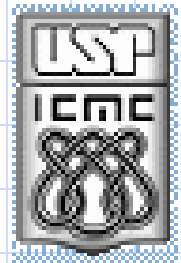


Análise de Agrupamento de Dados **(Aula 4 – Agrupamento de Textos)**

Thiago Ferreira Covões

Departamento de Ciências de Computação
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo



Agenda

3. Métodos para Agrupamento de Dados.

3.1 Métodos Hierárquicos

3.2 Métodos Particionais

4. Agrupamento de Textos

4. Mineração de Textos

- Quantidade abundante de dados está disponível na forma de textos
 - Artigos científicos
 - Livros
 - Páginas da Internet
- Área ampla, diversas oportunidades para MD
 - Classificação
 - Associação
 - Agrupamento
 - Podemos aplicar o que foi visto até agora para obtermos *insights* sobre uma coleção de documentos

Below every paper are TOP 100 most-occurring words in that paper and their color is based on LDA topic model with $k = 7$.
(It looks like 0 = theory, 1 = reinforcement learning, 2 = graphical models, 3 = deep learning/vision, 4 = optimization, 5 = neuroscience, 6 = embeddings etc.)

Toggle LDA topics to sort by: [TOPIC0](#) [TOPIC1](#) [TOPIC2](#) [TOPIC3](#) [TOPIC4](#) [TOPIC5](#) [TOPIC6](#)

Discriminatively Trained Sparse Code Gradients for Contour Detection

Ren Xiaofeng, Liefeng Bo

[\[pdf\]](#) [\[bibtext\]](#) [\[supplementary\]](#)
[\[rank by tf-idf similarity to this\]](#)
[\[abstract\]](#)



[set, algorithm, including] [average, approach, benchmark, evaluation] [comparing, normal, hierarchical] [contour, gpb, local, detection, depth, scg, color, image, oriented, matching, contrast, object, grayscale, precision, recognition, transform, work, learned, pooling, pixel, representation, double, global, learn, accuracy, scale, level, segmentation, figure, feature, nyu, globalization, scene, training, rich, single, automatically, apply, discriminative, codewords, ieee, half, directly, unsupervised, higher, chromaticity] [sparse, dictionary, gradient, pursuit, size, spectral, analysis, edge, step, sparsity] [power, coding, surface, natural] [code, learning, linear, data, orthogonal, dataset, svm, large, better, table, well, datasets]

Convolutional-Recursive Deep Learning for 3D Object Classification

Richard Socher, Brody Huval, Bharath Bath, Christopher Manning, Andrew Ng

[\[pdf\]](#) [\[bibtext\]](#) [\[supplementary\]](#)
[\[rank by tf-idf similarity to this\]](#)
[\[abstract\]](#)



[set, order, comparison, main, combining] [based, standard, water] [model, tree, number, structure, will, softmax, allows, parent, hierarchical] [depth, mns, object, cnn, feature, layer, recursive, unsupervised, mn, rgb, image, work, learn, single, convolutional, food, multiple, color, training, deep, recognition, pooled, scene, applied, pooling, convolution, trained, figure, raw, modality, architecture, scale, vision, computer, invariant, previous, compared, level, cnns, segmentation, box, improves, performed, produce, garlic, top] [random, size, matrix, high, sparse, method] [neural, input, network, surface] [learning, vector, performance, large, kernel, dataset, quality, art]

