Escola Superior de Agricultura Luiz de Queiroz - USP
LCE5859 - Métodos Estatísticos Multivariados
Eduardo Elias Ribeiro Junior (edujrrib@gmail.com)
10 de março de 2017

Análise de agrupamento aplicada aos resumos dos papers apresentados no evento SIGKDD2016

Eduardo Elias Ribeiro Junior*[1]
* Departamento de Ciências Exatas LCE ESALQ-USP

Resumo

Atualmente com o avanço no poder de armazenagem muitos conhecimento é produzido e mantido em diversas mídias (servidores, unidades de disco, pendrives, etc.). Na área acadêmica o produto relacionado à transmissão de conhecimento são artigos. Comumente, em eventos científicos, têm-se os artigos expostos que são agrupados em diferentes temas. Todavia esse agrupamento é em geral realizado por humanos e nem sempre são efetivos. Nesse artigo propõe-se uma metodologia para agrupamento de artigos científicos, com base nos textos contidos em seus títulos e resumos, baseados em técnicas estatísticas multivariadas. Os dados utilizados no trabalhos são referentes aos títulos e resumos de 203 artigos apresentados na 22nd SIGKDD Conference e foram obtidos via ferramentas de webscrapping. Após higienização dos textos foram obtidos 3173 termos distintos que foram organizados em uma matriz termo-documento de frequências. Utilizando componentes principais as 3173 colunas da matriz termo-documento foram reduzidas a 199 componentes principais. Com as componentes principais estimou-se, via estatísticas Gap, que apenas 6 grupos seriam adequados para agrupamento, o que mostra que a atribuição à 16 temas distintos feita pela organização do evento é desnecessária. Os grupos puderam ser interpretados e a análise destes mostrou que o agrupamento foi satisfatório. Todo o trabalho foi elaborado com o auxílio do software R e os códigos estão disponíveis em um material suplementar online.

Palavras-chave: *text-mining*, *web-scrapping*, *agrupamento*, *pca*, *k-means*, *estatísticas Gap*.

[1]Contato: <edujrrib@gmail.com>/<jreduardo.usp.br>

1 Introdução

Atualmente se produz muito conhecimento que, com o avanço da computação e poder de armazenagem, estão mantidos em diversos mídias (servidores, unidades de disco, pendrives, etc.) ao redor do mundo. Esse conhecimento é transcrito em diferentes produtos como artigos científicos, páginas web, blogs, vídeos entre outros. Todavia, quanto maior a quantidade de conteúdo maior a dificuldade para sua organização.

Na tentativa de organizar esses produtos há diversos esforços voltados à análise de texto, uma vez que, das mídias citadas as mais comuns são expressas em forma de texto (como artigos, posts, livros, etc.). Dificuldades para análise ou mineração de textos são frequentes e estão presentes desde a coleta dos dados até a escolha da técnica adequada para análise e interpretação dos resultados.

Na mineração de dados textuais pode-se listar duas estratégias principais i) *Bag of words* em que a semântica do texto é desfeita e a estrutura linguística ignorada, sob essa abordagem os textos são representados pela frequência ou ocorrência das palavras; e ii) *Natural Processing Language (NLP)* em que as palavras são caracterizadas pelo seu sentido morfológico e os textos são analisados considerando elementos da linguagem (BERRY; KOGAN, 2010). A estratégia via *bag o words*, embora simples, é extremamente útil e mais comum. Os principais objetivos sob essa estratégia são agrupamento, classificação e predição de textos (SILGE; ROBINSON, 2017) embora outros possam existir.

Uma aplicação direta da análise de texto pode ser pensada para a comunidade acadêmica. Comumente eventos da comunidade científica reúnem pesquisadores para exposição e discussão de seus recentes trabalhos. Esses trabalhos são submetidos ao evento e, em geral, são classificados por temas, o que facilita i) os participantes a localizarem seus interesses e ii) consultas ao acervo após finalização do evento. Todavia, há pouco rigor na atribuição desses temas e muitas vezes, pela má atribuição, esses acabam sendo dispensáveis. Assim a o agrupamento de textos para geração de temas com trabalhos homogêneos tem extrema relevância para científicos de eventos.

Nesse artigo apresentamos uma análise textual dos resumos dos artigos apresentados na *22nd SIGKDD Conference*, maior evento de maior evento de Knowledge Discovery and Data Mining promovido pela Association for Computing Machinery (ACM), com o objetivo de agrupar os artigos pela similaridade de seus textos. Isso é realizado via abordagem *bag of words* utilizando técnicas multivariadas para redução de dimensionalidade e agrupamento. Após agrupamento os grupos são interpretados como temas e contrastados com os temas criados pela organização do evento.

