Agrupamento de mensagens do Twitter

Fabrício J. Barth

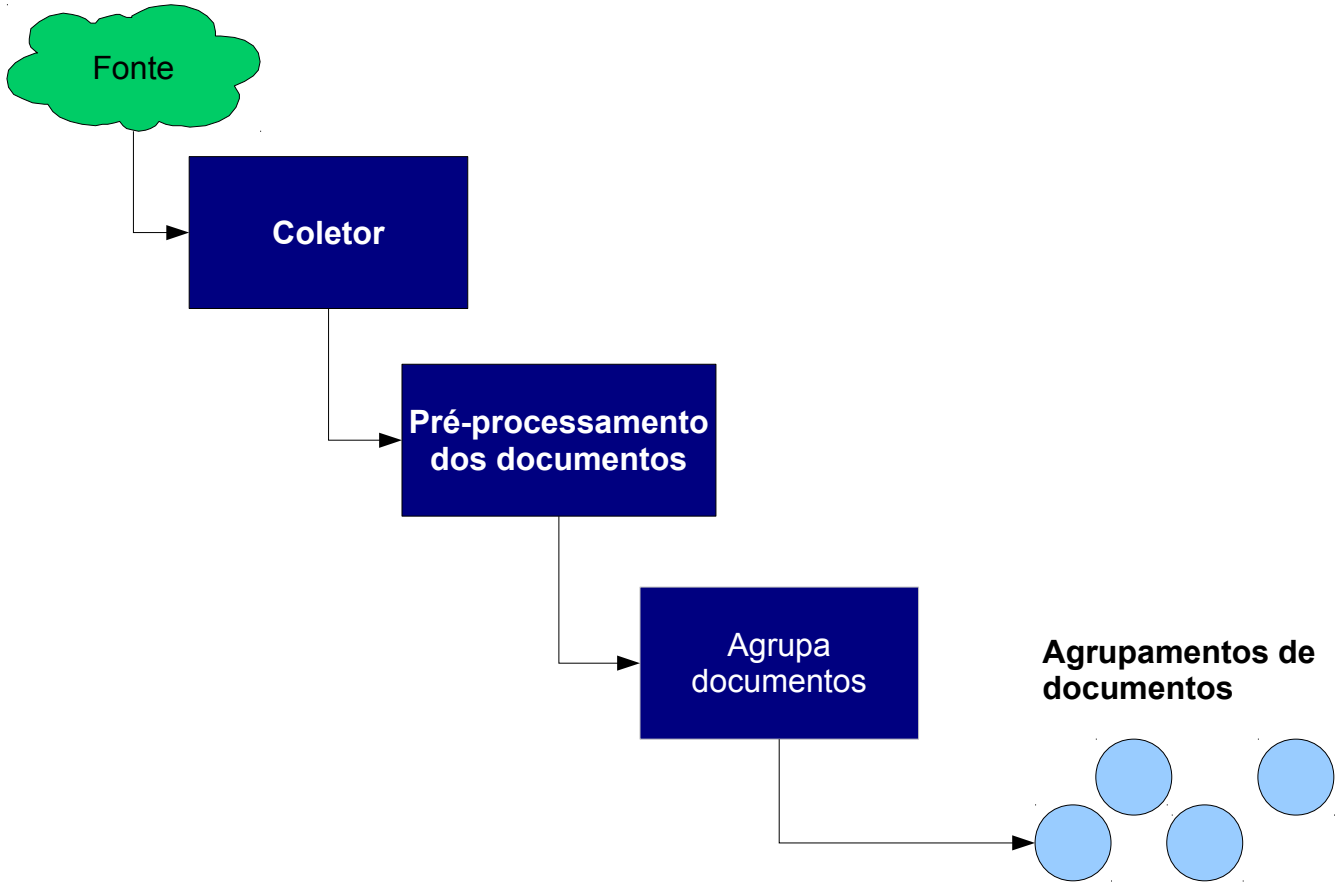
Faculdade BandTec e VAGAS Tecnologia

Junho de 2013

Sumário

- Componentes para uma solução
- Coletor
- Pré-processamento dos documentos
- Agrupamento dos documentos
- Análises

Componentes para uma solução...



Coletando dados do twitter com o R

```
library(twitteR)
cred <- OAuthFactory$new(
                consumerKey="XXXX",
                consumerSecret="XXXX",
                requestURL="https://api.twitter.com/oauth/request_token",
                accessURL="https://api.twitter.com/oauth/access_token",
                authURL="http://api.twitter.com/oauth/authorize")

cred$handshake()
registerTwitterOAuth(cred)

about_protesto <- searchTwitter('#protestosp', n=250)
about_protesto_2 <- searchTwitter('sp', n=90)
prefeituraTweets <- userTimeline('Prefeitura_SP')

text <- c(about_protesto, about_protesto_2, prefeituraTweets)

df <- twListToDF(text)
save(df, file="../data/protesto.rda")
```

Pré-processamento dos dados

Formato de um documento

... Esta disciplina tem como objetivo apresentar os principais conceitos da área de Inteligência Artificial, caracterizar as principais técnicas e métodos, e implementar alguns problemas clássicos desta área sob um ponto de vista introdutório.
A estratégia de trabalho, o conteúdo ministrado e a forma dependerão dos projetos selecionados pelos alunos.
Inicialmente, os alunos deverão trazer os seus Projetos de Conclusão de Curso, identificar intersecções entre o projeto e a disciplina, e propor atividades para a disciplina. ...

