Trabalho Individual 1

David Burth Kurka*

2 de outubro de 2013

1 Introdução

O objetivo deste trabalho é classificar um grande conjunto de documentos de texto, de forma não supervisionada.

O conjunto de documentos é formado por 18.828 textos de emails e sabe-se que podem ser agrupados em algum número entre 1 e 100 grupos. É parte do trabalho descobrir o número de grupos adequado.

Este relatório está dividido da seguinte forma: Na sessão 2 descreve-se como foi feita a extração de características; Na sessão 3, é descrito o processo para escolha do número de grupos dos dados; A sessão 4 descreve os resultados da classificação geral, apoiados pelos dados detalhados no apêndice 5; Finalmente, em 5 são trabalhadas as conclusões com o trabalho.

2 Extração de Características

Aqui descreveremos como foi feita a seleção de características, a partir dos 18.828 documentos trabalhados. Este talvez seja o passo mais importante do processo, já que a extração de características adequada permitirá a identificação correta dos documentos da coleção e assim viabilizará o agrupamento de documentos semelhantes. Busca-se também um número adequado de características, para que o problema se torne tratável.

Para a extração de características foi utilizada a biblioteca em R *tm.plugin.dc*[1], extensão da biblioteca de mineração de textos *tm*, capaz de manipular documentos de textos, fazer operações e construir matrizes de termos de documentos, de forma distribuida, sendo portanta apropriada para o montante de arquivos trabalhados. As características extraidas nesse projeto são todas baseadas nas palavras individuais existentes em cada documento. Cada documento portanto, é caracterizado por um conjunto de frequências de termos, que serão descritos a seguir.

A primeira operação consiste em extrair todos os documentos do diretório e carregar seus termos no *corpus*. Em seguida, são feitas diversas operações no

*RA: 070589, email: david.kurka@gmail.com

[1]Disponível em `http://cran.r-project.org/web/packages/tm.plugin.dc/`

conteúdo do documento, com a finalidade de excluir informações não relevantes para a classificação, dos mesmos. As seguintes operações são realizadas no *corpus*:

- Remoção de caractéres de pontuação, números e espaços em branco;
- Transformação de caractéres em caixa alta, para caixa baixa;
- Remoção de *stopwords* (palavras frequentes, sem valor semântico);
- Remoção de derivações de palavras (*stemming*)

Feito este processo, obtemos um conjunto de 127.620 termos distintos, em todos os documentos. Então, foi procedido a construção de das matrizes de termos dos documentos, agregando todos as características de todos os documentos. Nesta etapa, termos com menos de 2 caractéres foram eliminados, assim como termos que aparecem menos de 10 ou mais que 10.000 vezes em todos os documentos. Essa última restrição ocorre pois, como o objetivo é agrupar grupos distintos de mensagens, não desejamos termos que estejam presentes em muitos grupos, nem em apenas pequena fração dentro de um grupo.

Finalmente, efetuamos a ponderação e a normalização das características. Para a ponderação, cada termo foi ponderado pela "frequência inversa dos documento" (IDF), que dá menos valor para termos presentes em muitos documentos e mais valor para termos presentes em poucos, aumentando o poder de discriminação das palavras [1]. A normalização faz com que o módulo do vetor de características de cada documento seja unitário, permitindo assim a comparação entre documentos de tamanhos diferentes.

Assim, concluimos o processo de extração de características, obtendo uma matriz de termos com 18.828 linhas (correspondentes aos 18.828 documentos) e com (apenas) 1.637 colunas, correspondentes às características extraídas, reduzidas e normalizadas.

3 Escolha do número de grupos (K)

Outro fator importante para a clusterização é a definição de quantos grupos distintos K existem nos dados. Essa definição pode ser um pouco subjetiva, pois podem ser interpretadas diversos critérios para agrupar documentos, gerando valores variados de K. Entretanto, buscaremos aqui, um valor entre 1 e 100 (definido pela proposta), que consiga encontrar um equilibro entre representatividade por cluster e baixa disperção.

Diferentes métodos para identificar o K foram explorados. Em todos eles, utilizou-se um subconjunto de 1000 documentos, retirados aleatoriamente do conjunto global, para permitir maior eficiência no processamento.

3.1 Método do "cotovelo"

A primeira estratégia utilizada para identificar o valor de K foi o chamado método do "cotovelo", no qual é calculada uma função de erro para agrupamentos

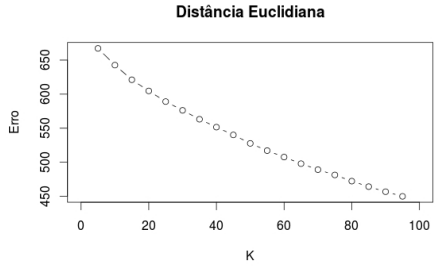


Figura 1: Curva erro versus número de nós, utilizando k-means com distância euclidiana. Note que não é possível identificar o "cotovelo"da figura.

com valores variáveis de K e espera-se encontrar um momento em que o aumento de grupos diminua o erro em uma taxa menor que o observado até então. Esse efeito é identicado como um "cotovelo" no gráfico $erro \times K$. A função de erro é definida como a soma das dispersões de todos os agrupamentos.

