Text Clustering - Análise de resultados

Eduardo C. Dias — Nº USP: 9778686
Lígia H. Yamamoto — Nº USP: 9490946
Rodrigo S. Hirama — Nº USP: 9894428
Victor G. O. M. Nicola — Nº USP: 9844881

Conjuntos de dados escolhidos

- BBC
  - 2250 textos;
  - 5 categorias: *business*, *entertainment*, *politics*, *sport* e *tech*.
- 20 NewsGroups
  - 5856 textos;
  - 6 categorias: *comp.os.ms-windows.misc*, *rec.autos*, *sci.med*, *sci.space*, *soc.religion.christian* e *talk.politics.guns*.

Pré-processamento

- Conversão para letras minúsculas;
- Remoção de Stop Words;
- Remoção de números;
- Stemming;
- Remoção de termos esparsos.

Pré-processamento

Remoção de termos esparsos

- BBC: termos com frequência relativa menor que 0.05, 0.1, 0.15.
- 20NewsGroups: termos com frequência relativa menor que 0.01, 0.05, 0.1, 0.15.

K-Means e suas variações

NORMALIZAÇÃO DOS DADOS PARA K-MEANS

- De início, os resultados apresentaram-se pouco espalhados entre os *clusters* (poucos *clusters* com praticamente todos os documentos).
- A aplicação da normalização sobre os dados (transformando-os em vetores unitários) corrigiu este problema.

$$norm(x) = \frac{x}{\sqrt{\sum_{i=1}^{m} x_i^2}}$$

DEFININDO RESULTADO SATISFATÓRIO

- Para a realização dos testes foram identificados como “resultados satisfatórios” aqueles em que a maioria dos documentos de cada categoria está em *clusters* diferentes.
- Exemplo de resultado satisfatório:

	business	entertainment	politics	sport	tech
2	1	1	6	115	229
3	88	38	56	340	85
5	270	1	8	2	78
4	147	346	33	5	3
1	4	0	314	49	6

DEFININDO CONCEITO DE PRECISÃO

- A precisão é calculada para um *cluster* C_i e sua conhecida categoria[1] S_j da seguinte maneira:

$$precisao(C_i, S_j) = \frac{|\ C_i \bigcap S_j\ |}{|\ C_i\ |}$$

Baseado em: E. Rendón, I. Abundez, A. Arizmendi and E. Quiroz, "Internal versus External cluster validation indexes", in International Journal of Computers and Communications - Vol. 5, 2011, pp. 27-34.

[1] Importante notar que é possível apenas calcular a precisão de resultados satisfatórios.

K-MEANS

- Foram testadas diversas configurações de pré-processamento, e as representações TF-IDF e Binária apresentaram melhores resultados, em comparação à TF;
- O conjunto com os refinamentos *Remove numbers*, *stop words*, *stemming com toLower e remove sparse terms 0.95*, embora apresente dimensionalidade muito menor (cerca de 650 dimensões, em comparação ao total de 29069 dimensões do conjunto original), proporcionou um grau de precisão satisfatório.

BBC: K-MEANS

PRECISÃO MÉDIA DO K-MEANS PADRÃO NO CORPUS BBC, UTILIZANDO CÁLCULO DE DISTÂNCIAS POR SIMILARIDADE COSSENO

	Business	Politics	Sport	Tech	Entertainment
Binária	0.89957	0.93775	0.94665	0.91586	0.83082
TF	0.82316	0.84113	0.87818	0.76169	0.64994
TF-IDF	0.89651	0.89715	0.97706	0.88736	0.76670

De 30 testes realizados para o cálculo de precisão média, houve para as representações:

- Binária: 23 agrupamentos satisfatórios
- TF: 13 agrupamentos satisfatórios
- TF-IDF: 16 agrupamentos satisfatórios

BBC: K-MEANS

PRECISÃO MÉDIA DO K-MEANS PADRÃO NO CORPUS BBC, UTILIZANDO CÁLCULO DE DISTÂNCIA EUCLIDIANA

	Business	Politics	Sport	Tech	Entertainment
Binária	0.93249	0.93491	0.93831	0.89787	0.84925
TF	0.79628	0.79553	0.87342	0.74537	0.59463
TF-IDF	0.90062	0.69178	0.94438	0.86599	0.59654

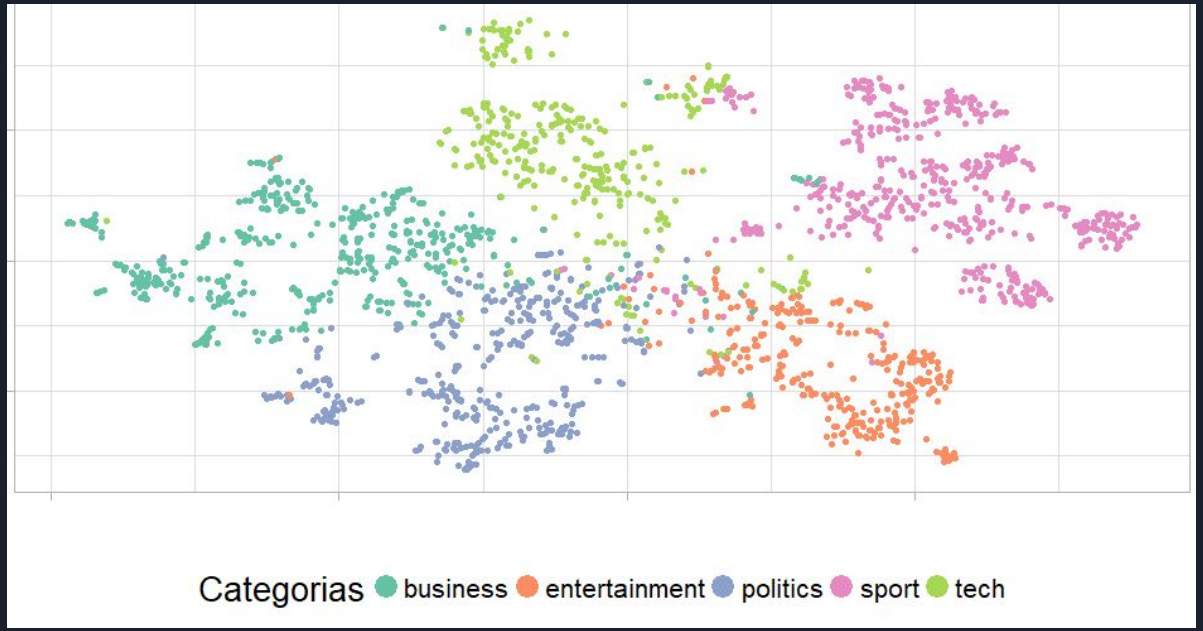
De 30 testes realizados para o cálculo de precisão média, houve para as representações:

- Binária: 21 agrupamentos satisfatórios
- TF: 19 agrupamentos satisfatórios
- TF-IDF: 19 agrupamentos satisfatórios

