.

Em seguida, gerou-se os agrupamentos, calculou-se o Coeficiente de Silhueta, o Índice de Calinski-Harabaz e o Índice de Davies-Bouldin e comparou-se os resultados que indicaram um desempenho levemente superior ao utilizar *PCA* em relação a não utilização do mesmo, o resultado pode ser observado na tabela I.

Por fim, analisou-se os agrupamentos obtidos, e observou-se que eles eram melhor distribuídos e mais coesos do que os obtidos sem a utilização do *PCA*.

Table I
Métricas

	Sem PCA	**Com PCA**
Coeficiente de Silhueta	0.084	0.084
Índice de Calinski-Harabaz	14.423	14.865
Índice de Davies-Bouldin	2.369	2.439

IX. Conclusões

O maior problema existente nos algoritmos de aprendizado não supervisionado ou algoritmos de agrupamento é a dificuldade em se encontrar os melhores agrupamentos possíveis. O sucesso nessa tarefa depende de uma análise prévia dos dados, uma boa representação da informação, uma boa avaliação utilizando as métricas existentes e a análise dos agrupamentos gerados pelo algoritmo.

A utilização do algoritmo *Principal Component Analysis* (*PCA*) além de reduzirem o número de features podem melhorar melhorar levemente o desempenho dos algoritmos de agrupamento, entretanto, é necessário estar atento a variância escolhida, porque ela pode impactar negativamente nos resultados.

REFERÊNCIAS

[1] ScikitLearn. 2018. Disponível em: <http://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient>. Acesso em 05/11/2018.

[2] ScikitLearn. 2018. Disponível em: <http://scikit-learn.org/stable/modules/clustering.html#calinski-harabaz-index>. Acesso em 05/11/2018.

[3] ScikitLearn. 2018. Disponível em: <http://scikit-learn.org/stable/modules/clustering.html#davies-bouldin-index>. Acesso em 05/11/2018.

[4] ScikitLearn. 2018. Disponível em: <http://scikit-learn.org/stable/modules/clustering.html#mean-shift>. Acesso em 05/11/2018.

[5] ScikitLearn. 2018. Disponível em: <http://scikit-learn.org/stable/modules/clustering.html#affinity-propagation>. Acesso em 05/11/2018.

[6] ScikitLearn. 2018. Disponível em: <http://scikit-learn.org/stable/modules/clustering.html#k-means>. Acesso em 05/11/2018.

[7] ScikitLearn. 2018. Disponível em: <http://scikit-learn.org/stable/modules/decomposition.html#pca>. Acesso em 05/11/2018.

[8] ScikitLearn. 2018. Disponível em: <http://scikit-learn.org/stable/modules/classes.html#clustering-metrics>. Acesso em 05/11/2018.

[9] Avila, Sandra. K-Means. 18 de setembro de 2018. Notas de Aula.

[10] Avila, Sandra. Técnicas de Agrupamento. 20 de setembro de 2018. Notas de Aula.

[11] Avila, Sandra. Artificial Neural Networks. 25 de setembro de 2018. Notas de Aula.

[12] Avila, Sandra. Artificial Neural Networks. 27 de setembro de 2018. Notas de Aula.

[13] Brownlee Jason (United States). “A Gentle Introduction to the Bag of Words Model“. Disponível em: <https://machinelearningmastery.com/exploding-gradients-in-neural-networks/>. Acesso em: 05/11/2018.

[14] TensorFlow. “Vector Representations of Words“. Disponível em:<https://www.tensorflow.org/tutorials/representation/word2vec>. Acesso em: 05/11/2018.