Na primeira tentativa utilizou-se o algoritmo da biblioteca padrão k-means, com o cálculo de distância euclidiana. Foram calculados agrupamentos para k variando de 5 à 95, sendo o erro calculado como a média de três execuções. Entretanto, como se pode ver na figura 1, o gráfico não apresenta diferenças de angulação em seus pontos, não sendo possível observar os desejados "cotovelos".

Como solução para esse impasse, utilizou-se os resultados de Strehl et al. [2], que observam que medidas de distância de cosseno e jaccard extendido obtêm melhor resultados no agrupamento de textos. Dessa forma, foi utilizado o algoritmo de agrupamento genérico *pam*[2], e calculado o k-medoides utilizando as distâncias acima, variando de 5 à 95, com o erro dado pela média de três execuções.

As figuras 2 e 3 mostram os resultados obtidos. Apesar da interpretação do gráfico também dar margem para múltiplas interpretações, notamos que para $K = 30$, há um ponto de inflexão, que podemos considerar como o cotovelo.

3.2 Agrupamento hierárquico

Para avaliar e ter mais embasamento para a escolha do k, também foi realizado o agrupamento hierárquico dos 1000 documentos amostrados e analisadas suas propriedades.

Utilizando a função *hclust*, geramos o dendograma, da figura 4.

Fazendo um corte no dendograma, para $K = 30$, como o sugerido na regra do cotovelo, vemos que os dados se distribuem de forma não uniforme, havendo concentrações de documentos em alguns grupos, enquanto outros grupos possuem poucos documentos. O histograma da distribuição de documentos, por nós pode ser visto na figura 5.

[2]disponível em `http://cran.r-project.org/web/packages/cluster/`

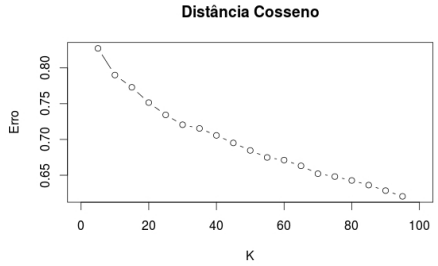


Figura 2: Curva calculada a partir de k-medoides, com cosseno. Note que há pontos de inflexão para k=30 e k=35.

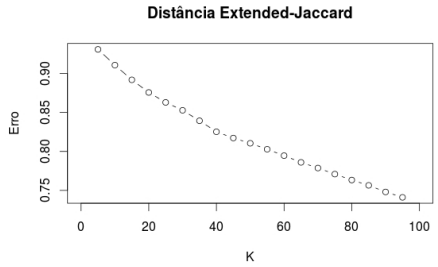


Figura 3: Curva calculada a partir de k-medoides, com calculo de distância jaccard extendido. Note que há pontos de inflexão para k=30.

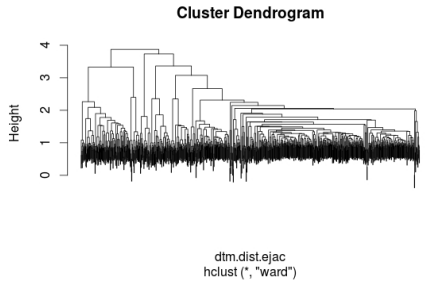


Figura 4: Dendograma do agrupamento hierárquico da amostra.

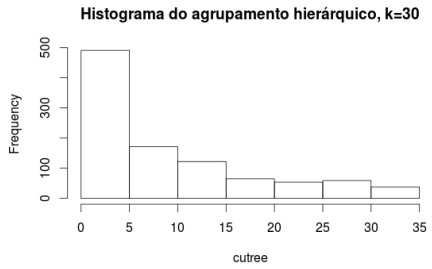


Figura 5: Histograma dos elementos por agrupamento, do agrupamento hierárquico, no corte k=30.

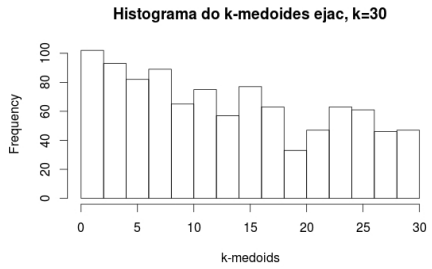


Figura 6: Histograma com a distribuição de elementos para k-medoides, com k=30.