BBC: K-MEANS++

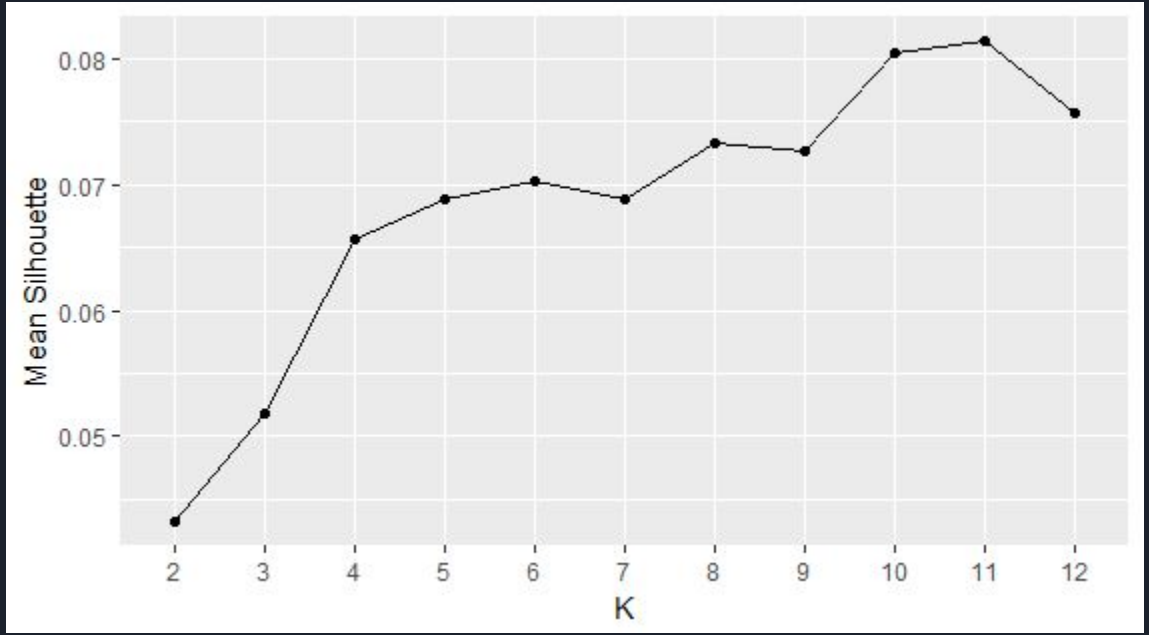
- Os resultados de precisão média foram muito semelhantes aos obtidos com K-means padrão, com uma melhora pouco expressiva.

BBC: VISUALIZAÇÃO



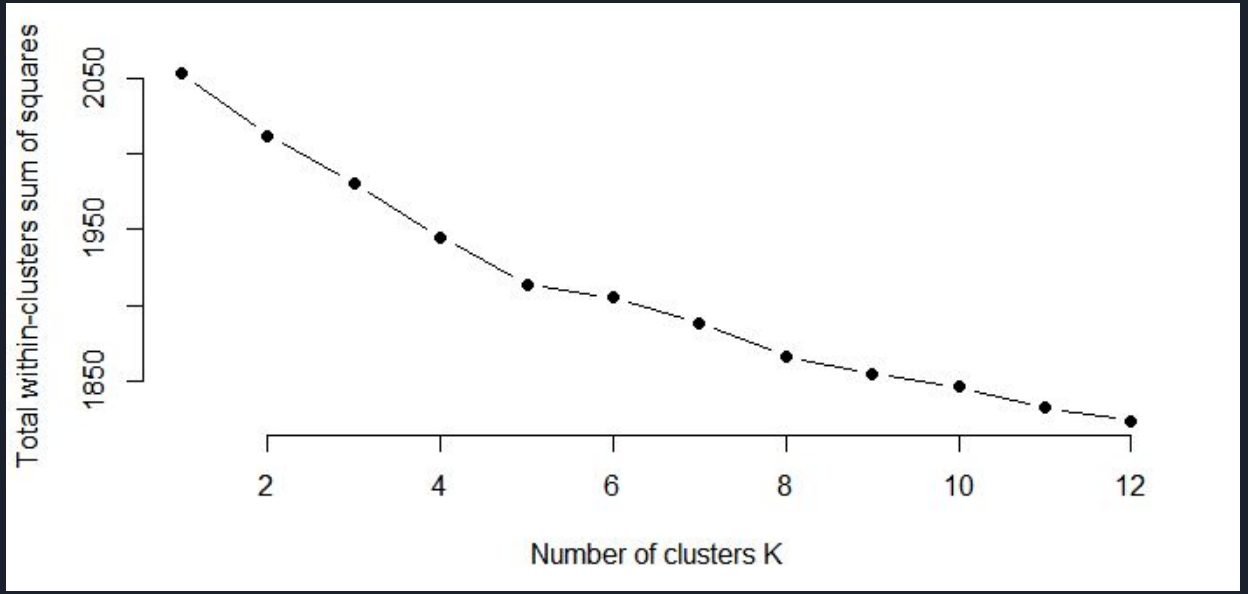
t-SNE no corpus BBC com representação TF-IDF e pré- processamento com Remove numbers, stop words e stemming e remove Sparse Terms 0,95, utilizando o K-means++ com distância cosseno e como critérios de parada a convergência ou 20 iterações.

BBC: SILHOUETTE



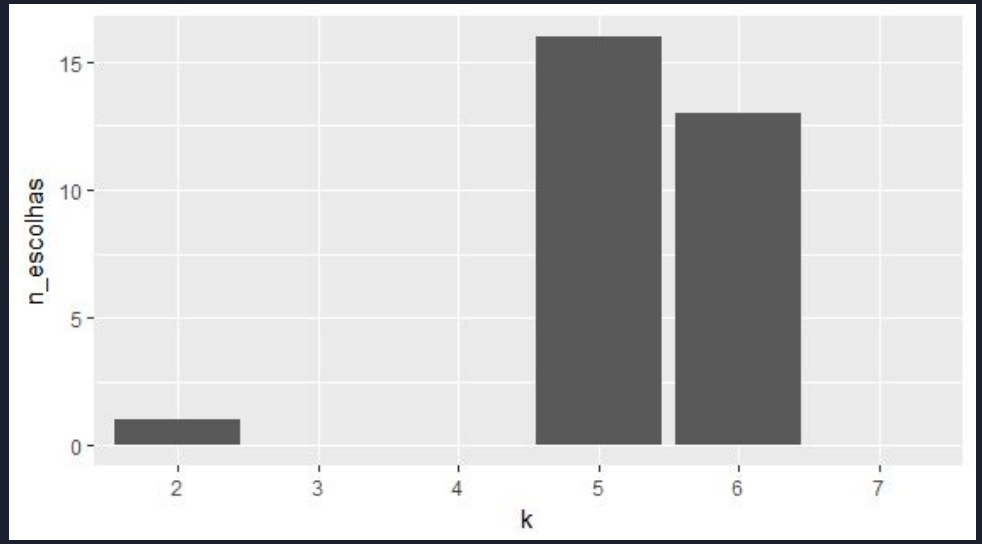
Silhouette para o K-means++ com distância cosseno e único critério de parada a convergência, no corpus BBC com representação TF-IDF e pré-processamento com Remove numbers, stop words e stemming e remove Sparse Terms 0.95 (valores médios com base em 30 testes).

