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, diversos 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-occuring 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

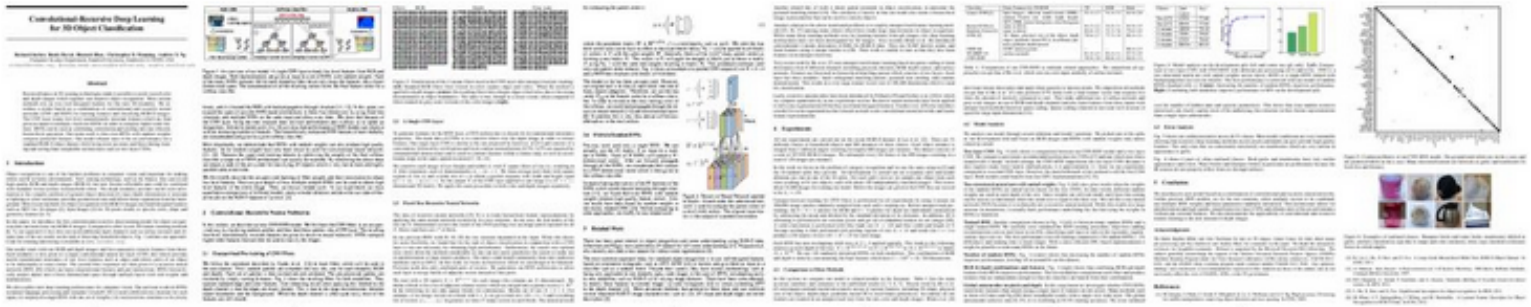
Discriminatively Trained Sparse Code Gradients for Contour Detection
Ren Xiaofeng, Liefeng Bo

[pdf] [bibtex] [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] [bibtex] [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, rnns, object, cnn, feature, layer, recursive, unsupervised, rnn, 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/

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 $\boldsymbol{x}_i$ 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)

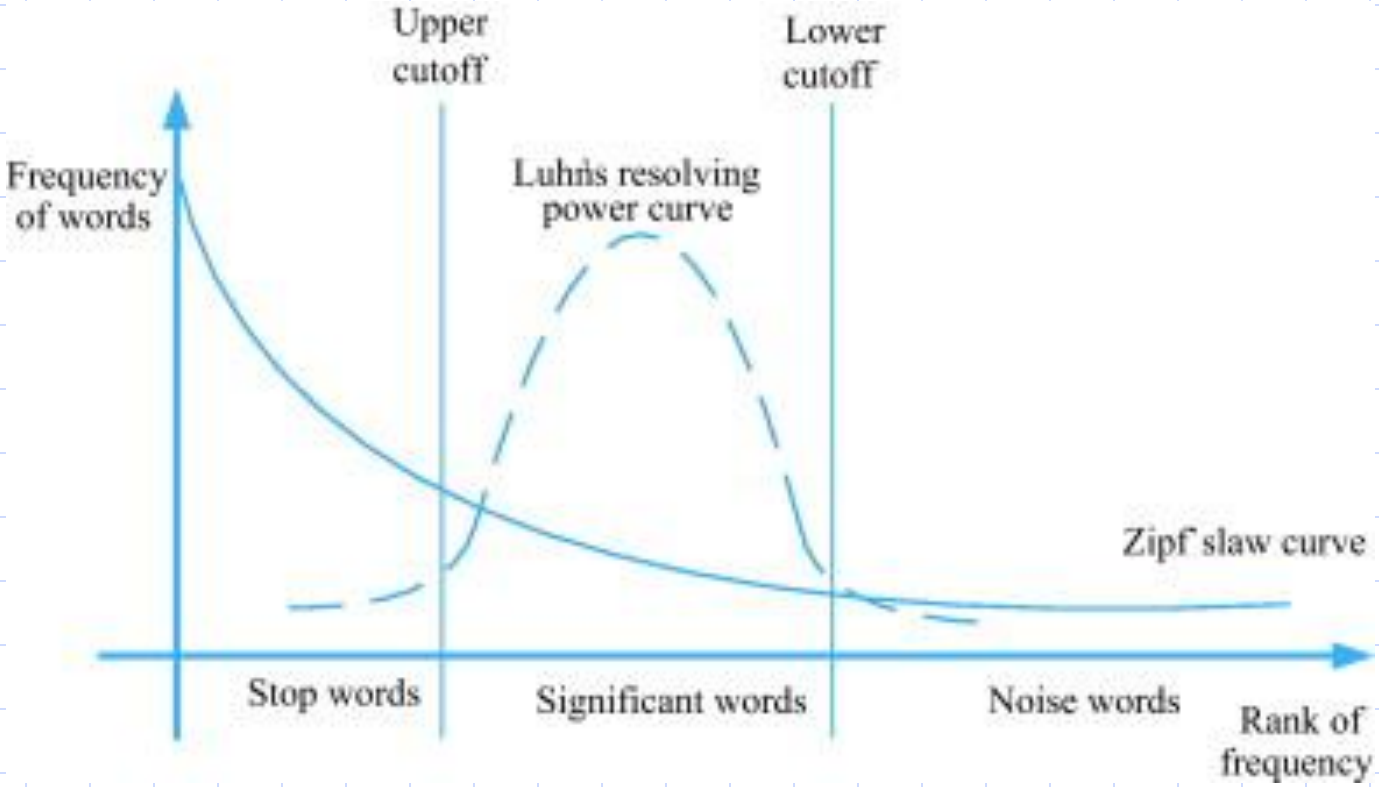
o	algoritmo	de	redes	neurais	funcionou	nos	dados	do	twitter
1	1	1	1	1	1	1	1	1	1

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 × 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)
    - $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i / \|\boldsymbol{x}_i\|; \quad \|\boldsymbol{x}_i\| = \sqrt{\sum_{j=1}^{n} (x_{ij})^2}$

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

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

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

- Os termos que documentos têm em comum são mais importantes do que os que estão ausentes
  - Atributo assimétrico

- Transformar em distância: $1\text{-}\cos(\boldsymbol{x}, \boldsymbol{y})$, [considerando que **x** e **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 Euclidana
  - Assume que os dados foram normalizados pela L2-norm

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

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

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