Apesar de não haver garantias de que os agrupamentos tenham números similares de elementos, havia essa expectativa, de modo que o resultado do agrupamento hierárquico pareceu negativo. Porém, foi feita a mesma análise de número de individuos por grupo com o resultado do k-medoides, onde foi obtido o histograma da figura 6, que apresenta um equilíbrio maior entre os elementos dos grupos e nos permite manter a escolha de $K = 30$.

4 K-means geral

Finalmente, definido K como 30, podemos finalmente agrupar todos os documentos e verificar o resultado.

Como na extração de características, conseguimos reduzir o vetor de característica para apenas 1.637 dimensões, podemos empregar algoritmos padrões de clusterização, disponíveis no repositório de R[3]. Utilizou-se portanto, a função

[3]Chegou a ser implementada uma versão própria do k-means, capaz de lidar com dimensões maiores que o padrão. Entretanto, como as implementações padrões são mais eficientes, optou-se por usá-las. O código desenvolvido, porém, encontra-se no projeto, como a obsoleta função *scalable-kmeans()*

kcca da biblioteca *flexclust*[4],

Seguindo os princípios usados na determinação de k, foi usado o algoritmo k-means, com distâncias calculadas por jaccard extendido, com K=30. Os resultados obtidos foram:

- Quantidade de elementos por cluster:

Cluster	1	2	3	4	5	6	7	8	9	10
#elem	633	486	376	990	513	520	1188	868	626	535
Cluster	11	12	13	14	15	16	17	18	19	20
#elem	968	876	485	484	793	822	809	460	826	423
Cluster	21	22	23	24	25	26	27	28	29	30
#elem	823	309	416	418	543	632	289	374	742	601

- Erro (divergência) por ponto

Cluster	1	2	3	4	5
Erro	627.6160	481.2034	372.9201	976.1856	507.1914
Cluster	6	7	8	9	10
Erro	514.3976	1172.3677	857.9733	617.1381	527.4032
Cluster	11	12	13	14	15
Erro	957.3333	861.5899	478.5700	478.0657	783.7111
Cluster	16	17	18	19	20
Erro	806.8532	799.1256	453.2202	812.5462	417.9416
Cluster	21	22	23	24	25
Erro	809.3983	305.3916	411.5809	411.6316	535.0430
Cluster	26	27	28	29	30
Erro	625.4407	284.1693	368.3567	734.9063	594.7018

- Divergência média = 619.4658 ± 222.2271
- Divergência total = 18583.97

Porém, mais importante que a análise quantitativa é a análise qualitativa dos clusters. No sessão anexa 5 encontram-se os centroides e os três documentos mais próximos deles. Além disso, para cada cluster foi elaborada uma nuvem de palavras (wordcloud), com os termos presentes no mesmo. Este recurso nos permite fazer observações e conclusões muito interessantes.

É possível observar que a clusterização conseguiu separar bem alguns documentos. Por exemplo, o agrupamento 1 claramente trata de temas ligados à saúde, enquanto o agrupamento 29 trata de assuntos relacionados à astronomia. Outros temas identificados foram homossexualidade (24), ciclismo (25), literatura (23), carros (15), hardware (13), jogos (7), política (3), entre outros.

Porém, também observa-se que há temas associados, como nos agrupamentos 6 e 28, que tem muitos termos relacionados à etnias e nacionalidades, ou nos

[4]disponível em `http://cran.r-project.org/web/packages/flexclust/flexclust.pdf`

agrupamentos 17 e 30, que tratam aparentemente de cristianismo. Nesse caso, pode ser que se tratam de subdivisões dentro do tema (como grupos de países, ou religiões diferentes), gerando vocabulários ligeiramente diferentes, apesar de dentro do mesmo tema.

5 Conclusão

Este trabalho explorou a manipulação de documentos de texto, com a finalidade de agrupar textos relacionados.

Percebeu-se que as etapas que necessitaram maior processamento intelectual e de máquina foram as etapas de extração de características e a etapa de definição do número de clusters.

Os desafios encontrados relacionaram-se com a capacidade computacional de lidar com tantos arquivos e dados, a imprecisão dos métodos para definição do número de agrupamentos K e adaptação à linguagem de programação R.

Porém, transpassados os desafios, pode ser feito o agrupamento com todos os dados e obteve-se resultados interessantes, como os mostrados na última sessão. Entretanto, acredita-se que ainda possam ser feitos ajustes em todas as etapas, para se conseguir resultados mais precisos e grupos mais detalhados.

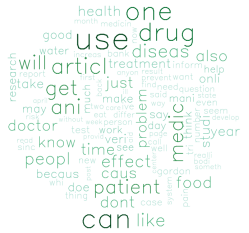
A Apêndice - Centroides e nuvens de palavras

A seguir estão representados, para cada agrupamento, o nome do arquivo centroide do grupo, o *"subject"* do documento e as mesmas informações para os três vizinhos mais próximos.