BBC: ELBOW METHOD



Elbow Method para o K-means++, com representação TF-IDF e pré- processamento com Remove numbers, stop words e stemming com toLower e remove sparse terms 0.95, para o corpus BBC.

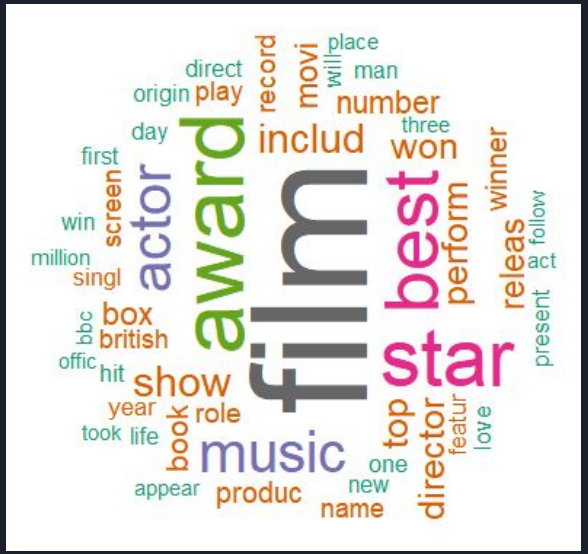
BBC: X-means



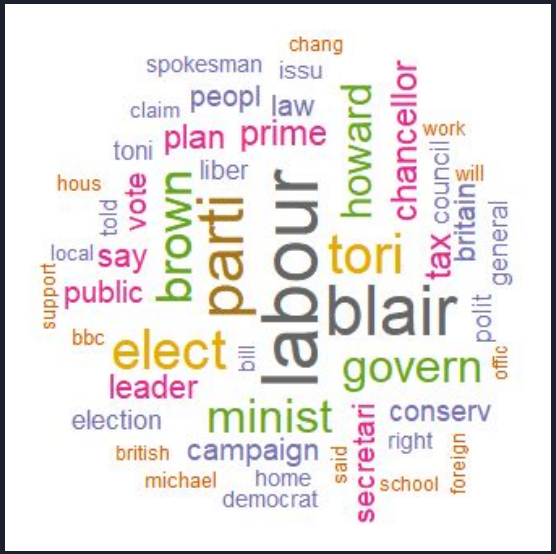
Quantidade de vezes que cada valor de k foi escolhido em um total de 30 testes. Parâmetros: distância cosseno com critérios de parada 20 iterações ou convergência, no corpus BBC com representação binária e pré-processamento com Remove numbers, stop words e stemming e remove Sparse Terms 0,95.