O artigo é organizado em cinco seções. Essa primeira seção enfatiza a importância da análise de textos e sua relevância em eventos científicos. Na Seção 2 o conjunto de textos é descrito assim como destacado o procedimento de *web scrapping* para sua obtenção. Na Seção 3 são descritos os métodos utilizados na análise dos dados e na Seção 4 apresentados os resultados da aplicação dos métodos bem como algumas discussões. Por fim a Seção 5 é destinada às considerações finais obtidas desse trabalho e à apresentação de possíveis direções para pesquisas futuras.

2 Conjunto de dados

O conjunto de dados analisados referem-se aos artigos apresentados na 22nd SIGKDD Conference realizada entre os dias 13 e 17 de agosto de 2016 sob organização da Association for Computing Machinery (ACM). As informações sobre os artigos aceitos no SIGKDD estão disponíveis no sítio eletrônico `http://www.kdd.org/kdd2016`. A Figura 1 ilustra a disposição das informações no sítio eletrônico do evento. Os dados utilizados na análise correspondem aos textos dos títulos e resumos dos artigos (destacados em vermelho na figura), além dos tópicos atribuídos aos artigos pelos organizadores do evento (destacados em azul).

Para extração dos dados utilizou-se as ferramentes para raspagem de dados web disponíveis pelo pacote `rvest` (WICKHAM, 2016) do software R. O processo de extração se deu em três passos, devido a disposição das informações conforme Figura 1:

1. Obtenção dos links para as páginas dos tópicos;
2. Obtenção dos links para as páginas dos artigos;
3. Obtenção das informação de título, resumo e tópicos para cada página de cada artigo.

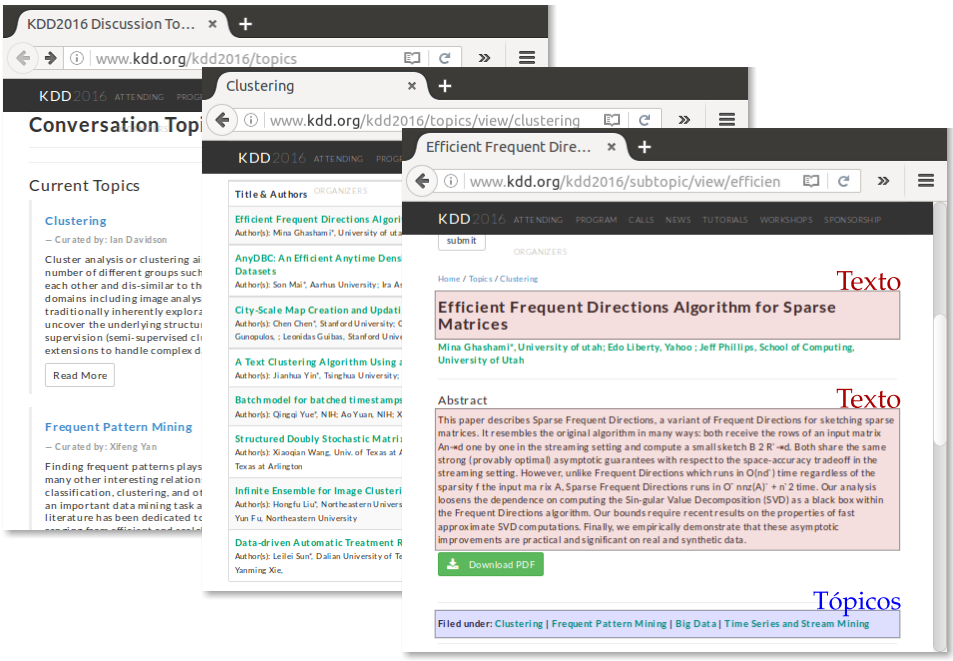

Figura 1: Sítio do SIGKDD2016 de onde foram extraídos os títulos, resumos e tópicos.

Ao todo foram 332 páginas consultadas, referentes à 203 artigos. Na Tabela 1 são apresentadas os 16 tópicos definidos pelos organizadores do SIGKDD e o número de artigos pertencentes a cada respectivo tópico. Note que nessa tabela a soma de artigos não representa o número de artigos únicos apresentados no evento, pois mais de um tópico pode ser atribuído ao mesmo artigo (veja a Figura 1). Observe também que há uma moderada predominância de artigos cujo foram atribuídos os três primeiros tópicos, ainda com apenas os 9 tópicos mais frequentes têm-se aproximadamente 86% de todas as atribuições. Das atribuições de tópicos pelos organizadores foram 115, 60, 18, 8, 1, 1 artigos com 1, 2, 3, 4, 5 e 6 tópicos atribuídos, respectivamente

Tabela 1: Frequência de artigos em cada tópico definido no evento.