Além disso, cada cluster possui uma "nuvem de palavras", imagem na qual palavras com mais frequência no cluster aparecem maiores, enquanto palavras de pouca frequência, menores. Esse artefato permite a visualização empírica das características do cluster.

A.1 Cluster 1

Nuvem de palavras:



Centróide:

```
Arquivo: "e46bb9ce30050f14175c04b652be5344.txt"
Subject: Re: Alarm systems: are they worthwhile?
```

Vizinho mais próximo #1:

```
Arquivo: "c23d3914f7ab6def8321f994f043fc11.txt"
Subject: Re:   Shaft-drives and Wheelies
```

Vizinho mais próximo #2:

```
Arquivo: "33b7786bbf7cd7225695812377b92d1d.txt"
Subject: ADCOM GTP500II IR sensor & repeater spec's?
```

Vizinho mais próximo #3:

```
Arquivo: "20e022bf19f3794f09546968eec815fe.txt"
Subject: Problems with HP Backgrounder-- Help!!
```

A.2 Cluster 2

Nuvem de palavras:

Centróide:

```
Arquivo: "33807a74b1cd875880d769ef1623afb8.txt"
Subject: Image Analysis for PC
```

Vizinho mais próximo #1:

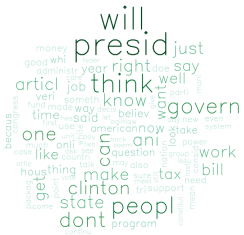
```
Arquivo: "739f5efbaa2a9623ed77312de6c1a1d2.txt"
Subject: VideoBlaster & PC SPEAKER
```

Vizinho mais próximo #2:

```
Arquivo: "16c890815ba81e5cde1f209c02700c5c.txt"
Subject: Magellan Update - 04/16/93
```

Vizinho mais próximo #3:

```
Arquivo: "0b192837dcc6ef1b78c16dc5464b645f.txt"
Subject: Re: islamic authority over women
```



A.3 Cluster 3

Nuvem de palavras:
Centróide:

```
Arquivo: "16f72ff66bc26e0f748282b93c4e6290.txt"
Subject: Re: Permanent Swap File
```

Vizinho mais próximo #1:

```
Arquivo: "5054b8b0b34441c0b93c84ee3f0e7820.txt"
Subject: Re: POV previewer
```

Vizinho mais próximo #2:

```
Arquivo: "fe7a6534b9aeecc68f4055619899a7a7.txt"
Subject: Re: USA McWeekly Stats
```

Vizinho mais próximo #3:

```
Arquivo: "270f0ee168c411dd9c1dbbc3f72d83d0.txt"
Subject: Re: Booting from B drive
```

A.4 Cluster 4

Nuvem de palavras:

Centróide:

```
Arquivo: "38bfd558f55c5d0a0a35ce11e79ad3ca.txt"
Subject: Headphones - AKG 340s ($200) For Sale
```

Vizinho mais próximo #1:

```
Arquivo: "a53c2bebe12f68d1f5bb95009c442276.txt"
Subject: Re: Thoughts on a 1982 Yamaha Seca Turbo?
```

Vizinho mais próximo #2:

```
Arquivo: "804c784ad68cb51523917f777b4165c6.txt"
Subject: ADB Mouse II (ergo) -- when?
```

Vizinho mais próximo #3:

```
Arquivo: "c29a79e20098fb1616e1c2bbba5aac18.txt"
Subject: Conference on Manned Lunar Exploration.  May 7  Crystal City
```

A.5 Cluster 5

Nuvem de palavras:

Centróide:

```
Arquivo: "d8b7c95887019aac8f1eb48715dbb072.txt"
Subject: Re: Tie Breaker....(Isles and Devils)
```

Vizinho mais próximo #1:

```
Arquivo: "d85e0b3f25bb0fb1e039f99ea01d94e7.txt"
Subject: Re: Kind, loving, merciful and forgiving GOD!
```

Vizinho mais próximo #2:

```
Arquivo: "4d59caa38bfe6c7b519ebd29e90bf917.txt"
Subject: RFI:Art of clutchless shifting
```

Vizinho mais próximo #3:

```
Arquivo: "83e84d8c78affd0f147f03f5177f6d85.txt"
Subject: Re: How does a pitcher get a save?
```

A.6 Cluster 6

Nuvem de palavras:
Centróide:

```
Arquivo: "f7681b48981539e171826d126c9e48ee.txt"
Subject: Re: Volume
```