BBC: WORLD CLOUDS, COM K=5

Business



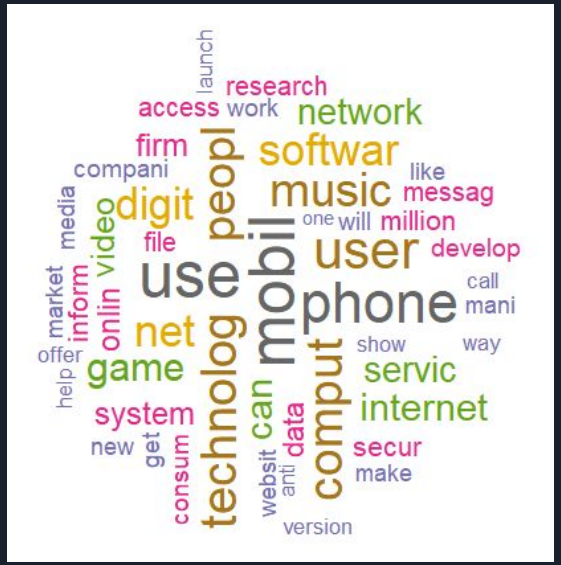
Entertainment



Politics

BBC: WORLD CLOUDS, COM K=5

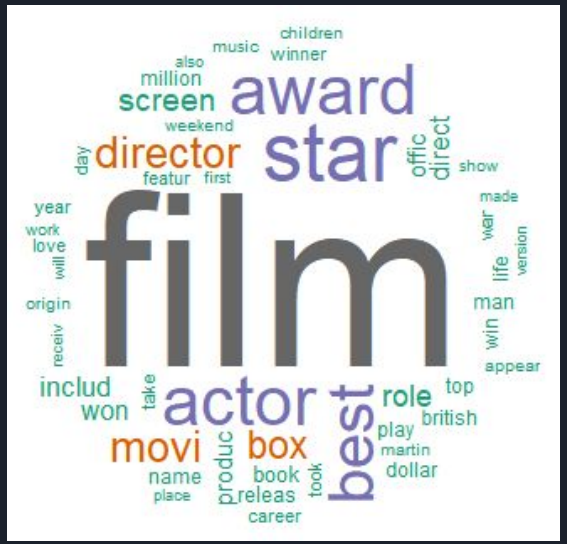
Sport



Tech

BBC: WORLD CLOUDS, COM K=6

A categoria *Entertainment* foi dividida em duas subcategorias: *Film* e *Music*

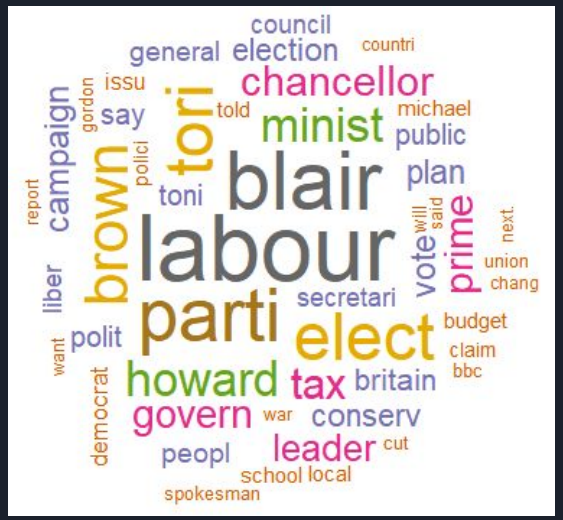


Film

Music

BBC: WORLD CLOUDS, COM K=6

A categoria *Politics* foi dividida em duas subcategorias: *Politics* e *Law*



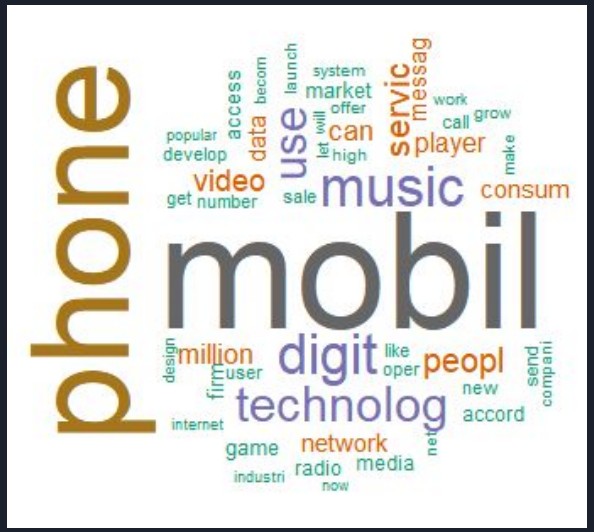
Politics

Law

BBC: WORLD CLOUDS, COM K=6

A categoria *Tech* foi dividida em duas subcategorias: *Tech* e *Mobile*

Tech



Mobile

20 NEWSGROUPS: K-MEANS

- A representação TF-IDF gerou os melhores resultados, a TF não teve boa precisão, e a binária não conseguiu dividir bem as categorias entre os *clusters*;
- O refinamento *Remove numbers*, *stop words e stemming com toLower e remove sparse terms 0.95* (com 330 dimensões) foi utilizado, entretanto, apresentando resultados não satisfatórios.
- A alternativa com remoção de termos esparsa 0.99 (com 1858 dimensões) foi utilizada melhorando os agrupamentos. Os testes apresentados a seguir utilizam esta configuração.

20 NEWSGROUPS: K-MEANS

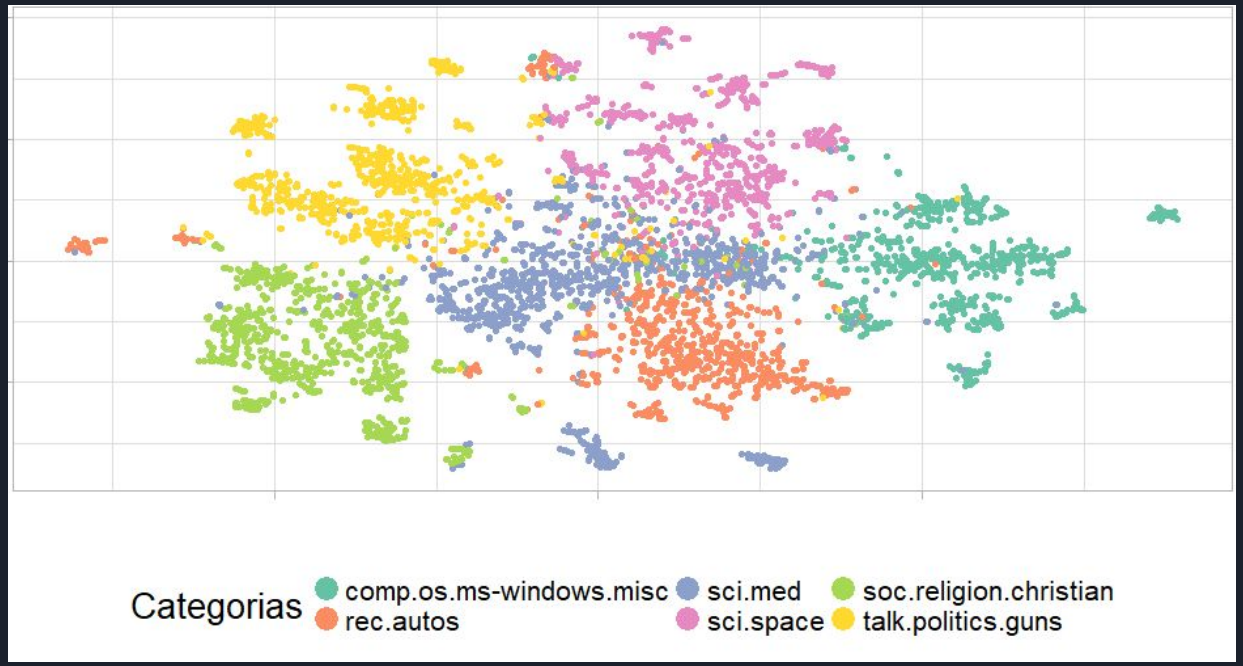
Tabela V
PRECISÃO MÉDIA DO K-MEANS PADRÃO NO CORPUS 20 NEWSGROUPS, PARA O PRÉ-PROCESSAMENTO *Remove numbers, stop words e stemming com toLower e remove sparse terms 0.99*

	Windows	Autos	Med	Space	Religion	Guns
Bin	-	-	-	-	-	-
TF	0.5870	0.4123	0.3778	0.4401	0.6980	0.3719
TF-IDF	0.8972	0.9066	0.9062	0.8794	0.9317	0.7346

De 30 testes realizados para o cálculo de precisão média, houve para as representações:

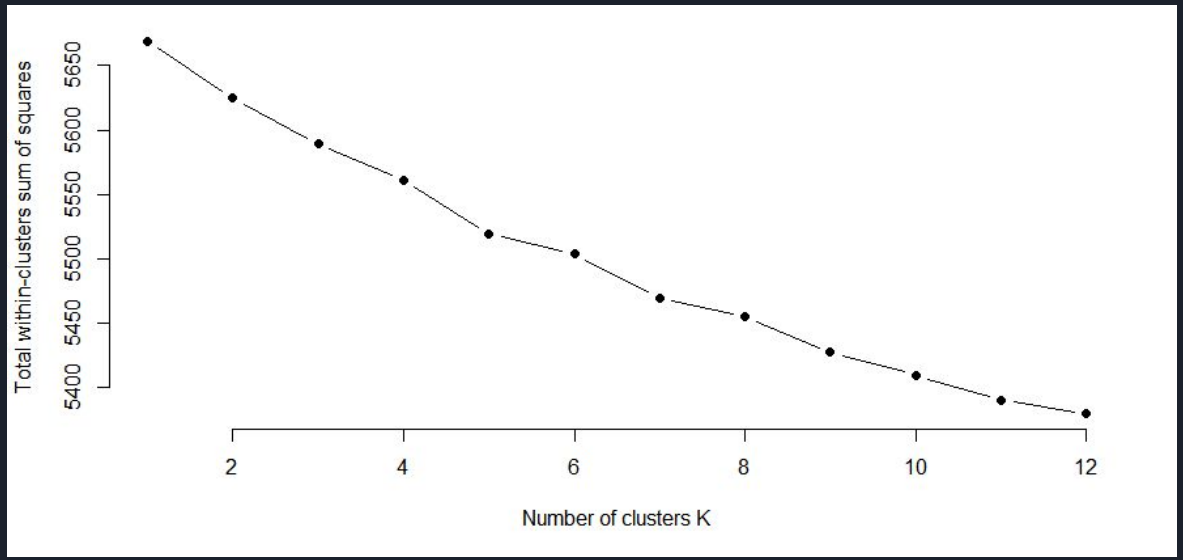
- Binária: nenhum agrupamento satisfatório
- TF: 5 agrupamentos satisfatórios
- TF-IDF: 12 agrupamentos satisfatórios

