Agrupando títulos de notícias de jornal em grupos similares com o k-means

Gustavo Ciotto Pinton, RA117136*
Lucas Rath Maia, RA117751†

Abstract

Neste documento, são propostos dois métodos capazes de agrupar títulos de jornais em categorias de acordo com o seu tema. Ambos os métodos utilizam a técnica de clusterização denominada de k-means, porém a maneira com que os atributos foram gerados variou. O primeiro método utilizou a técnica de bag-of-words aliada ao cálculo da métrica TFIDF sobre word n-grams, com n indo de 1 a 5, enquanto que o segundo baseou-se no uso de matrizes previamente calculadas, chamadas de word embeddings. No total, foram disponibilizados um milhão de títulos publicados entre 2003 e 2017 do jornal australiano ABC (Australian Broadcasting Corporation).

1. Introdução

Técnicas de agrupamento de elementos em grupos com características semelhantes possuem cada vez mais interesse nos dias de hoje com a ampla utilização das redes sociais. Elas permitem, por exemplo, que propagandas e campanhas sejam direcionadas a determinados grupos de maneira que o efeito sobre os elementos que o compõem seja potencializado. Outro exemplo do grande poder deste tipo de técnica foi o recente caso da empresa *Cambridge Analytica* que utilizou um grande volume de perfis da rede social Facebook para categorizar indivíduos em grupos com orientações políticas semelhantes com o objetivo de influenciar eleições de diversos países, incluindo a dos Estados Unidos [1].

Tendo em vista tal capacidade, propomos neste documento a clusterização de títulos publicados entre 2003 e 2017 do jornal australiano ABC (*Australian Broadcasting Corporation*), a partir da técnica de aprendizado não supervisionado denominada *k-means*. Tal técnica consiste no cálculo de k centroídes, de maneira a otimizar as distâncias de cada amostra a seu determinado centroíde. Assim como qualquer método de aprendizado de máquina, a otimização implica a minimização de uma função de custo J, conforme representado na equação 1, logo abaixo, em que M é a quantidade de amostras disponíveis, $\boldsymbol{c}^{(i)}$ é o índice do centróide ao qual a amostra $\boldsymbol{x}_i$ pertence e $\boldsymbol{\mu}_j$ é o centróide k, com $\boldsymbol{x}^i$ e $\boldsymbol{\mu}_j \in \mathbb{R}^n$. Por fim, a quantidade de centróides a ser utilizada é obtida a partir de um método tal como o *elbow method*, de modo que o k escolhido fique entre uma região de distorção estabilizada e outra em que a mesma métrica cai rapidamente.

$$J(\boldsymbol{c}^{(1)}, ..., \boldsymbol{c}^{(m)}, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k) = \frac{1}{M} \sum_{i=1}^{M} \|\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{c}^{(i)}}\|^2 \quad (1)$$

Antes de utilizarmos o *k-means*, é necessário obtermos atributos pertinentes e que descrevam corretamente cada um dos títulos das notícias, isto é, necessita-se traduzir uma frase em um vetor $\boldsymbol{x}_i$ numérico aplicável ao *k-means*. Neste documento, serão apresentados dois mecanismos de obtenção destes atributos. O primeiro é baseado na técnica de *bag-of-words* [2] aliada ao cálculo da métrica *TFIDF* sobre *word n-grams* gerados a partir de cada título. Neste caso, separa-se os títulos em *tokens* e, de acordo com n, inclui-se na *bag-of-words* toda combinação sequencial contendo n destes *tokens*. Por exemplo, para a frase *I love cat* e $n = 2$, consideraríamos *I love* e *love cat*. Uma vez inseridas todas as combinações na *bag-of-words*, a próxima etapa, para cada título, é calcular a frequência (ou *term frequency*) que cada combinação aparece neste determinado título multiplicada pelo fator *inverse document frequency*, que avalia a importância desta combinação entre todos os títulos. Em outras palavras, dá-se menos importância a *n-grams* que aparecem em muitos títulos pelo simples fato de que eles não acrescentam informação nova e específica, não sendo, portanto, importantes para a diferenciação e agrupamento dos respectivos títulos. Palavras como *the*, por exemplo, podem ter uma frequência alta em um título, porém também podem receber um fator *IDF* extremamente baixo, tendo em vista que aparecem em muitas amostras distintas. A equação 2 representa uma maneira para calcular a métrica *TFIDF*, em que $f_{t,d}$ é a frequência do termo (ou *n-gram*) t no título d, M é o número total de amostras e m_t é o número de amostras em que t aparece.