Vizinho mais próximo #1:

```
Arquivo: "11eb419f4de8be9d7813a52e1e5b1a0d.txt"
Subject: Re: Who's next? Mormons and Jews?
```

Vizinho mais próximo #2:

```
Arquivo: "ee4c97a95b001a4850aea8064b87c6f5.txt"
Subject: Re: Being right about messiahs
```

Vizinho mais próximo #3:

```
Arquivo: "b1077ceaa7109d77feb327ef429622cd.txt"
Subject: Re: Clipper considered harmful
```

A.7 Cluster 7

Nuvem de palavras:

Centróide:

```
Arquivo: "aacdd4b132ac68bfea86b7eff8cad710.txt"
Subject: Re: Is "Kermit" available for Windows 3.0/3.1?
```

Vizinho mais próximo #1:

```
Arquivo: "23f0c291edb96ed337f84d5dd0bdd9aa.txt"
Subject: Re: SATANIC TOUNGES
```

Vizinho mais próximo #2:

```
Arquivo: "75fc2a5644d62871721a68fb9a1298bf.txt"
Subject: bike for sale in MA, USA
```

Vizinho mais próximo #3:

```
Arquivo: "2f8f254b36512ab319f0d1a967f44954.txt"
Subject: Re: Non-lethal alternatives to handguns?
```

A.8 Cluster 8

Nuvem de palavras:

Centróide:

```
Arquivo: "7ae6a52b787777a141086e98eaf1966b.txt"
Subject: Pseudocollisions in MD5
```

Vizinho mais próximo #1:

```
Arquivo: "70f8b1847df15f4bfb490762325f427e.txt"
Subject: Widget source code needed
```

Vizinho mais próximo #2:

```
Arquivo: "f34bf2b326b22daf25b5741635b23464.txt"
Subject: Re: Turkey-Cyprus-Bosnia-Serbia-Greece (Armenia-Azeris)
```

Vizinho mais próximo #3:

```
Arquivo: "69d715e9e63a420a88f8f22b68497574.txt"
Subject: Re: VLB bus master problem?
```

A.9 Cluster 9

Nuvem de palavras:
Centróide:

```
Arquivo: "87747261d7778ed44944c688c3906147.txt"
Subject: Re: japanese moon landing?
```

Vizinho mais próximo #1:

```
Arquivo: "a87f5771b07cb8f2fd59786347232568.txt"
Subject: Re: Encapsulated Postscript and X
```

Vizinho mais próximo #2:

```
Arquivo: "3c50b509dfc0a4b6dede4394e4b7562d.txt"
Subject: Where to get ATI card video drivers/fonts?
```

Vizinho mais próximo #3:

```
Arquivo: "f89db2f81a0bdec0e4365d9c0279a3fc.txt"
Subject: Re: "Accepting Jeesus in your heart..."
```

A.10 Cluster 10

Nuvem de palavras:

Centróide:

```
Arquivo: "7343b1eeb5d02b510794215e981ecb95.txt"
Subject: Re: STRONG & weak Atheism
```

Vizinho mais próximo #1:

```
Arquivo: "888a386f47298eb3e83e880e6c26609a.txt"
Subject: Re: Taurus/Sable rotor recall
```

Vizinho mais próximo #2:

```
Arquivo: "939cae13a68e6ee9269f50b2d9f39f84.txt"
Subject: DON'T BUY FROM T.C. COMPUTERS! ! !
```

Vizinho mais próximo #3:

```
Arquivo: "88a2f4e15d08fdfa24c14dcfced1b57d.txt"
Subject: Re: Highlights
```

A.11 Cluster 11

Nuvem de palavras:

Centróide:

```
Arquivo: "ac98113f5178cdba26aa0331028d89f4.txt"
Subject: Re: Pink Noise
```

Vizinho mais próximo #1:

```
Arquivo: "f337526cdb2f6207973d7722fab9ede9.txt"
Subject: Re: Nature of God (Re: Environmentalism and paganism)
```

Vizinho mais próximo #2:

```
Arquivo: "04333055cf40207af092c94e0bb7a703.txt"
Subject: Re: Boom!  Whoosh......
```

Vizinho mais próximo #3:

```
Arquivo: "1cc8bf4a17a9f0658e75a0e09bcdf55c.txt"
Subject: ### 68040 25Mz FOR SALE : ABSOLUTELY NEVER USED ###
```

A.12 Cluster 12

Nuvem de palavras:
Centróide:

```
Arquivo: "1c8e89c0e399c648e0a27064a40fb119.txt"
Subject: Ohio Legislative Alert -- H.B. 278
```

Vizinho mais próximo #1:

```
Arquivo: "7b87a361fd08184648a39a311fcfbaee.txt"
Subject: Re: Hockey guest spots...
```