20 NEWSGROUPS: VISUALIZAÇÃO



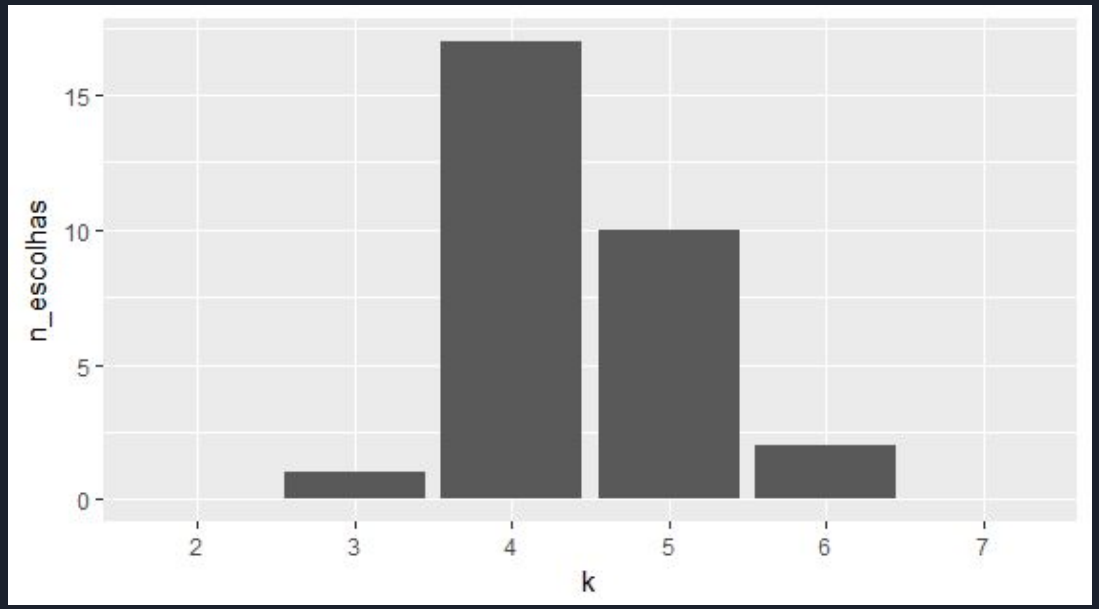
t-SNE no corpus 20 Newsgroups com representação TF-IDF e pré-processamento com Remove numbers, stop words e stemming e remove Sparse Terms 0,99, utilizando o K-means++ com distância cosseno e como critérios de parada a convergência ou 20 iterações.

20 NEWSGROUPS: ELBOW METHOD



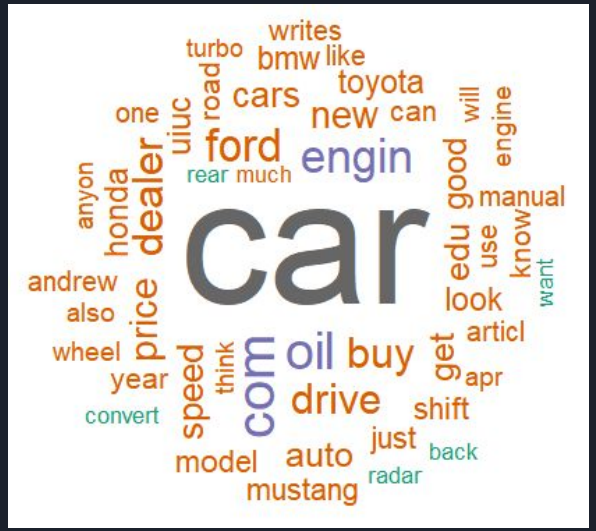
Elbow Method para o K-means++, com representação TF-IDF e pré-processamento com Remove numbers, *stop words* e stemming com toLower e remove sparse terms 0.95, para o corpus 20 Newsgroups.

20 NEWSGROUPS: Xmeans

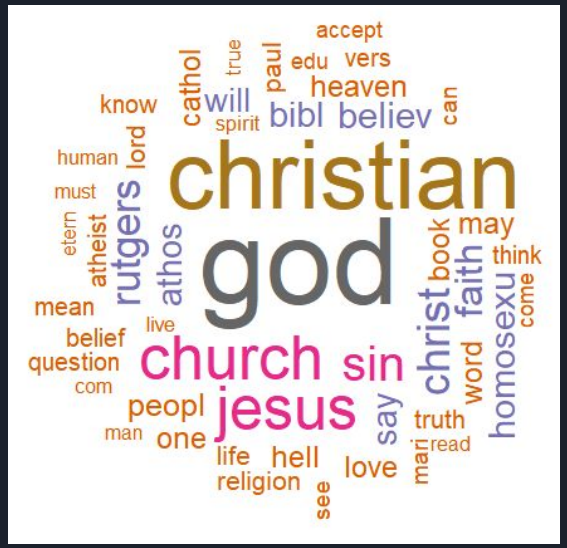


Quantidade de vezes que cada valor de k foi escolhido em um total de 30 testes. Parâmetros: distância cosseno com critérios de parada 20 iterações ou convergência, no corpus 20 Newsgroups com representação binária e pré-processamento com Remove numbers, stop words e stemming e remove Sparse Terms 0,95.