<http://cs.stanford.edu/people/karpathy/nipspreview/>

search.carrot2.org/stable/search?query=futebol+brasil&results=100&source=web&algorithm=lingo&view=foamtree&skin=fancy-compact&EToolsDocumentSource.langua

futebol brasil

Web Wiki Bing News Images Jobs PubMed RPUT

Folders Circles **FoamTree**

Top 77 results of about 487000 for **futebol brasil**

- [Confederação Brasileira de Futebol](#)
Site Oficial da Confederação Brasileira de Futebol.
<http://www.cbf.com.br/> [Bing, Google, Teoma, Yahoo]
- [Futebol - UOL Esporte](#)
Futebol - UOL Esporte. ... Bernardo Gentile/UOL Esporte. Clubes sugerem jogos pelo **Brasil** após perderem Engenhão, mas Ferj veta - Denilton Dias/Vipcomm ...
<http://esporte.uol.com.br/futebol/> [Google, Teoma]
- [Futebol do Brasil - Wikipédia, a enciclopédia livre](#)
O **Futebol no Brasil** foi introduzido por Charles Miller, um jovem **brasileiro** que, após viagem pela Inglaterra, trouxe consigo duas bolas de **futebol** e passou a ...
http://pt.wikipedia.org/wiki/Futebol_do_Brasil [Bing, Google, Teoma, Yahoo]
- [Campeonato Brasileiro - Esporte Interativo - Yahoo!](#)
Tudo sobre o Campeonato **Brasileiro**. Encontra as últimas notícias, tabelas e resultados. Cobertura completa do Campeonato **Brasileiro** no Yahoo! Esporte ...
<http://br.esporteinterativo.yahoo.com/futebol/campeonato-brasileiro/> [Google, Teoma]
- [Só Futebol Brasil](#)
Só **Futebol Brasil** - Todas as Camisas, todas as paixões · International Orders Central de Telefone Sobre a Só **Futebol Brasil** Meu cadastro Meus pedidos ...
<http://www.sofutebolbrasil.com/> [Bing, Google, Teoma, Yahoo]
- [futebol | globoesporte.com](#)
No globoesporte.com você encontra a melhor cobertura sobre o **Futebol** e Outros Esportes, no **Brasil** e no Mundo: Notícias, Vídeos e muito mais.
<http://globoesporte.globo.com/futebol/> [Google, Teoma]
- [Globoesporte](#)

Query: futebol brasil - Source: Web (77 results, 2745 ms) - Clusterer: Lingo

v3.6.1-SNAPSHOT | build | 2012-06-22 22:55 © 2002-2013 Stanislaw Osinski, Dawid Weiss

<http://search.carrot2.org/>

4. Agrupamento de Textos

- Trabalhar com textos apresenta diversos desafios:
 - Ausência integral/parcial de estrutura
 - Idioma
 - Qualidade dos dados
 - Twitter
 - Mensagens de fóruns

4.1 Pré-processamento

- Vimos como agrupar dados no formato tabela atributo-valor
- Como gerar uma tabela desse tipo a partir dos textos
 - Nossos x_i são documentos
 - E os nossos atributos?
 - Caracteres? Pouco descritivo...
 - Palavras? Perda de contexto...
 - N-gramas? Nem sempre ajuda...
 - Vamos chamar de *termos*...

4.1 Pré-processamento

➤ Exemplo:

“O algoritmo de redes neurais funcionou nos dados do Twitter”

Frequência de palavras (1-gram)