$$\text{TFIDF}(t) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \log\left(\frac{M}{m_t}\right) \quad (2)$$

Como o leitor deve estar imaginando, a matriz de atributos gerada a partir do método descrito acima é extremamente esparsa, isto é, preenchida em sua quase totalidade por 0. Tendo em vista que o *k-means* não é muito eficiente para este tipo de estrutura de dados, precisamos de algo que reduza a dimen-

*Is with the Institute of Computing, University of Campinas (Unicamp). **Contact**: gustavociotto@gmail.com
†Is with the Institute of Computing, University of Campinas (Unicamp). **Contact**: lucasrm25@gmail.com

sionalidade dos vetores de atributos. Nesse contexto, utiliza-se ou o PCA ou o SVD para tal redução. Ambos utilizam o cálculo de autovetores e autovalores sobre a matriz de autocovariância com o intuito de calcular as n direções de maior variância, porém o segundo, ao contrário do primeiro, não centraliza os dados antes de computar a decomposição dos valores singulares, sendo muito mais eficiente para matrizes esparsas. Neste relatório, será utilizada, portanto, a técnica de SVD.

A segunda técnica de processamento de atributos é denominada *word embeddings*. Neste caso, uma matriz $\boldsymbol{W}_{P\times D}$ relaciona uma determinada palavra $i \in [1, P]$ a um vetor de dimensão D da ordem de centenas de elementos. Multiplica-se $\boldsymbol{W}$ com a matriz $\boldsymbol{F}_{M\times P}$ em que cada linha contém P elementos, em que cada um deles representa a quantidade de ocorrências da palavra i no respectivo título. Obtém-se, assim, a nova matriz $\boldsymbol{V}_{M\times D}$ contendo os atributos a serem utilizados no *k-means*. Neste documento, a matriz $\boldsymbol{W}$ foi gerada a partir de um algoritmo chamado *GloVe* [3] e dados retirados da *Wikipedia* no ano de 2014. Uma característica interessante da matriz gerada por esse algoritmo é que palavras com contextos correlacionados terão vetores apontando para direções semelhantes. Por exemplo, *king* e *queen* têm vetores parecidos. Além disso, verifica-se a relação vetorial *king - queen = man - woman*. Em suma, o próprio algoritmo se encarrega de adicionar o significado semântico de palavras correlacionadas, possibilitando uma clusterização eficiente.

É importante ressaltar que antes dos dois processamentos apresentados, transformam-se eventuais letras maiúsculas em minúsculas e retiram-se dos documentos todas *stop words*, isto é, palavras que não acrescentam nenhuma informação adicional tal como artigos e preposições. Adicionalmente, optamos por reduzir do vocabulário palavras com frequência acima de uma porcentagem determinada a partir de testes. Isso foi realizado para evitar que os métodos gerassem clusters viciados em apenas palavras iguais e não significados.

Por fim, por limitações computacionais, resolvemos utilizar apenas uma fração dos dados, isto é, apenas o ano de 2017, para o agrupamento. Assume-se, portanto, que os temas não variam muito de um ano para o outro.

2. Atividades

As próximas subseções visam explicar as escolhas dos diversos parâmetros adotados pelo autor.