20 NEWSGROUPS: WORLD CLOUDS, COM K=6



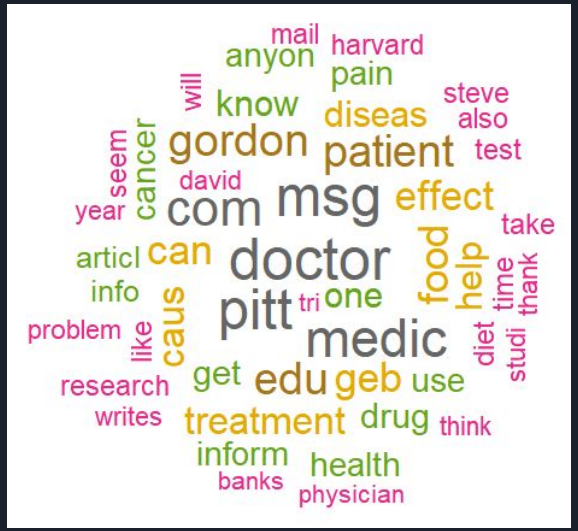
rec.autos



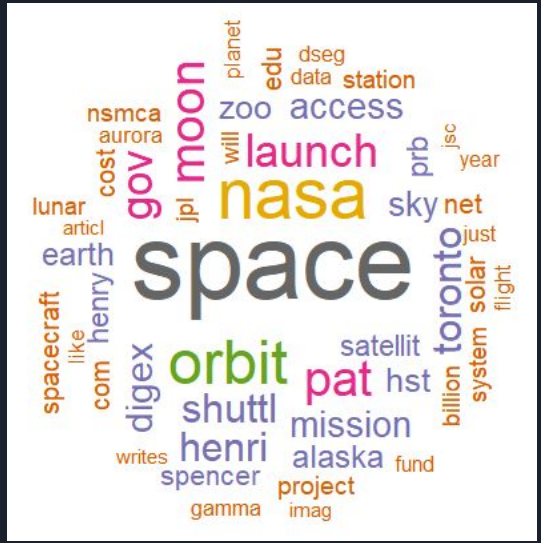
soc.religion.christian

talk.politics.guns

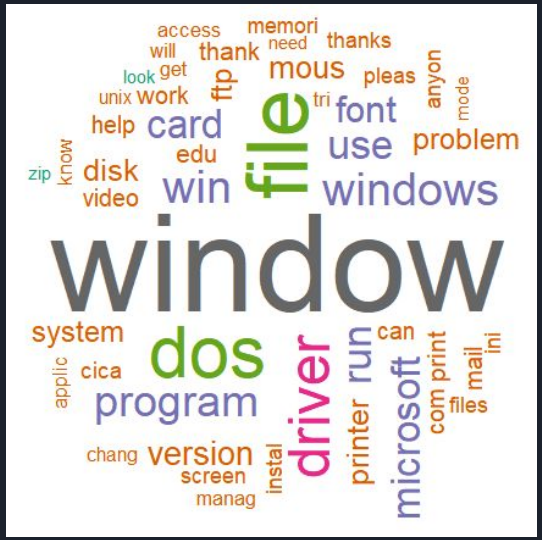
20 NEWSGROUPS: WORLD CLOUDS, COM K=6



sci.med



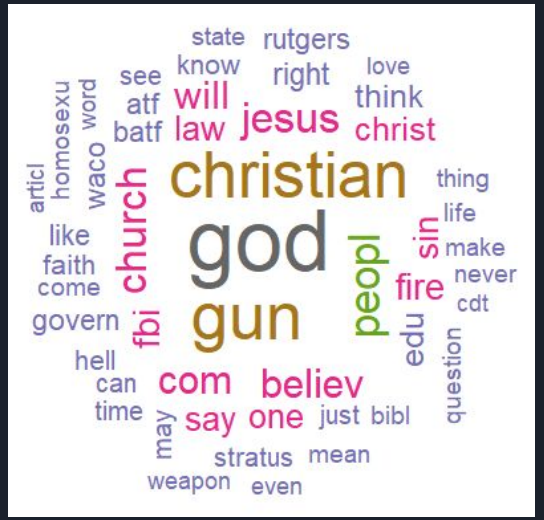
sci.space



comp.os.ms-windows.misc

20 NEWSGROUPS: WORLD CLOUDS, COM K=5

As categorias *soc.religion.christian* e *talk.politics.guns* foram unidas em uma única categoria, denominada *Ethic*

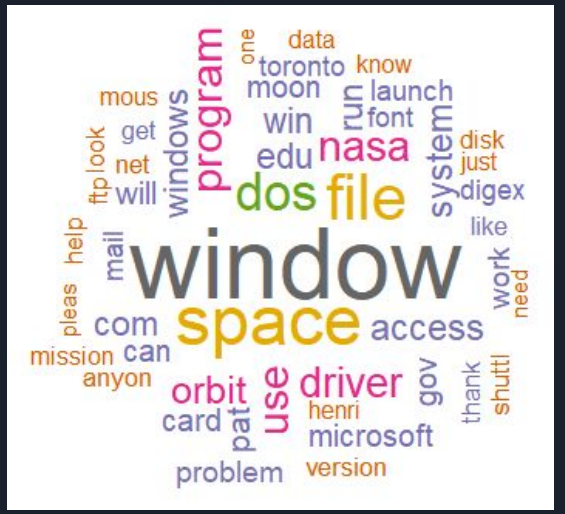


Ethic

20 NEWSGROUPS: WORLD CLOUDS, COM K=4

As categorias *talk.politics.guns* e *sci.med sci.med* foram unidas em uma categoria: *Treatment*. Além disso, as duas categorias *comp.os.ms-windows.misc* e *sci.space* foram unidas em uma categoria: *Astronautics*.

Treatment



Astronautics

SOM (Self-Organizing Maps)

SOM

Estratégia de testes

- Teste de parâmetros;
- Seleção de conjuntos;
- Script com testes padronizados.

SOM

Parâmetros fixos:

- Número de épocas: 150;
- Taxa de aprendizagem inicial: 1;
- Função de atualização da taxa de aprendizado: linear;
- Função de atualização de pesos: Gaussiana;
- Função de raio de vizinhança: decrescente.

SOM

Variação de parâmetros:

- Tamanho do espaço matricial: 10, 12, 15, 20, 25 e 30;
- Raio de vizinhança inicial: 1 até (tamanho do espaço matricial)/5.

SOM

Visualização dos resultados

- Matriz de categorias;
- Matriz U;
- Erros de quantização e distorção topológica por época.

SOM - BBC

Médias de erros dos conjuntos testados - BBC						
	85		90		95	
	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão
ErrorQBIN	3,669	0,100	5,076	0,103	6,779	0,127
ErrorQTF	7,058	0,202	9,433	0,254	12,347	0,294
ErrorQTFIDF	0,067	0,002	0,112	0,003	0,184	0,004
ErrorTBIN	0,157	0,092	0,137	0,087	0,123	0,073
ErrorTTF	0,212	0,094	0,204	0,099	0,198	0,086
ErrorTTFIDF	0,246	0,112	0,223	0,092	0,203	0,072

SOM - BBC

Matriz de categorias (20x20 e raio de vizinhança 3 - TF-IDF)

business(1)

politics (3)

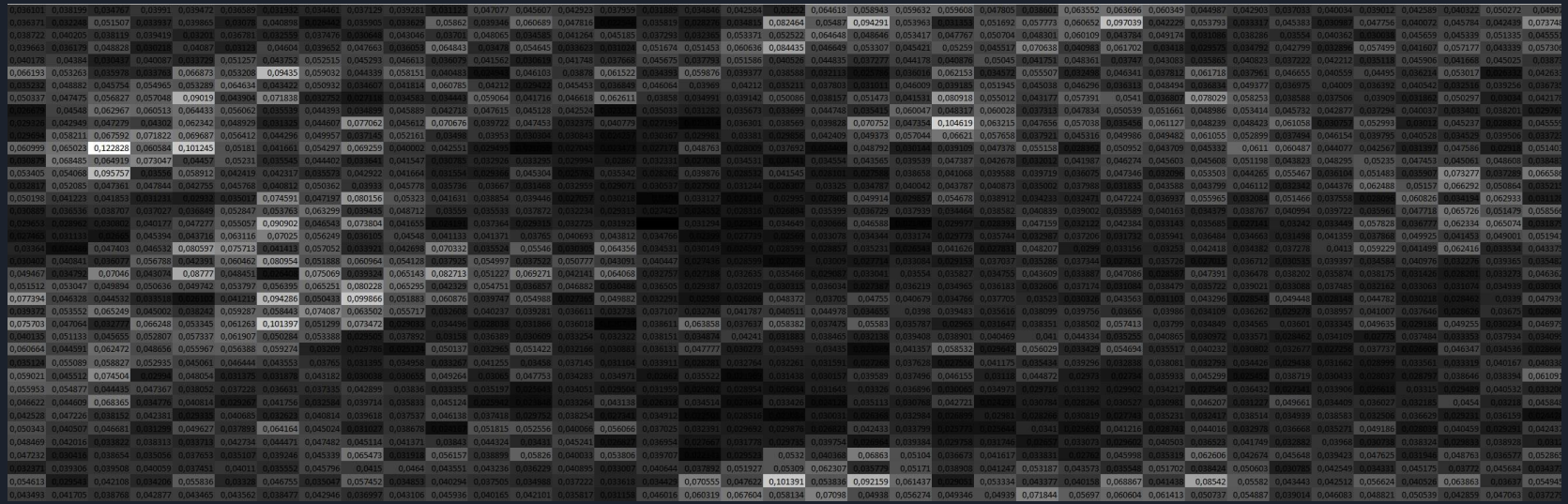
tech(5)

entertainment(2)

sports(4)