	Tópico	Nº de artigos	Freq. absoluta	Freq. acumulada
1	big-data	56	0.169	0.169
2	mining-rich-data-types	55	0.166	0.334
3	graph-mining-and-social-networks	45	0.136	0.470
4	dimensionality-reduction	31	0.093	0.563
5	classification	26	0.078	0.642
6	recommender-systems	20	0.060	0.702
7	time-series-and-stream-mining	19	0.057	0.759
8	deep-learning	17	0.051	0.810
9	frequent-pattern-mining	17	0.051	0.861
10	semi-supervised-learning	10	0.030	0.892
11	optimization-techniques	9	0.027	0.919
12	clustering	8	0.024	0.943
13	large-scale-machine-learning-systems	8	0.024	0.967
14	privacy-preserving-data-mining	5	0.015	0.982
15	outlier-and-anomaly-detection	4	0.012	0.994
16	data-reliability-and-truthfulness	2	0.006	1.000
	Total	332	1.000	

3 Metodologia

Muito da análise de texto, não fundamentada em modelos estatísticos, é realizada via técnicas multivariadas. Para atender o objetivo proposto nesse trabalho faz-se uso de técnicas multivariadas de redução de dimensionalidade via componentes principais (*PCA*) e de análise de agrupamento via algoritmo de *k-means*. Nessa seção são apresentados os métodos aplicados ao conjunto de textos a fim agrupá-los de forma satisfatória.

3.1 Organização de textos para análise

Uma importante e não trivial etapa no processo de análise de textos é a adequação dos textos para análise. Conforme descrito na Seção 1 a abordagem via *bag of words*, transforma os textos em uma matriz $\mathbb{X}$, cujo linhas são os documentos, colunas são os termos e os valores representam as frequências dos termos nos documentos. Para obtenção dos termos faz-se a higienização dos textos que consiste na remoção das palavras de parada (preposições, artigos, conjunções, etc.), da pontuação, dos espaços em branco e dos números. Além da remoção desses caracteres, faz-se a radicalização das palavras, para que palavras de mesmo sentido sejam associadas ao mesmo termo, por exemplo as `clustering`, `cluster`, `clusterization`, `clustered` possuem todas o mesmo radical, `cluster`, portanto ficam todas associadas ao termo `cluster`. Existem vários algoritmos para radicalização, nesse trabalho utilizou-se o algoritmo de Poter (PORTER, 2001).

3.2 Redução de dimensionalidade via componentes principais

Não raramente a matriz termo-documento de frequências é retangular possuindo um número de colunas (termos) muito maior que o número de linhas (documentos), além de ser razoavelmente esparsa. Assim é de interesse em análise de texto a redução do número de colunas da matriz termo-documento $\mathbb{X}$. Uma técnica multivariada, cujo um dos objetivos é a redução de dimensionalidade é chamada de análise de componentes principais, que consiste construção de novas variáveis denominadas componentes principais a partir da decomposição da matriz de covariância ou correlação de $\mathbb{X}$ (FRIEDMAN; HASTIE; TIBSHIRANI, 2001).

Para apresentação do método tome $\mathbf{R}$ como a matriz de correlações de $\mathbb{X}$, de dimensão $p \times p$. $\mathbf{R}$ pode ser escrita como $\mathbf{E}\Lambda\mathbf{E}^t$ ($\mathbf{R} = \mathbf{E}\Lambda\mathbf{E}^t$), em que Λ é a matriz diagonal dos autovalores λ_i, $i = 1, \ldots, p$ da matriz $\mathbb{X}$, ou seja, $\Lambda = \text{diag}(\underline{\lambda})$; e $\mathbf{E}$ é a matriz dos autovetores $\underline{e}_i$, $i = 1, \ldots, p$. Assim as novas variáveis, denominadas componentes principais são dadas por $\underline{z}_i = \mathbb{X}\underline{e}_i$, para $i = 1, \ldots, p$.

Note que são calculadas tantas componentes quanto for o número de variáveis, porém apenas um subconjunto dessas componentes já é suficiente para explicar boa parte da variação total dos dados originais $\mathbb{X}$. O percentual da variância explicada pela i-ésima componente é dada por $\lambda_i / \sum_{j=1}^{p} \lambda_j$.