Conjunto de Exemplos - Atributo/Valor

Doc.	**apresent**	**form**	**tecnic**	**caracteriz**	$\cdots$
d_1	0.33	0.33	0.33	0.33	$\cdots$
d_2	0	0.5	0.2	0.33	$\cdots$
d_3	1	0.6	0	0	$\cdots$
d_4	0.4	0.3	0.33	0.4	$\cdots$
d_5	1	0.4	0.1	0.1	$\cdots$
d_n	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Atributo/Valor usando vetores

Como representar os documentos?

$$\overrightarrow{d_i} = (p_{i1}, p_{i2}, \cdots, pin) \tag{1}$$

- Os atributos são as palavras que aparecem nos documentos.
- Se todas as palavras que aparecem nos documentos forem utilizadas, o vetor não ficará muito grande?

Diminuindo a dimensionalidade do vetor

- Como filtrar as palavras que devem ser usadas como atributos?
- Em todos os idiomas existem átomos (palavras) que não significam muito. **Stop-words**

Esta disciplina **tem como** objetivo apresentar **os** principais conceitos **da** área **de** Inteligência Artificial**,** caracterizar **as** principais técnicas **e** métodos**, e** implementar alguns problemas clássicos **desta** área **sob um** ponto **de** vista introdutório**.**

. . .

Diminuindo ainda mais a dimensionalidade do vetor

- Algumas palavras podem aparecer no texto de diversas maneiras: **técnica**, **técnicas**, **implementar**, **implementação**...
- **Stemming** - encontrar o radical da palavra e usar apenas o radical.

Atributo/Valor usando vetores

- Já conhecemos os atributos.
- E os valores?
  - ⋆ **Booleana** - se a palavra aparece ou não no documento (1 ou 0)
  - ⋆ **Por freqüência do termo** - a freqüência com que a palavra aparece no documento (normalizada ou não)
  - ⋆ **Ponderação tf-idf** - o peso é proporcional ao número de ocorrências do termo no documento e inversamente proporcional ao número de documentos onde o termo aparece.

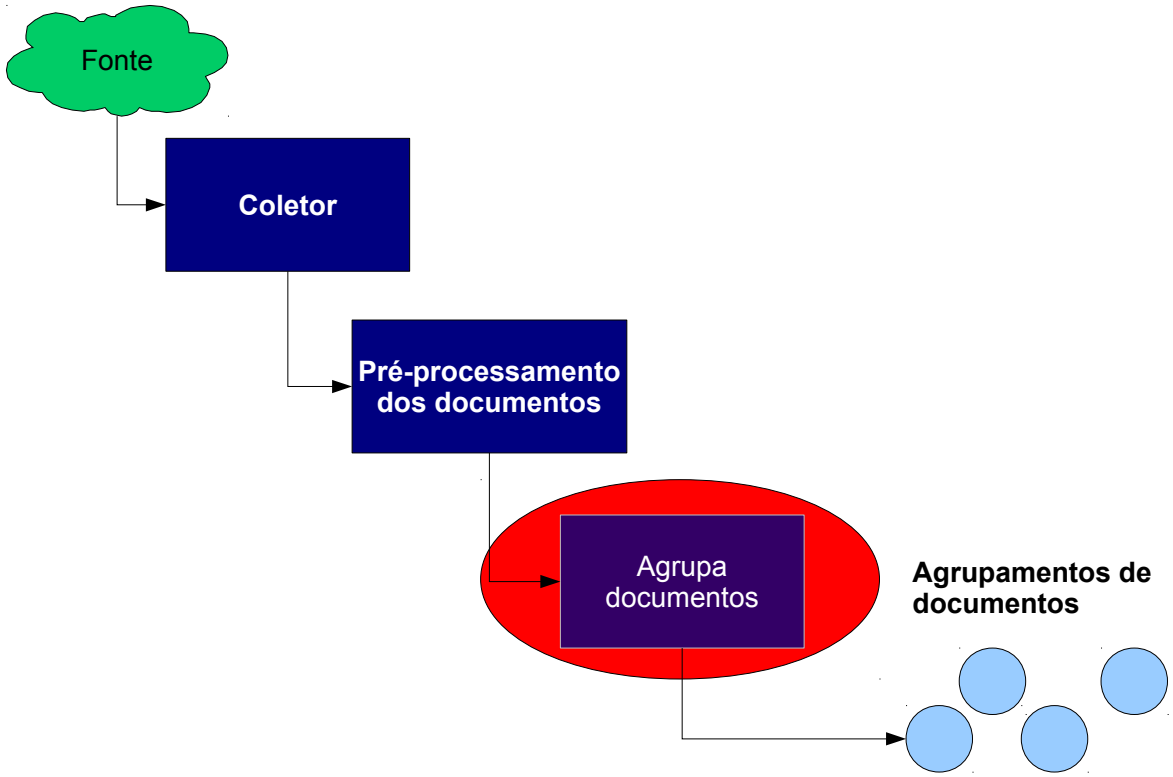
Por freqüência do termo

(apresent,0.33) (form,0.33) (tecnic,0.33) (caracteriz,0.33)
(projet,1.0) (introdutori,0.33) (objet,0.33) (inteligente,0.33)
(conclusa,0.33) (selecion,0.33) (intersecco,0.33) (classic,0.33)
(identific,0.33) (conceit,0.33) (trabalh,0.33) (disciplin,1.0)
(traz,0.33)