2.1. Cálculo dos atributos com bag-of-words

Conforme discutido na seção **Introdução**, utilizamos apenas *word n-grams* para o cálculo da métrica *TFIDF*. Após testes, determinou-se que os melhores resultados foram para n no intervalo [1,5]. A figura 1 nos ajuda a entender um pouco o porquê deste resultado: a maioria dos títulos contém entre 4 e 7 palavras, sendo assim, *n-grams* com n mais elevados nos ajudam a representar melhor o significado de cada sentença e impedem que os *clusters* se viciem em apenas palavras semelhantes. Para esta configuração e ano de 2017, os 44182 vetores de atributos obtidos possuíram 5885 elementos. Também a partir de testes, escolhemos reduzir tal dimensionalidade para

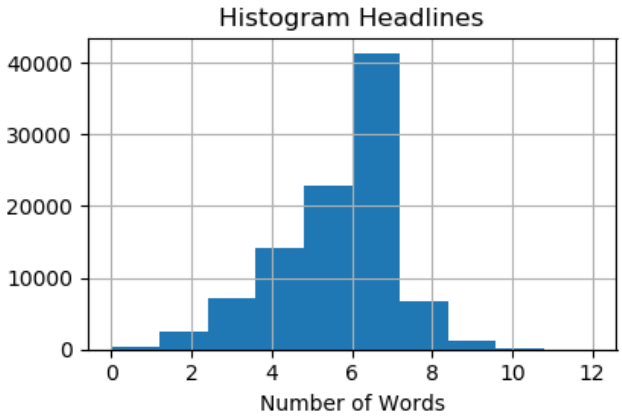

Figura 1. Histograma do número de palavras nos títulos das notícias para o ano de 2017. Pré-processamento (eliminação de *stop words*, letras maiúsculas) já realizado.

50 utilizando a técnica de *Single Value Decomposition* (SVD), conforme discutido na seção anterior. Para os dois passos, empregamos bibliotecas já implementadas no módulo *python sklearn* [4].

2.2. Cálculo dos atributos com word embeddings

A partir do algoritmo GloVe [3], escolhemos o *dataset* com 400 mil palavras no vocabulário e vetores de 300 elementos. Neste caso, evidentemente, utilizamos apenas $n = 1$ para o cálculo dos *word n-grams* e obtivemos vetores de atributos com 300 elementos para uso direto no *k-means*.

3. Soluções propostas

Essa seção é dedicada à discussão das soluções propostas.

3.1. Bag-of-words e SVD

A primeira etapa é determinar o número de *clusters* a ser utilizados. Para tal, realizamos a processamento para k de 2 até 80 e, a partir do *elbow method*, determinamos experimentalmente qual deles é a melhor opção. A figura 2 possui os resultados das distorções, definidas como a soma das distâncias ao quadrado entre cada amostra e o seu centroíde dominante, para todo o intervalo sugerido. Verifica-se que, apesar da grande variação, o *cotovelo* é encontrado para $k = 40$, ponto em que a soma das distâncias estabilizam-se em torno de aproximadamente 1375 (salvo apenas algumas exceções como $k = 50$ e k inferiores a $\sim$70). Ressalta-se que para todos os cálculos, o número de iterações máximos do *k-means* foi 1000.

Tendo escolhido o valor de k, podemos utilizar uma ferramenta de visualização para verificar se os *clusters* foram bem definidos. Uma alternativa é o **t-SNE**, que consiste em uma técnica de redução de dimensionalidade adequada para a visualização de *datasets* com muitas dimensões. A figura 3 representa o resultado da clusterização para $k = 40$ para apenas 2000 amostras (tal limitação foi devida aos recursos computacionais disponíveis). Observa-se a presença de grupos muito bem separados e definidos nas partes mais externas do