A redução de dimensionalidade ocorre por tomar a matriz de componentes principais $\mathbf{Z}$, com m componentes, necessárias para explicar um percentual arbitrário da variação total, $m << p$, ao invés da matriz $\mathbb{X}$. Embora MANLY (2008) sugira que a análise de agrupamento realizada sobre componentes principais deva ser evitada, em problemas *"small n large p"* essa é uma solução comumente utilizada.

3.3 Agrupamento K-means

Métodos de agrupamento visam agrupar as observações de um conjunto de dados em grupos, cujo as observação de um grupo sejam similares e as observações de grupos distintos sejam distintas. Existem métodos de agrupamentos hierárquicos e não-hierárquicos. Nos hierárquicos normalmente se define o número de grupos a partir de um dendrograma, essa escolha é geralmente subjetiva e priorizasse, nesses casos, a interpretação dos grupos formados. Para os não-hierárquicos o número de grupos deve ser informado previamente e então um algoritmo atribuirá as observação aos grupos.

O método não-hierárquico mais comumente utilizado é o chamado método de k-means (HARTIGAN; WONG, 1979). Nesse trabalho utiliza-se método de agrupamento não-hierárquico supra-citado. O algoritmo 1 descreve o procedimento para formação dos grupos via *k-means*. Note que nos dados de

entrada do algoritmo o número de grupos a serem formados e os dados devem ser especificados.

Algoritmo 1: Agrupamento não-hierárquico de k-means.

Entrada:
k : número de grupos a serem formados (escalar inteiro);
$\mathbf{X}$: conjunto de dados (matriz $n \times p$).
Saída:
$\underline{g}$: a atribuição dos grupos a cada observação (vetor de tamanho n);

1 Atribua aleatoriamente as observações aos grupos;
2 **while** *As atribuições dos grupos não for a mesma.* **do**
3 **(a)** Para cada grupo k calcule o seu centroide. O centroide do k-ésimo grupo é dado pelo vetor média das p variáveis calculado das observação pertencentes a esse grupo;
4 **(b)** Atribua cada observação ao grupo, cujo tem o centroide mais próximo, a proximidade é dada distância euclidiana entre o centroide e o vetor de variáveis da observação
5 **end**

Conforme visto no algoritmo 1 o número de grupos k deve ser informado a priori, para avaliação do agrupamento realizado com k grupos pode-se verificar as medidas de homogeneidade dos grupos e heterogeneidade entre grupos. A homogeneidade é dada pelo inverso da soma de quadrados das observações alocadas em um mesmo grupos e a heterogeneidade pela soma de quadrado das observações alocadas em grupos distintos.

O principal problema em análise de agrupamento é a escolha do número de grupos, k, a serem formados. TIBSHIRANI; WALTHER; HASTIE (2001) propuseram a estatística Gap para estimação de k. Essa estatística é calculada usando as medidas de homogeneidade do agrupamento em contraste com medidas de variáveis simuladas, em geral, por uma distribuição uniforme. A escolha do número de grupos é então realizada comparando as estatística Gap para um k. Para maiores informações sobre o algoritmo consulte TIBSHIRANI; WALTHER; HASTIE (2001) e DUDEK (2016).

3.4 Recursos computacionais

As análises presentes nesse artigo são todas realizadas no *software* R (versão 3.3.3) e estão disponíveis no sítio eletrônico do autor `https://jreduardo.github.io/lce5859-mem/`.

4 Resultados e discussão

Após extração dos dados (Seção 2) realizou-se a higienização dos textos que consistiu na remoção das palavras de parada (preposições, artigos, conjunções, etc.), da pontuação, dos espaços em branco e dos números. Além da remoção desses caracteres, realizou-se a radicalização das palavras restantes utilizando o algoritmo de Poter (PORTER, 2001). Do processo de higienização descrito, restaram `r length(tm::Terms(dtm_texts))` palavras distintas em todos os textos.

Na Figura 2 são apresentados os 5% termos, provenientes da radicalização, mais frequentes em todos os textos. Note que os termos mais frequentes são realmente àqueles utilizadas no ambiente de Knowledge Discovery and Data Mining. Destaque para as palavras **model** e **data**, o que reflete a característica do evento em discutir estudos aplicados.