Vizinho mais próximo #2:

```
Arquivo: "f0e65c72ad3446b948ff657b8723e476.txt"
Subject: Re: BMW heated grips
```

Vizinho mais próximo #3:

```
Arquivo: "b557274062592697df8b6ef3fa06a657.txt"
Subject: Re: gadgets vs widgets
```

A.13 Cluster 13

Nuvem de palavras:

Centróide:

```
Arquivo: "b7810259c821ef0cab4ac219d0545e0e.txt"
Subject: Re: Atheists and Hell
```

Vizinho mais próximo #1:

```
Arquivo: "c4e6e93cd0f3f52660ef7a988b96cb7a.txt"
Subject: Re: Sandberg, Runs, RBIs (was: Re: Notes on Jays vs. Indians Series)
```

Vizinho mais próximo #2:

```
Arquivo: "ff31a76da136d1213e4bdc1fb2953593.txt"
Subject: Re: IDE vs SCSI
```

Vizinho mais próximo #3:

```
Arquivo: "d4c04477dc67a4bd4f96809c64eef133.txt"
Subject: Looking for drawing packages
```

A.14 Cluster 14

Nuvem de palavras:

Centróide:

```
Arquivo: "2321c14b1a03a2803cfac73115b603c0.txt"
Subject: Pregnency without sex?
```

Vizinho mais próximo #1:

```
Arquivo: "cf5432269571a4a4e50d0041bd1707b4.txt"
Subject: Re: So Far , So Good (THE RED SOX)
```

Vizinho mais próximo #2:

```
Arquivo: "5948471ad83d654cace03ff0a1e5bb47.txt"
Subject: ***Wanted : 386DX-33 motherboard
```

Vizinho mais próximo #3:

```
Arquivo: "6a748a5fa98b2b244155f6fed7e6ac5f.txt"
Subject: Re: Deuterocanicals, eps. Sirach
```

A.15 Cluster 15

Nuvem de palavras:
Centróide:

```
Arquivo: "415ce617495178681f342417d8563986.txt"
Subject: Reboot when I start windows.
```

Vizinho mais próximo #1:

```
Arquivo: "84a9779a27a0615443f232306ff845d3.txt"
Subject: Re: Slavery (was Re: Why is sex only allowed in marriage: ...)
```

Vizinho mais próximo #2:

```
Arquivo: "800314fe97008e43c8e2b25f9144e8cc.txt"
Subject: Re: Oops! Oh no!
```

Vizinho mais próximo #3:

```
Arquivo: "2611baa186a7ba22642ce058ed332804.txt"
Subject: Re: Identify this bike for me
```

A.16 Cluster 16

Nuvem de palavras:

Centróide:

```
Arquivo: "507529c9e24e1c1c77af448e93bf7f59.txt"
Subject: Clinton caves in: reduces jobs bill
```

Vizinho mais próximo #1:

```
Arquivo: "b1986e33939535ec6a46e68af84c71ea.txt"
Subject: Re: Blinking Cursor in Xterm???
```

Vizinho mais próximo #2:

```
Arquivo: "71a511b4de1339b5d42bdbeb8d090d54.txt"
Subject: Problems with Open Windows
```

Vizinho mais próximo #3:

```
Arquivo: "fbeb03049ed58b22694b0befb44ea0df.txt"
Subject: disp140 [0/7]
```

A.17 Cluster 17

Nuvem de palavras:

Centróide:

```
Arquivo: "506378ab551f5090d1c3d21b8c3ef22e.txt"
Subject: Re: intolerance - eternal life - etc
```

Vizinho mais próximo #1:

```
Arquivo: "a27ed78a5c5193d635f1cbcd810a466a.txt"
Subject: Re: Secret algorithm [Re: Clipper Chip and crypto key-escrow]
```

Vizinho mais próximo #2:

```
Arquivo: "aca92095ae5e039098a67bb0131c4951.txt"
Subject: help
```

Vizinho mais próximo #3:

```
Arquivo: "4cc2a44472457caf344065922a7fdfbf.txt"
Subject: Forsale: Drum things.. CHEAP!
```

A.18 Cluster 18

Nuvem de palavras:
Centróide:

```
Arquivo: "b404b3d09d9fbefdf48999f639b606ce.txt"
Subject: Re: Fractals? what good are they?
```

Vizinho mais próximo #1:

```
Arquivo: "e740a9c7881ad21f89deebb925d12de3.txt"
Subject: Re: vitamin A and hearing loss
```

Vizinho mais próximo #2:

```
Arquivo: "167501f9e105e814abb9ab58a8fe0599.txt"
Subject: Re: help - how to construct home-built battery for 3rd grade sci report
```

Vizinho mais próximo #3:

```
Arquivo: "372696f620606ff0781ea53d6c5c6fce.txt"
Subject: Re: Changing oil by self.
```