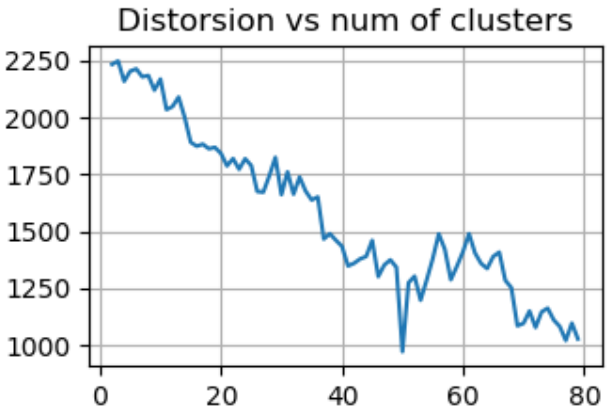

Figura 2. Distorção para cada valor de k variando de 2 até 80 para atécnicas de *bag-of-words* e SVD.

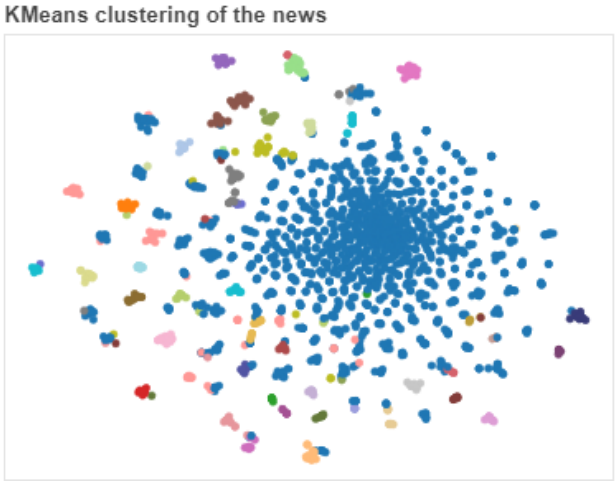

Figura 3. Representação obtida a partir da técnica de t-SNE para 40 *clusters*.

gráfico e um grupo predominante no centro. Este último possui títulos, ao contrário dos demais, com temas não muito definidos, sendo considerados pelo autor deste documento como **Geral**. Abaixo, são apresentadas as tuplas (W, Q) para alguns dos *clusters* obtidos neste método, em que W é um *word n-gram* e Q é a quantidade de vezes que W aparece. Tal listagem ocorre em ordem decrescente em relação a Q e considera apenas os cinco maiores Q. Além disso, são adicionadas três títulos escolhidos randomicamente no respectivo *cluster* e uma breve explicação sobre o seu tema.

- **Cluster 1**: 31831 títulos no total. (house, 292), (health, 280), (school, 270), (woman, 268), (labor, 265). Frases: (i) *political donations in queensland to be revealed within 7 days*, (ii) *should tourists boycott myanmar again*, (iii) *nsw hot weather boils with penrith record*. As frases neste *cluster* não possuem nenhum tema específico em comum, sendo, deste modo, classificados no tema **Geral**.

- **Cluster 3**: 199 títulos no total. (scorecentre, 159), (nrl, 127), (afl, 75), (raiders, 29), (broncos, 29), (bulldogs, 26). Frases: (i) *nrl scorecentre warriors broncos sharks bulldogs*, (ii) *nrl top five: april 3*, (iii) *nrl scorecentre*. O assunto deste *cluster* é facilmente identificado como **Liga de Rugby Australiana**, contendo títulos com a abreviação *nrl* e nomes de diversos times.

- **Cluster 5**: 638 títulos no total. (murder, 360), (charged, 353), (alleged, 47), (woman, 36), (trial, 33), (stabbing, 31). Frases: (i) *rockhampton man ian coombe charged with fraud*, (ii) *former teacher charged over alleged historical sex assaults*, (iii) *borce ristevski faces court charged with murdering wife karen*. Este *cluster* agrupa títulos relacionados a **Crimes e assassinatos**.