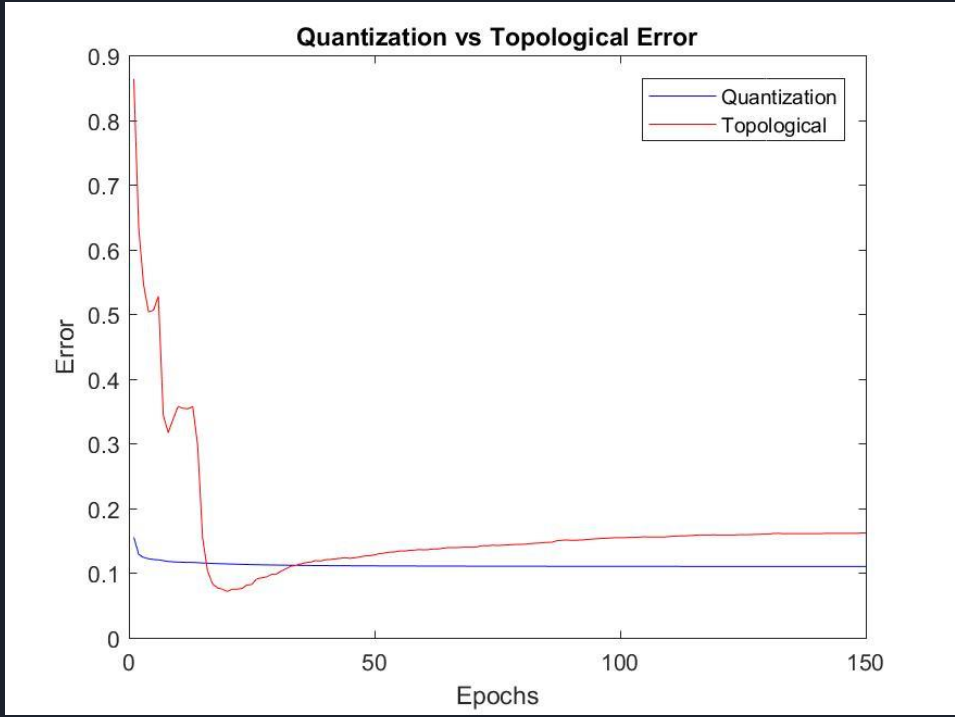
SOM - BBC

Matriz U (20x20 e raio de vizinhança 3)



SOM - BBC

Erros de quantização e distorção topológica

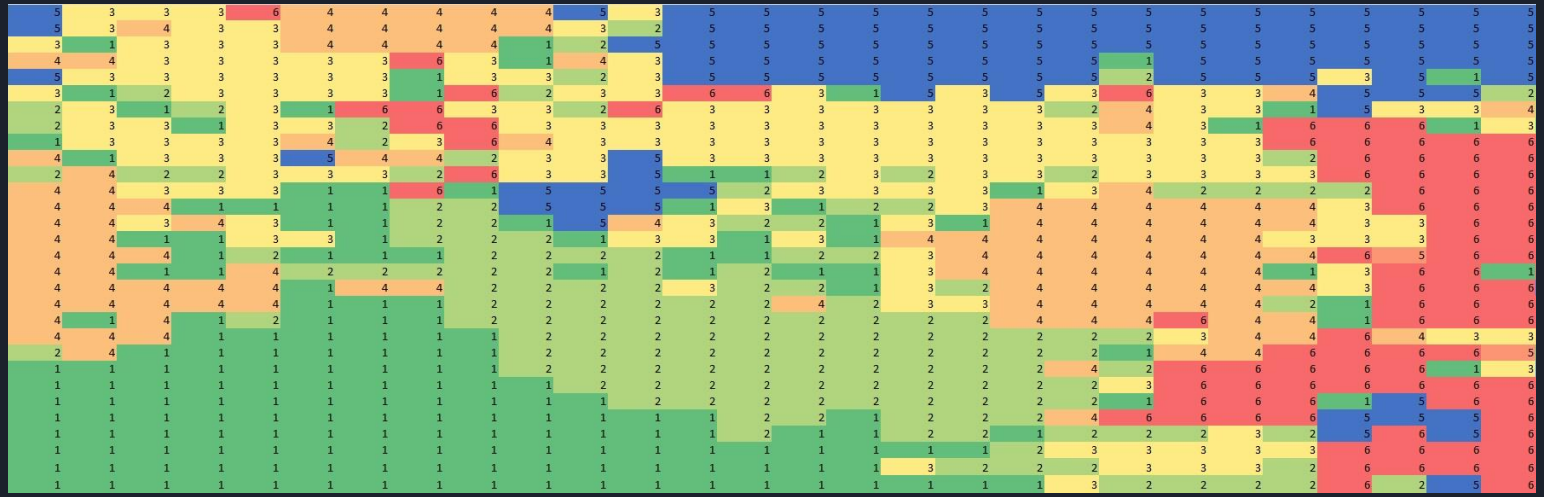


SOM - 20 News Groups

Médias de erros dos conjuntos testados - 20 news groups						
	85		90		95	
	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão
ErrorQBIN	2,212	0,079	3,055	0,067	4,532	0,077
ErrorQTF	3,958	0,129	5,251	0,139	7,501	0,177
ErrorQTFIDF	0,060	0,002	0,099	0,003	0,187	0,004
ErrorTBIN	0,290	0,106	0,280	0,113	0,252	0,108
ErrorTTF	0,309	0,091	0,296	0,092	0,280	0,092
ErrorTTFIDF	0,386	0,114	0,395	0,108	0,370	0,105

SOM - 20 News Groups

Matriz de categorias (30x30 e raio de vizinhança 4 - TF-IDF)



comp.os.ms-windows.misc (1)

sci.med (3)

soc.religion.christian (5)