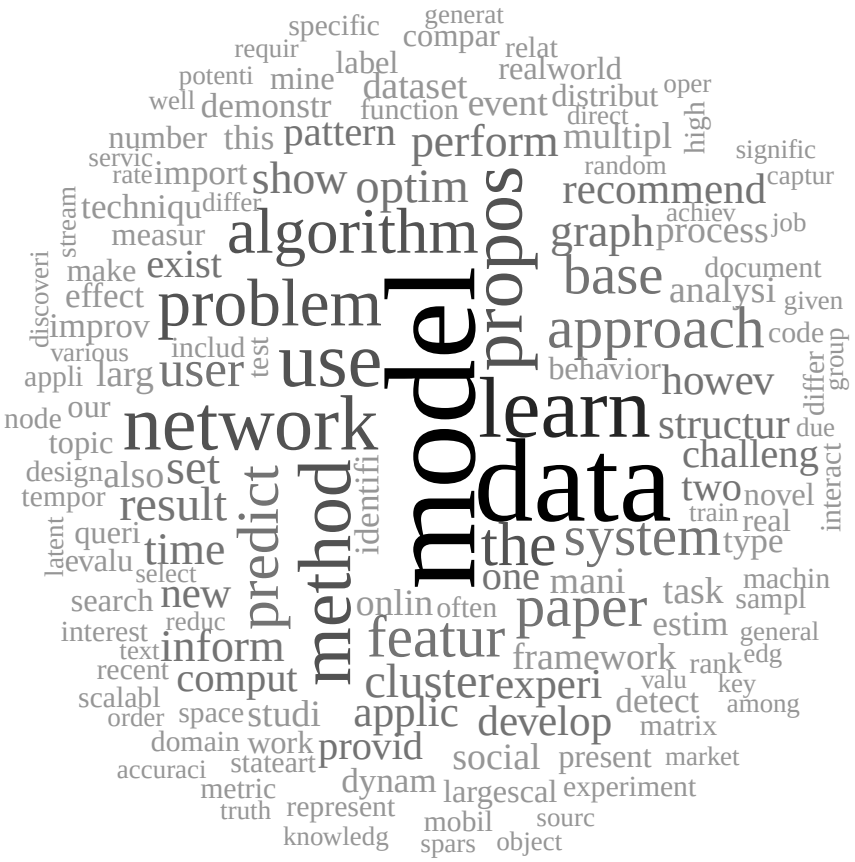

Figura 2: Nuvem dos 5% termos mais frequentes nos textos extraídos dos resumos e títulos dos artigos apresentados no SIGKDD 2016.

Os termos, provenientes da radicalização palavras, e os documentos, nesse caso os artigos, formam uma matriz denominada matriz termo-documento ($\mathbb{X}$). Nessa matriz as linhas são os artigos e as colunas os termos, os valores representam a frequência do termo no artigo. O gráfico à esquerda na Figura 3 ilustra essa matriz, apenas com os 30 termos mais frequentes. Note que há muitos valores iguais ou próximos a zero, mesmo dentre os 30 termos mais frequentes, isso ocorre pois há muitos termos específicos dentro da área de Knowledge Discovery e Data Mining.

A dimensão da matriz $\mathbb{X}$ é 203×3173. Como o número de colunas é muito maior que o número de linhas, qualquer análise de agrupamentos (que utilize distâncias) ficará prejudicada devido a chamada "maldição da dimensionalidade" (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). Assim procedeu-se com a construção de componentes principais para redução de dimensionalidade. No gráfico à direita da Figura 3 é apresentada a evolução do percentual da variância explicada pelo número de componentes principais. Note que com 199 componentes já explica-se quase 100% da varição total. Portanto, foram tomadas 199 componentes, uma redução de 2974 dimensões colunas da matriz $\mathbb{X}$.