A.19 Cluster 19

Nuvem de palavras:

Centróide:

```
Arquivo: "3b176e071233d390da3033f7fff36ec8.txt"
Subject: Sun 4.1.3, OpenWindows 3.0 problem: static linking and X libraries
```

Vizinho mais próximo #1:

```
Arquivo: "3f4d4ffde871acd32607fca5c05cdd5c.txt"
Subject: Re: FTP PC/TCP ver 2.04 FOR SALE cheap
```

Vizinho mais próximo #2:

```
Arquivo: "2533d723efa9441a4fedaf23a95b4331.txt"
Subject: help: Splitting a trimming region along a mesh
```

Vizinho mais próximo #3:

```
Arquivo: "d934eb44a7bbd2cd40a5cf5b55b60dfd.txt"
Subject: Re: 386 Motherboard advice needed
```

A.20 Cluster 20

Nuvem de palavras:

Centróide:

```
Arquivo: "ae7e02c9523154763e902a3d48e306b9.txt"
Subject: Re: Rosicrucian Order(s) ?!
```

Vizinho mais próximo #1:

```
Arquivo: "c7557c8479ae371f857760859b4db04f.txt"
Subject: Re: From Israeli press. Madness.
```

Vizinho mais próximo #2:

```
Arquivo: "b000a71f6586984a320947f48aa34448.txt"
Subject: 4X4 On/Off-Road Rally - Joliet Il.
```

Vizinho mais próximo #3:

```
Arquivo: "5734785b2becdbde7d48e9ed405366e8.txt"
Subject: Re: Blood Cholesterol -  Gabe Mirkin's advice
```

A.21 Cluster 21

Nuvem de palavras:
Centróide:

```
Arquivo: "1325c72eaccc78bad80e3d53f80651d9.txt"
Subject: Re: To All My Friends on T.P.M., I send Greetings
```

Vizinho mais próximo #1:

```
Arquivo: "78ca53c878372c2fc2aae79837aa2b48.txt"
Subject: Re: It's a rush... (was Re: Too fast)
```

Vizinho mais próximo #2:

```
Arquivo: "e43edf0ae546769544428ef4bed8d47f.txt"
Subject: Pointer..Xlib
```

Vizinho mais próximo #3:

```
Arquivo: "ec4e0086ab8ef5c619f3665d00853cca.txt"
Subject: Re: Public Service Translation No.2
```

A.22 Cluster 22

Nuvem de palavras:

Centróide:

```
Arquivo: "688dbb722313c47eb32311629cbb2013.txt"
Subject: Olivetti XT
```

Vizinho mais próximo #1:

```
Arquivo: "d59b21644bc408285ce8b20475f2c025.txt"
Subject: Re: prayers and advice requested on family problem
```

Vizinho mais próximo #2:

```
Arquivo: "f7309e862ff03d4189182710f18e58f3.txt"
Subject: Re: looking for hot Mac 3D anim software
```

Vizinho mais próximo #3:

```
Arquivo: "824506046ee8b233f36cadc62208cb83.txt"
Subject: xrolo/SPACRC/SunOS4.1.1/audio
```

A.23 Cluster 23

Nuvem de palavras:

Arquivo: "d24fdca2a59e3e036198f4b5a2c03da7.txt"Centróide:

```
Arquivo: "2b204bec5452d76945d40433b6324b2d.txt"
Subject: Re: Where can I get a New York taxi?
```

Vizinho mais próximo #1:

```
Arquivo: "ce1684d5a49179ee3c957ba8c7dc4764.txt"
Subject: Re: Asynchronous X Windows?
```

Vizinho mais próximo #2:

```
Arquivo: "9f8a5551a47a115a653c406cafd158fa.txt"
Subject: Re: WARNING.....(please read)...
```

Vizinho mais próximo #3:

```
Subject: Re: Let's play the name game!
```

A.24 Cluster 24

Nuvem de palavras:
Centróide:

```
Arquivo: "1984f68d3b9984729c537d50110f1957.txt"
Subject: Re: Militello update
```

Vizinho mais próximo #1:

```
Arquivo: "c2252b1fbf08231ffca452eaf016bd55.txt"
Subject: CLINTON: Public Schedule of the President 4.5.93
```

Vizinho mais próximo #2:

```
Arquivo: "f167ecb6bac720e8356da92872d99605.txt"
Subject: Re: WARNING.....(please read)...
```

Vizinho mais próximo #3:

```
Arquivo: "f2ccc2a2847b8e2b1d3f5787bf43c8ab.txt"
Subject: Re: SAD MAC CODE 0F0064 ???
```

A.25 Cluster 25

Nuvem de palavras:

Centróide:

```
Arquivo: "bc40f365885bc3d84f04b4a988688dba.txt"
Subject: NHL Playoff Stats & Scores
```