- **Cluster 20**: 127 títulos no total. (violence, 127), (domestic, 110), (victims, 16), (demand, 6), (help, 6). Frases: (i) *russian women hide domestic violence scars with tattoos*, (ii) *candlelight vigil in hobart for domestic violence victims*, (iii) *catalonia how did it come to violence in the streets*. Este *cluster* é dominado pelas palavras *violence* e *domestic* que aparecem muito mais que as demais. Portanto, o tema do grupo é **Violência doméstica**.

- **Cluster 21**: 323 títulos no total. (marriage, 325), (vote, 40), (survey, 37), (bill, 25), (postal, 22), (gay, 15). Frases: (i) *margaret court marriage bible isnt meant to be read so literally*, (ii) *the same sex marriage debate*, (iii) *same sex marriage bill debate moves to amendments*. O tema predominante neste *cluster* é **Casamento de pessoas do mesmo sexo**, apesar de que apenas a palavra *marriage* aparece muito mais que as demais. Nota-se também que *marriage* ocorre mais de uma vez em um mesmo título, tendo em vista que suas aparições superam o próprio número de títulos no *cluster*.

- **Cluster 30**: 406 títulos no total. (crash, 321), (killed, 71), (fatal, 68), (plane, 63), (dead, 56), (car, 53), (driver, 49). Frases: (i) *uber suspends self driving car program after arizona crash*, (i) *plane crash on swan river during australia day*, (iii) *police searching for driver of truck that hit sydney home*. A principal temática deste grupo, conforme palavras e frases, é **Acidente em meios de transporte**.

- **Cluster 38**: 179 títulos no total. (media, 163), (social, 115), (video, 10), (campaign, 8), (facebook, 8). Frases: (i) *film media union to look at safety after bliss n eso shooting*, (ii) *pepsi pulls kendall jenner ad amid social media outcry*, (iii) *formula one relaxes social media rules for teams*. Neste caso, notícias com o tema **Rede social** são agrupadas no *cluster*.

3.2. Word embeddings

Assim como na solução anterior, é necessário selecionar a quantidade k de *clusters* a ser utilizados com auxílio do *elbow method*. Neste caso, temos vetores de atributos com 300 elementos gerados pela multiplicação da matriz de entrada por uma outra pré-processada a partir de dados do algoritmo GloVe, conforme discutido na seção **Introdução**. A

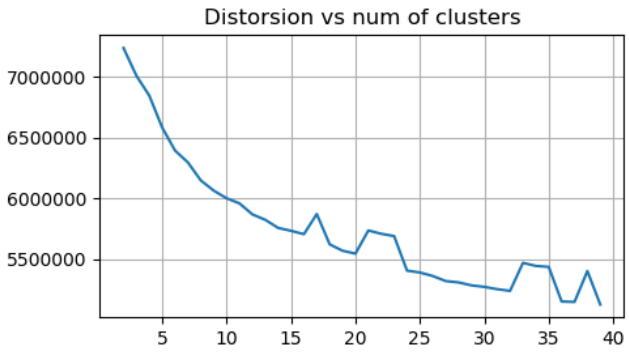

Figura 4. Distorção para cada valor de k variando de 2 até 80 para a técnica de *word embedding*.

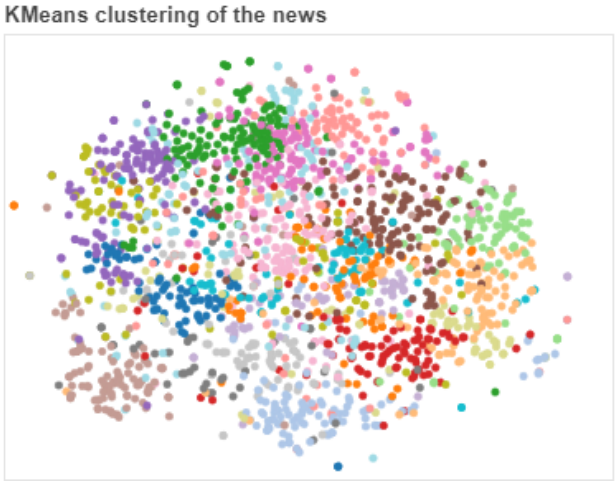

Figura 5. Representação obtida a partir da técnica de t-SNE para 20 *clusters* com *features* calculados a partir dos *word embeddings*.