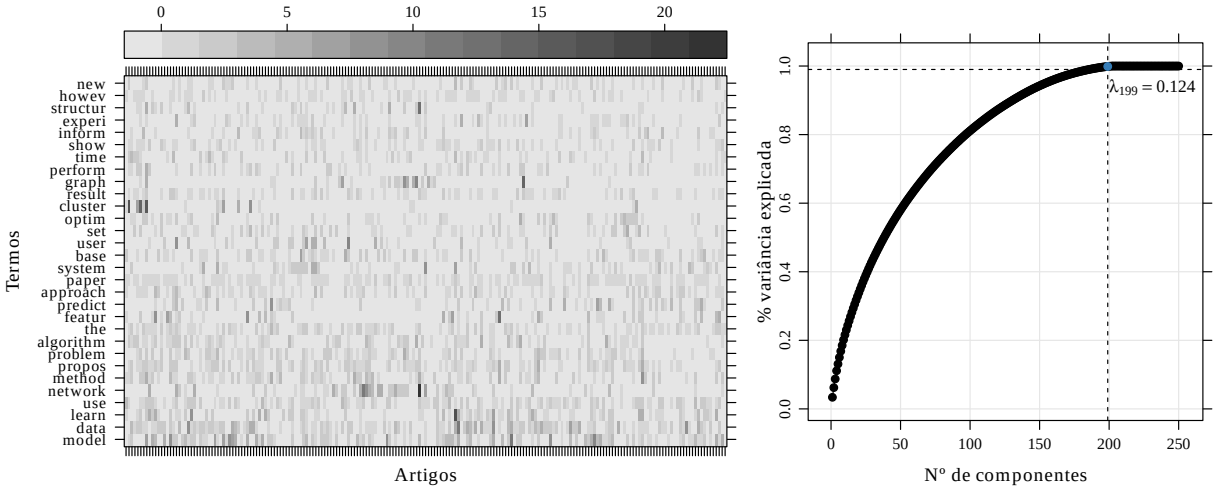

Figura 3: Representação da matriz termo-documento de frequências com os 50 termos mais frequentes (à esquerda) e proporção acumulada da variância explicada pelo número de componentes considerados (à direita)

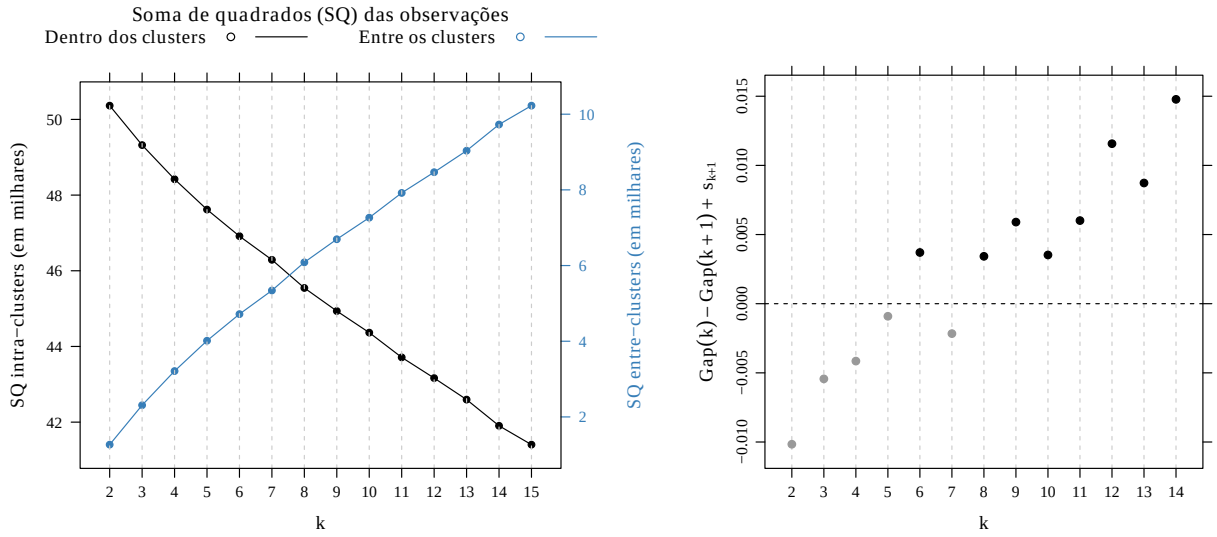

Figura 4: Medidas de qualidade de agrupamento. Soma das distâncias euclidianas intra e entre clusters (esquerda) e diferenças de indíces Gap (direita).

Com as 199 componentes realizou-se o agrupamento não hierárquico *k-means*, conforme algoritmo proposto por HARTIGAN; WONG (1979). Como agrupamentos não hierárquicos necessitam que o número de grupos a serem formados seja conhecido, os artigos foram agrupados em $2, 3, \ldots, 15$ grupos. Os resultados dos agrupamentos são apresentados na Figura 4. No gráfico à esquerda são exibidas as somas de quadrados $\sum_{i=j}(x_i - x_j)^2$ das observações alocadas em um mesmo grupo (SQ intra-cluster, em preto), e a soma de quadrados das observações alocadas em um grupos diferentes (SQ entre-cluster, em azul). Para um bom agrupamento espera-se que a distância intra-cluster seja baixa e a distância entre-cluster alta, porém essas são medidas inversamente proporcionais e assim desejamos um bom compromisso entre as duas medidas para a escolha do número de grupos. Embora as escalas sejam distintas (valores dos eixos y), por esse gráfico, o número de grupos adequado está em torno de 7 ou 8. Calculando os estatísticas Gap (TIBSHIRANI; WALTHER; HASTIE, 2001) o número de grupos indicado é 6, o menor k em que a condição $Gap(k) - Gap(k+1) + s_{k+1} > 0$ (gráfico à direita da Figura 4), o que está concordante com o gráfico de distâncias intra e entre-clusters.