Vizinho mais próximo #1:

```
Arquivo: "41d418c8778488d51f4a1ac5556f13e4.txt"
Subject: Re: what to do with old 256k SIMMs?
```

Vizinho mais próximo #2:

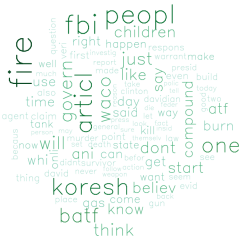
```
Arquivo: "47fc45fb3b525a8c59136fb6faaa8d4c.txt"
Subject: Re: Feds Caught in Another Lie
```

Vizinho mais próximo #3:

```
Arquivo: "4714249b6c42f046acf362f8508a2f07.txt"
Subject: COMPUTER/AUDIO/VIDEO/FURNITURE needed as of 4/18
```

A.26 Cluster 26

Nuvem de palavras:



Centróide:

```
Arquivo: "c690c6e5c1743d52f75fd7a1efb3005f.txt"
Subject: Re: Proventil Inhaler
```

Vizinho mais próximo #1:

```
Arquivo: "caf0dea0f2a9c9b582e622ddeeedf17f.txt"
Subject: Easter: what's in a name? (was Re: New Testament Double Standard?
```

Vizinho mais próximo #2:

```
Arquivo: "9f14b65c022d0d6045a660f066f4542a.txt"
Subject: Re: CD's For Sale
```

Vizinho mais próximo #3:

```
Arquivo: "64745931f5611bbfae81dd785d9835da.txt"
Subject: Re: Conductive Plastic, what happened?
```

A.27 Cluster 27

Nuvem de palavras:
Centróide:

```
Arquivo: "4c473463d5fe5a736277eeddc4277e4a.txt"
Subject: He has risen!
```

Vizinho mais próximo #1:

```
Arquivo: "b3d5054ce8acc76fb4be3e4da2471d12.txt"
From: stigaard@mhd.moorhead.msus.edu
```

Vizinho mais próximo #2:

```
Arquivo: "90d25083cff41ab68e49524bd0ebcddb.txt"
Subject: Re: Title for XTerm
```

Vizinho mais próximo #3:

```
Arquivo: "18d24bb92ead3761a36cb9b1f8c5105b.txt"
Subject: Tires For Sale
```

A.28 Cluster 28

Nuvem de palavras:

Centróide:

```
Arquivo: "5697682937fc72ad98ef612e52db5219.txt"
Subject: Re: Portland earthquake
```

Vizinho mais próximo #1:

```
Arquivo: "7e38777b6cb135e5b93e8f39b7ffa72b.txt"
Subject: Brand new H.P. toner for sale, cheap!
```

Vizinho mais próximo #2:

```
Arquivo: "7ebb8b3c4e2204f1a6ff3dbb42c70d51.txt"
Subject: Re: Pens fans reactions
```

Vizinho mais próximo #3:

```
Arquivo: "c11e7b206fb7ab3b9f2a979e097d8bc0.txt"
Subject: Re: Your opinion and what it means to me.
```

A.29 Cluster 29

Nuvem de palavras:

Centróide:

```
Arquivo: "aaf7efa6f0eb622bdf3cf84f1230a5ff.txt"
Subject: Help with fixed-frequency (52kHz?) VGA monitor
```

Vizinho mais próximo #1:

```
Arquivo: "f146063efd8c77a26ac20a8583c0e7ae.txt"
Subject: Re: CView answers
```

Vizinho mais próximo #2:

```
Arquivo: "d6baaf900cc00562b2ebd6529d453373.txt"
Subject: Forsale: Dynakit PAS-2x tube pre-amp
```

Vizinho mais próximo #3:

```
Arquivo: "cd7f677d1eb786b1c58f2df6fb633e05.txt"
Subject: Re: Zionist leaders' frank statements
```

A.30 Cluster 30

Nuvem de palavras:
Centróide:

```
Arquivo: "94674246c15f6002d6870e02caa9c114.txt"
Subject: Re: My Gun is like my American Express Card
```

Vizinho mais próximo #1:

```
Arquivo: "a83d7cd30ff11aebb82b0f605bda827e.txt"
Subject: Re: The Kuebelwagen??!!
```

Vizinho mais próximo #2:

```
Arquivo: "f2a776488eff74831161a3ebc97395c4.txt"
Subject: Date is stuck
```

Vizinho mais próximo #3:

```
Arquivo: "9b4881b51aa7b7355926983291b4b0df.txt"
Subject: Re: Societal basis for morality
```

Referências

[1] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. *KDD workshop on text ...*, pages 1–20, 2000.

[2] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. *Workshop on Artificial Intelligence for Web ...*, 2000.