figura 4 representa o resultado das distorções calculadas para k no intervalo de 2 a 80. Neste caso, escolhemos, $k = 20$ para o número de centróides. Destaca-se que a distorção deste método é muito superior àquela calculada na solução anterior, sendo aproximadamente $\frac{5.500.000}{1375} = 4000$ maior. A figura 5, por sua vez, possui o resultado da representação visual gerada pelo algortimo **t-SNE**. Observa-se que nenhum *cluster* é visualmente bem definido, isto é, os dados estão dispostos em uma grande nuvem sem nenhuma separação aparente. Tal fato explica a grande diferença de distorção encontrado neste caso em relação ao anterior.

Assim como na subseção anterior, destacam-se alguns *clusters* obtidos a partir do método de *word embeddings*:

- **Cluster 3**: 1871 títulos no total. (cancer, 117), (hospital, 82), (health, 70), (disease, 65), (child, 65). Frases: (i) *pressure on tweed hospital sends patients to qld*, (ii) *chef prepares recipes for cancer patients*, (iii) *a message about mens mental health*. O tema predominante neste *cluster* são **Problemas relacionados à saúde**.

- **Cluster 8**: 1767 títulos no total. (turnbull, 82), (minister, 81), (malcolm, 66), (house, 64), (interview, 63). Frases: (i) *premier opposition leader make final pitch wa election*, (ii) *obama breaks healthcare silence to slam republican proposal*, (iii) *why tasmanias western arthur range is such a tough walk*. As notícias neste *cluster* tratam de **Política**.

- **Cluster 15**: 2134 títulos no total. (market, 177), (wall, 153), (prices, 145), (street, 121), (dollar, 110). Frases: (i) *wage growth remains at record lows*, (ii) *veroguard manufacturing centre create 600 jobs northern adelaide*, (iii) *fossil beer the latest innovation to hit booming craft brewing*. Este *cluster* trata claramente de notícias ligadas ao **Mercado e setor econômico**.

- **Cluster 19**: 2075 títulos no total. (murder, 314), (charged, 207), (guilty, 205), (alleged, 158), (accused, 129). Frases: (i) *terrorist neil prakash denies charges reports*, (ii) *animal cruelty charges see woman jailed 31 dogs 43 cats*, (iii) *police officer charged with filming sex act with colleague*. Com base nas palavras mais frequentes neste *cluster*, conclui-se que o seu tema é **Crimes e assassinatos**.

4. Conclusões

Dois métodos de extração de atributos foram aplicados aos dados: *bag-of-words* com *word n-grams* e *word embeddings*. Para o primeiro, encontrou-se a partir do *elbow method* que o melhor número de *clusters* a ser utilizado seria $k = 40$. Com este k, obtivemos um distorção, isto é, a soma das distâncias ao quadrado entre cada amostra e o seu centroíde dominante, de aproximadamente 1375. Para o segundo método, adotamos $k = 20$ e chegamos a uma distorção 4000 vezes maior. Utilizando a ferramenta de visualização *t-SNE*, visualizamos *clusters* bem definidos para o primeiro método, enquanto que para o segundo, apenas uma nuvem de pontos sem separação aparente dos pontos, justificando, deste modo, a diferença observado na distorção. Em geral, foi possível a atribuição de temas aos *clusters* gerados em ambos os métodos.

Referências

[1] The cambridge analytica files: the story so far. https://tinyurl.com/y73a5zby. Accessed: 2018-05-12. 1

[2] A. Rocha, W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. R. B. Carvalho, and E. Stamatatos. Authorship attribution for social media forensics. *IEEE Transactions on Information Forensics and Security*, 12(1):5–33, Jan 2017. 1

[3] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 2

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 2