Note que o número de grupos estimado pelas estatísticas Gap, assim como visto nas medidas de

homogeneidade dos grupos e heterogeneidade entre grupos, mostram que os 16 tópicos criados pelos organizadores do evento são muito mais do que o necessário para agrupar artigos similares.

Os 6 grupos formados pelo algoritmo k-means contém 6, 36, 120, 3, 15, 23 elementos (respectivamente). Para permitir a interpretação dos grupos, exibi-se na Figura 5 os 5% termos mais frequentes em cada grupo. Note que o Grupo 3 é formado por um número de artigos muito maior que os demais, o que indica o que consequentemente acarreta em um número maior de termos e maior heterogeneidade do grupo. Isso fica claro ao observar a gama de termos apresentadas para esse grupo na Figura 5. A partir dessa figura também pode-se caracterizar os grupos:

- **Grupo 1:** artigos essencialmente sobre agrupamentos;
- **Grupo 2:** artigos essencialmente sobre propostas de modelagem de dados;
- **Grupo 3:** artigos diversos sobre análise de dados para Knowledge Discovery e Data Mining;
- **Grupo 4:** artigos essencialmente sobre reconhecimento de padrão e e análise de grafos;
- **Grupo 5:** artigos relacionados a métodos de aprendizado de máquina[2]; e
- **Grupo 6:** artigos essencialmente sobre redes neurais.

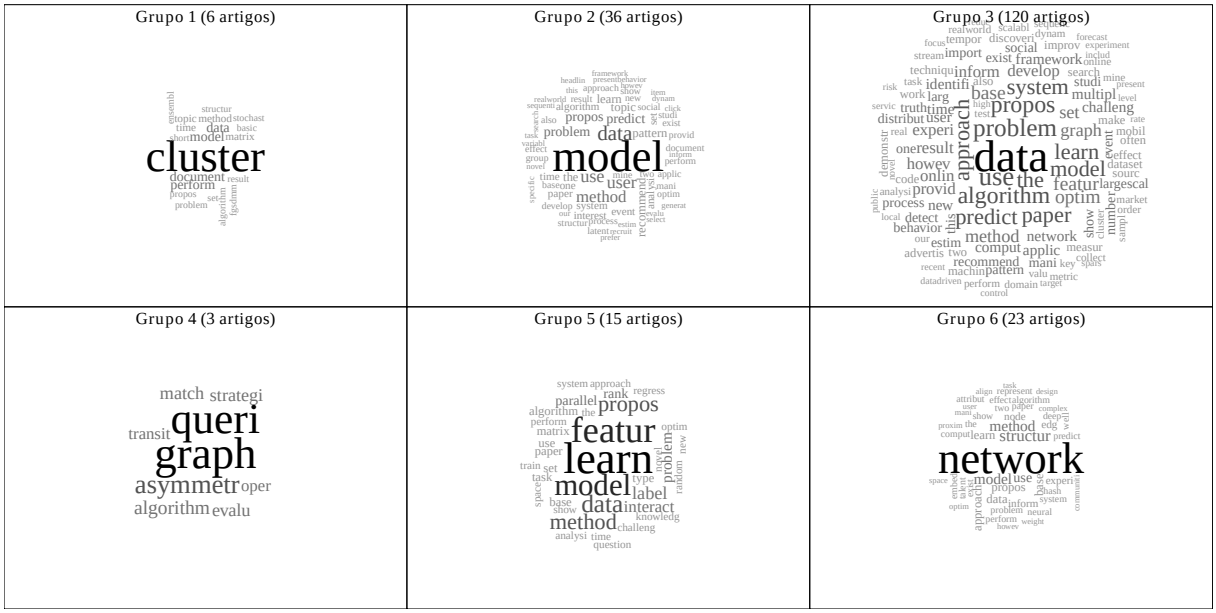

Figura 5: Nuvem com os 5% termos mais frequentes em cada grupo formado pelo algoritmo k-means.

Como resultado complementar, são apresentados na Tabela 2 os títulos de dois artigos escolhidos aleatoriamente de cada um dos grupos formados. Pode se observar que os títulos apresentados condizem com a representação dos grupos realizada na Figura 5. Porém vale ressaltar que o título somente uma pequena parte do conjunto de textos que caracterizam o artigo, a maior parte da informação provém do resumo.