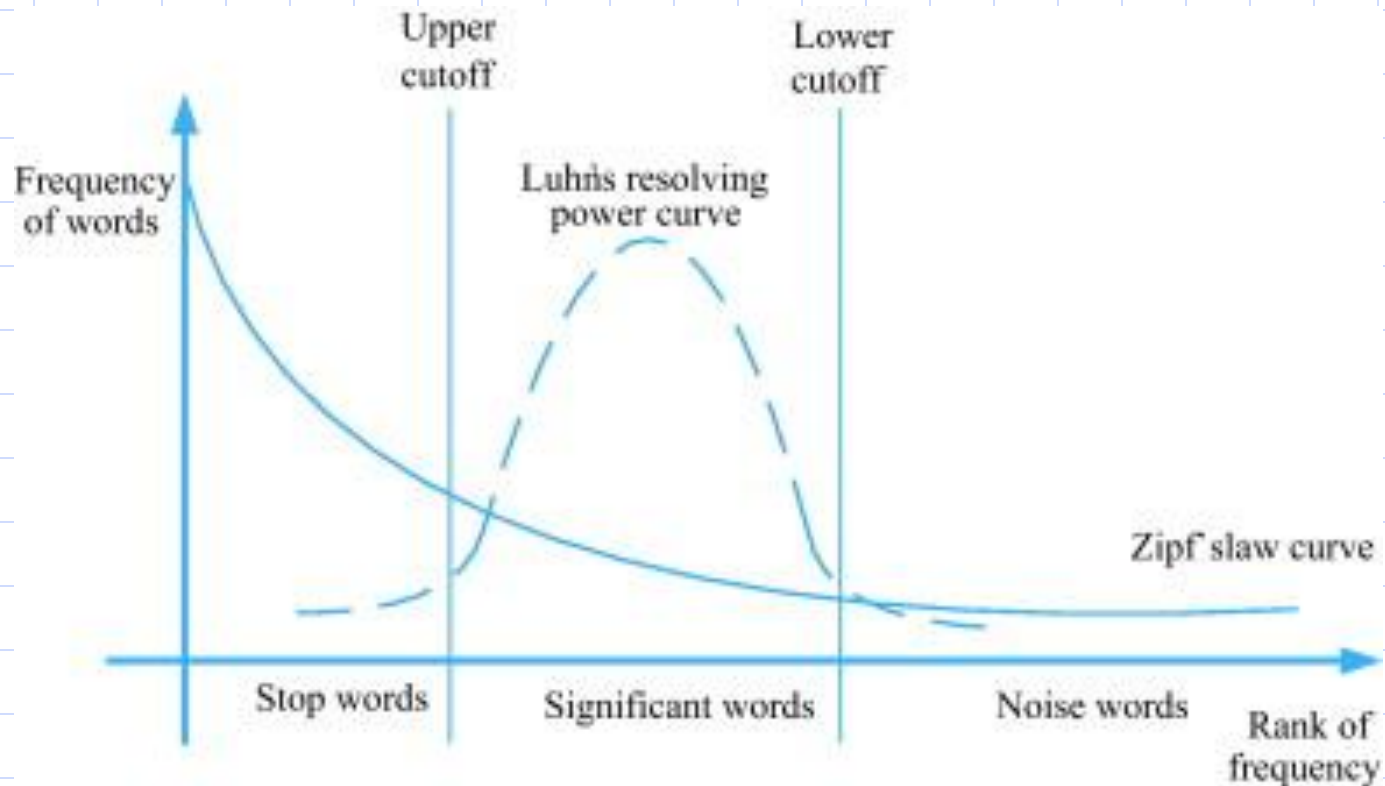
o	algoritmo	de	redes	neurais	funcionou	nos	dados	do	twitter
1	1	1	1	1	1	1	1	1	1

Frequência de bi-gramas (2-gram)

o algoritmo	algoritmo de	de redes	redes neurais	neurais funcionou	funcionou nos	...
1	1	1	1	1	1	

4.1 Pré-processamento

- Corte de Luhn e Lei de Zipf
 - Técnica para redução de termos baseado em sua frequência



4.1 Pré-processamento

- Procedimento padrão:
 - Remoção de *stop words*
 - Preposições, artigos...
 - Conversão de sinônimos
 - Automóvel, carro, veículo
 - Radicalização (Stemming)
 - Pedra, pedreiro, pedregulho
- Mesmo realizando essas remoções, o número de termos costuma ser da ordem de milhares até para conjuntos (corpus) pequenos de documentos.

4.1 Pré-processamento

- Como calcular os valores de cada atributo (x_{ij}):
 - Booleano: presente ou não, $x_{ij} \in \{0,1\}$;
 - Frequência do termo (tf): número de vezes que o termo apareceu no documento (valor absoluto ou relativo), $x_{ij} \in \{0, \dots, freqMax\}$ ou $x_{ij} \in [0,1]$
 - Log-Frequência (meio termo): $1+\log(tf)$ [$tf > 0$]
 - A frequência do termo na coleção traz alguma informação adicional?