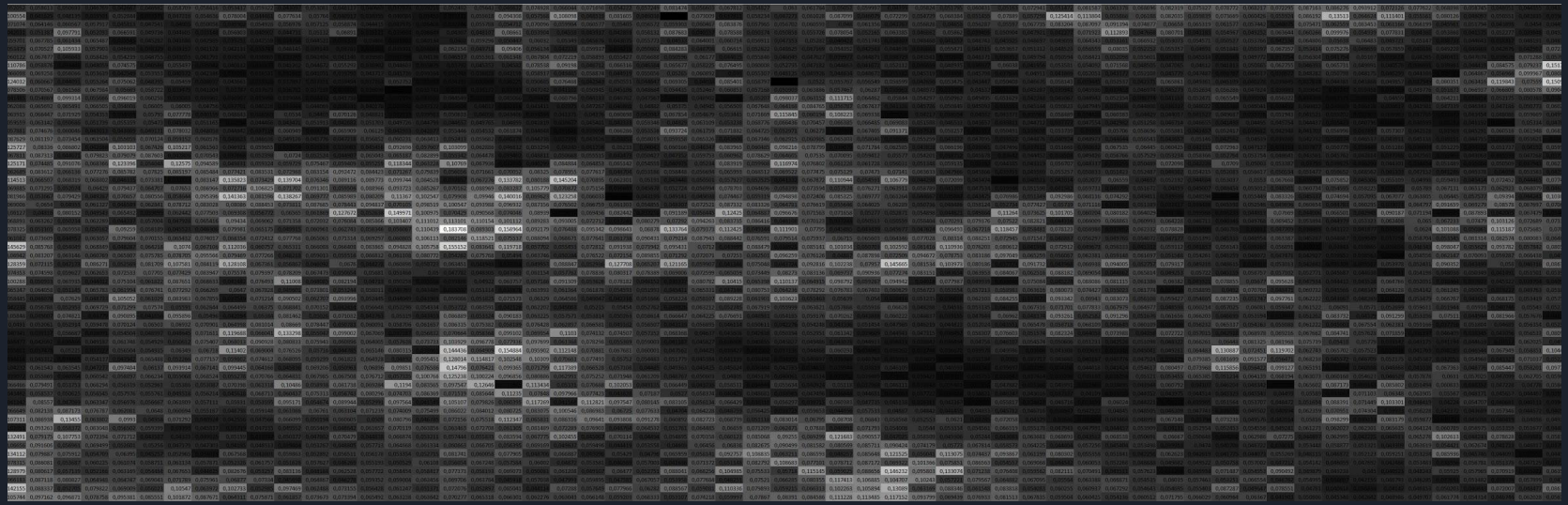
rec.autos (2)

sci.space (4)

talk.politics.guns (6)

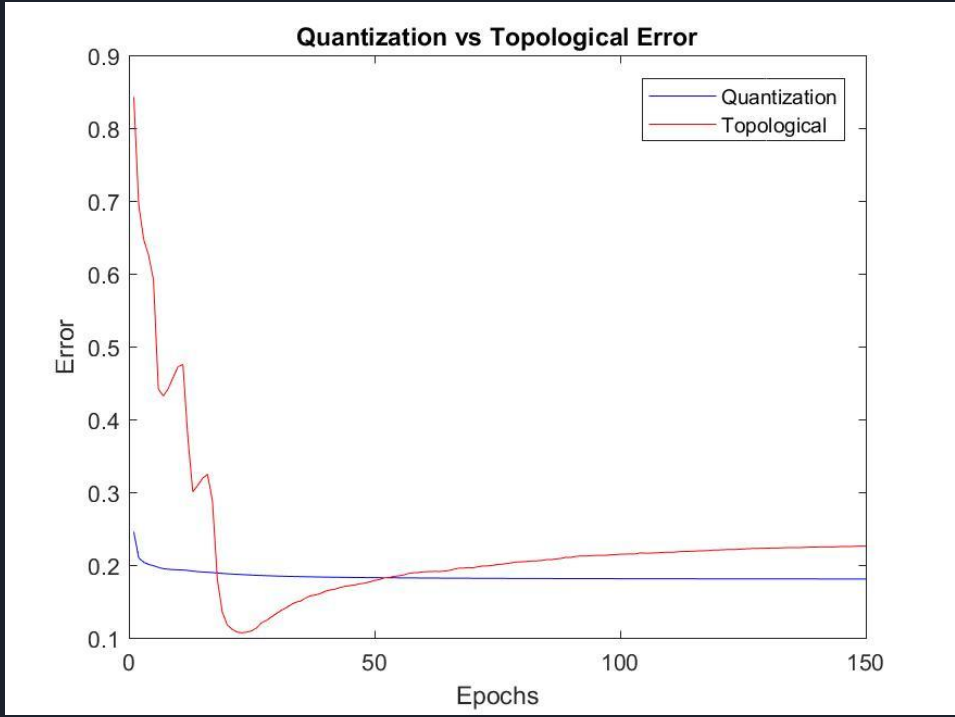
SOM - 20 News Groups

Matriz U (30x30 e raio de vizinhança 4)



SOM - 20 News Groups

Erros de quantização e distorção topológica



SOM

- Com mais dimensões, o erro de quantização cresce e o de distorção topológica decresce, em média;

- Com a matriz de categorias é possível visualizar o agrupamento no espaço matricial;

- Raios de vizinhança menores (1 e 2) produzem resultados piores;

- Quanto maior o número de dimensões, mais neurônios são necessários para visualizar bons resultados.

CONCLUSÃO

- Para menos dimensões, K-means e K-means++ apresentam desempenho superior em termos computacionais, com baixa perda de precisão. Apesar de apresentar resultados aceitáveis para poucas dimensões, o SOM é mais lento;

- A sensibilidade do BIC score para muitas dimensões faz com que o X-means apresente resultados menos satisfatórios nessas condições; é notável a diferença de perfomance em todos os algoritmos com o aumento das dimensões.

CONCLUSÃO

- Para todos os algoritmos (K-means, K-means++, X-means e SOM), a representação que gerou melhores resultados foi a TF-IDF, seguida pela Binária e, por último, a TF.

- Índice Silhouette foi substituído por outras medidas avaliativas, por gerar resultados não muito satisfatórios. No caso do X-means, foi utilizado o BIC score, e no caso de avaliação de qualidade de agrupamentos, foram utilizada a tabela de precisão média.

REFERÊNCIAS

- D. Arthur and S. Vassilvitskii, “K-means++: The advantages of careful seeding,” in SODA ’07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 2007, pp. 1027–1035.
- D. Pelleg and A. W. Moore, “X-means: Extending k-means with efficient estimation of the number of clusters,” in ICML ’00 Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2000, pp. 727–734.
- A. Trevino. (2016, Dec.) Introduction to k-means clustering. [Online]. Available: https://www.datascience.com/blog/k-means-clustering
- L. Maaten and G. Hinton, “Visualizing data using t-sne,” in Journal of Machine Learning Research - Vol. 9. JMLR, 2008, pp. 2579–2605.
- E. Rendón, I. Abundez, A. Arizmendi and E. Quiroz, “Internal versus External cluster validation indexes”, in International Journal of Computers and Communications - Vol. 5, 2011, pp. 27-34.
- W. Natita, W. Wiboonsak, and S. Dusadee, "Appropriate Learning Rate and Neighborhood Function of Self-organizing Map (SOM) for Specific Humidity Pattern Classification over Southern Thailand", in International Journal of Modeling and Optimization - Vol. 6, No. 1, 2016, pp. 61-65.

OBRIGADO!