Tabela 2: Amostras aleatórias de dois artigos em cada grupo

Índice	Grupo	Título do artigo
2	1	A Text Clustering Algorithm Using an Online Clustering Scheme for Initialization
8	1	Structured Doubly Stochastic Matrix for Graph Based Clustering
150	2	Fast Unsupervised Online Drift Detection Using Incremental Kolmogorov-Smirnov Test
275	2	Privacy-preserving Class Ratio Estimation
325	3	Developing a Data-Driven Player Ranking in Soccer using Predictive Model Weights
63	3	Minimizing Legal Exposure for High-Tech Companies through Collaborative Filtering Methods
174	4	Scalable Pattern Matching over Compressed Graphs via Dedensification

[2] o termo *feature* em machine learning é geralmente usado como sinônimo para variáveis preditoras.

168	4	Online Asymmetric Active Learning with Imbalanced Data
162	5	Scalable Fast Rank-1 Dictionary Learning for fMRI Big Data Analysis
131	5	Images Don't Lie: Transferring Deep Visual Semantic Features to Large-Scale Multimodal Learning to R
25	6	Inferring Network Effects from Observational Data
134	6	Large-Scale Item Categorization in e-Commerce Using Multiple Recurrent Neural Networks

5 Conclusões

Nesse artigo foram apresentados os procedimentos para extração de dados textuais de páginas web, nomeadamente, os títulos e resumos dos artigos apresentados na *22nd SIGKDD Conference*. Foram 332 páginas consultadas e 203 artigos, cujo título e resumos foram extraídos. Após pré-processamento os 203 artigos forneceram 3173 termos distintos. A coleção de 203 artigos, organizada em uma matriz termo-documento, de dimensão 203×3173 foi submetida a uma análise de componentes principais onde 199 componentes foram retidas para continuidade da análise, ou seja, a matriz termo-documento foi reduzido de 3173 colunas para 199. Com a nova matriz em que para cada artigo foi associado valores de 199 componentes, realizou-se análises de agrupamento por *k-means* para k's de 2 a 15. Verificou-se, pela estatística de Gap, que o número de grupos adequado a esses dados são 6, muito menor do que os 16 tópicos definidos pelo evento. A avaliação dos grupos mostrou que o agrupamento conseguiu juntar artigos com conteúdos similares podendo nomear esses grupos como

- **Grupo 1:** artigos essencialmente sobre agrupamentos;
- **Grupo 2:** artigos essencialmente sobre propostas de modelagem de dados;
- **Grupo 3:** artigos diversos sobre análise de dados para Knowledge Discovery e Data Mining;
- **Grupo 4:** artigos essencialmente sobre reconhecimento de padrão e e análise de grafos;
- **Grupo 5:** artigos relacionados a métodos de aprendizado de máquina; e
- **Grupo 6:** artigos essencialmente sobre redes neurais.

Com base nos resultados obtidos fica evidente que uma abordagem estatística ou heurística, para a definição de temas para trabalhos apresentados em eventos científicos, leva a uma melhor organização dos trabalhos, do que a atribuição subjetiva realizada pelos organizadores. A aplicação da metodologia apresentada também não se restringe a trabalhos acadêmicos, agrupamento de *posts* em blogs, de comentários em redes sociais, reclamações em central de atendimentos, entre outros são exemplos em que as análises discutidas no artigo podem ser replicadas com a devida adequação.

Como pesquisas decorrentes desse trabalho sugere-se uma abordagem probabilística para definição dos temas, BLEI (2012) apresenta alguns modelos de probabilísticos de tópicos que podem auxiliar na formação dos temas, em especial o modelo de alocação latente de Dirichlet parece satisfazer o objetivo formulado nesse artigo.

Referências

BERRY, M. W.; KOGAN, J. Text mining: applications and theory. **John Wiley & Sons**, 2010.

BLEI, D. M. Probabilistic Topic Models. **Commun. ACM**, v. 55, n. 4, p. 77–84, 2012.

DUDEK, M. W. A. **clusterSim: Searching for optimal clustering procedure for a data set**.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. **The elements of statistical learning**. Springer series in statistics Springer, 2001.

HARTIGAN, J. A.; WONG, M. A. A k-means clustering algorithm. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, v. 28, n. 1, p. 100–179, 1979.

MANLY, B. F. J. **Métodos Estatísticos Multivariados: uma introdução**. 3rd. ed.

PORTER, M. F. **Snowball: A language for stemming algorithms**, 2001. Disponível em: <http://snowball.tartarus.org/>

SILGE, J.; ROBINSON, D. **Text Mining with R: a tidy approach**. O'Reilly Media, 2017.

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 63, n. 2, p. 411–423, 2001.

WICKHAM, H. **rvest: Easily Harvest (Scrape) Web Pages**.