4.1 Pré-processamento

- Frequência na coleção:
 - Termos raros são mais informativos do que termos *muito* frequentes
 - No entanto, mesmo termos frequentes na coleção podem ser úteis para discernir entre alguns documentos
- Ponderação tf-idf:
 - Se df é o número de documentos na coleção que apresentam o termo e N o número de documentos
 - $ptf = 1 + \log(tf)$ [$tf > 0$]
 - $idf = \log(N/df)$
 - $tf-idf = ptf \times idf$

4.1 Pré-processamento

➤ Normalização

➤ Documentos extensos (por exemplo, teses de doutorado) costumam ser repetitivos e ter uma diversidade grande de palavras

➤ Isso afeta diretamente as frequências dos termos

➤ Por tal razão é comum normalizar os documentos (objetos)

➤ A normalização mais comum é a L2-norm (euclidiana)

➤ $\tilde{\mathbf{x}}_i = \mathbf{x}_i / \|\mathbf{x}_i\|$; $\|\mathbf{x}_i\| = \sqrt{\sum_{j=1}^n (x_{ij})^2}$

4.2 Medida de similaridade

- Obtido o conjunto de dados de forma adequada para o agrupamento é necessário definir como comparar objetos (documentos)
- Quais são as características dos dados?
 - Alta dimensionalidade
 - Esparsos (número de valores iguais a zero é considerável)
- Medida mais comum: medida do cosseno

4.2 Medida de similaridade

➤ Medida do cosseno

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

$$\mathbf{x} \cdot \mathbf{y} = \sum x_i y_i$$

➤ Os termos que documentos têm em comum são mais importantes do que os que estão ausentes

➤ Atributo assimétrico

➤ Transformar em distância: $1 - \cos(\mathbf{x}, \mathbf{y})$, [considerando que \mathbf{x} e \mathbf{y} têm apenas elementos não-negativos]

4.3 *Spherical k-means*

➤ Variante do algoritmo *k-means* para agrupar dados utilizando a medida do cosseno ao invés da distância Euclidiana

➤ Assume que os dados foram normalizados pela L2-norm

➤ Dada a alteração de como comparar objetos, o cálculo do centróide também é modificado:

$$\tilde{\mathbf{c}}_j = \frac{\mathbf{m}_j}{\|\mathbf{m}_j\|}, \quad \mathbf{m}_j = \frac{\sum_{\tilde{\mathbf{x}}_i \in C_j} \tilde{\mathbf{x}}_i}{N}$$

4.3 *Spherical k-means*

➤ Os passos básicos do algoritmo são os mesmos:

1. Selecionar k pontos (centróides iniciais);

2. Repetir até “convergir”:

- 2.1 Formar k grupos atribuindo cada ponto ao seu centróide mais próximo;

- 2.2 Re-computar o centróide de cada grupo;

4.3 *Spherical k-means*

➤ As dificuldades existentes no *k-means* ainda permanecem:

➤ Sensibilidade à inicialização

➤ Como comparar duas partições?

$$J = \sum_{i=1}^k \sum_{\tilde{x} \in C_i} \tilde{c}_i \cdot \tilde{x}$$

Referências

➤ Slides do Prof. Ronaldo Prati (UFABC) - <http://professor.ufabc.edu.br/~ronaldo.prati/DataMining/Mineracao-Textos.pdf>

➤ Inderjit S. Dhillon , Dharmendra S. Modha, Concept Decompositions for Large Sparse Text Data Using Clustering, **Machine Learning**, v.42 n.1-2, p.143-175, 2001

Onde estamos?

3. Métodos para Agrupamento de Dados.

3.1 Métodos Hierárquicos;

3.2 Métodos Particionais.

4. Agrupamento de textos:

- Medida de similaridade adequada:

- Adaptação do k-médias

Próxima aula: PROVA! Boa sorte!

Na sequência do curso: Regressão