Executando esta etapa no R

http://rpubs.com/fbarth/agrupamentosTwitter

Componentes para uma solução...



Wiki2Group - http://trac.fbarth.net.br/wikiAnalysis

Algoritmos para Agrupamento

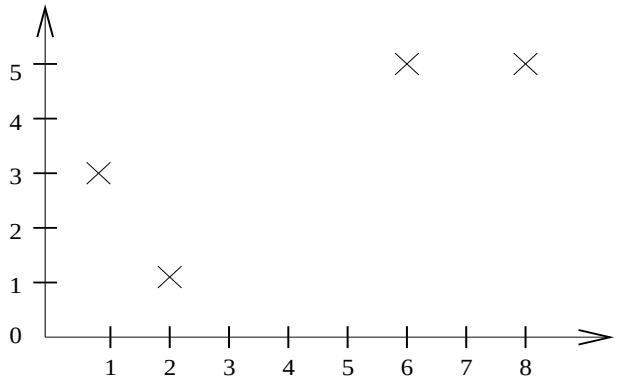
Definições de Algoritmos de Agrupamento

- O objetivo dos algoritmos de agrupamento é colocar os objetos **similares** em um **mesmo grupo** e objetos **não similares** em **grupos diferentes**.
- Normalmente, objetos são descritos e agrupados usando um conjunto de **atributos e valores**.
- Não existe nenhuma informação sobre a classe ou categoria dos objetos.

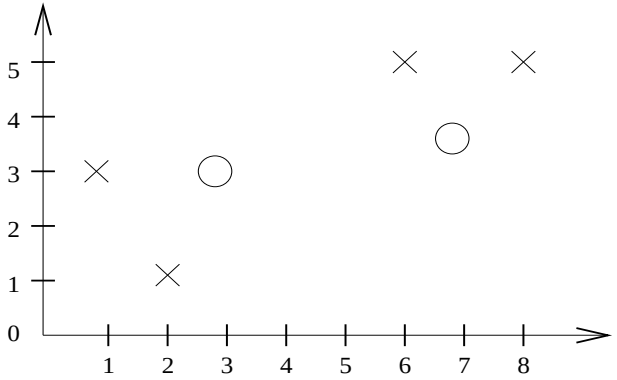
Algoritmos para Agrupamento - *K-means*

- **K** significa o número de agrupamentos (que deve ser informado à priori).
- Sequência de ações **iterativas**.
- A parada é baseada em algum critério de qualidade dos agrupamentos (por exemplo, similaridade média).

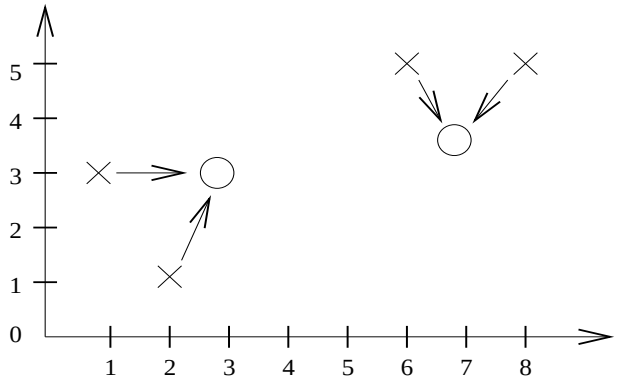
Algoritmo para Agrupamento - *K-means*



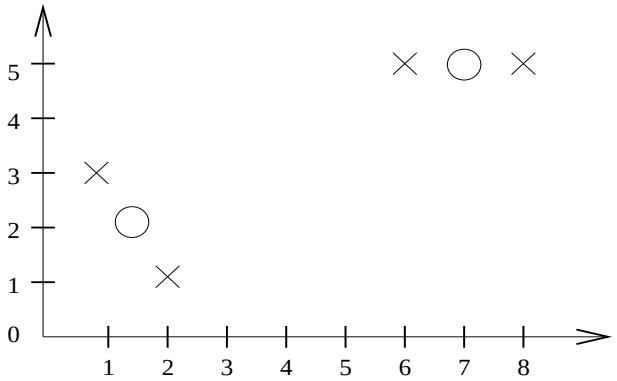
(1) Objetos que devem ser agrupados



(2) Sorteio dos pontos centrais dos agrupamentos



(3) Atribuição dos objetos aos agrupamentos



(4) Definição do centro do agrupamento

Algoritmo **K-means**

- A medida de distância pode ser a distância Euclidiana:

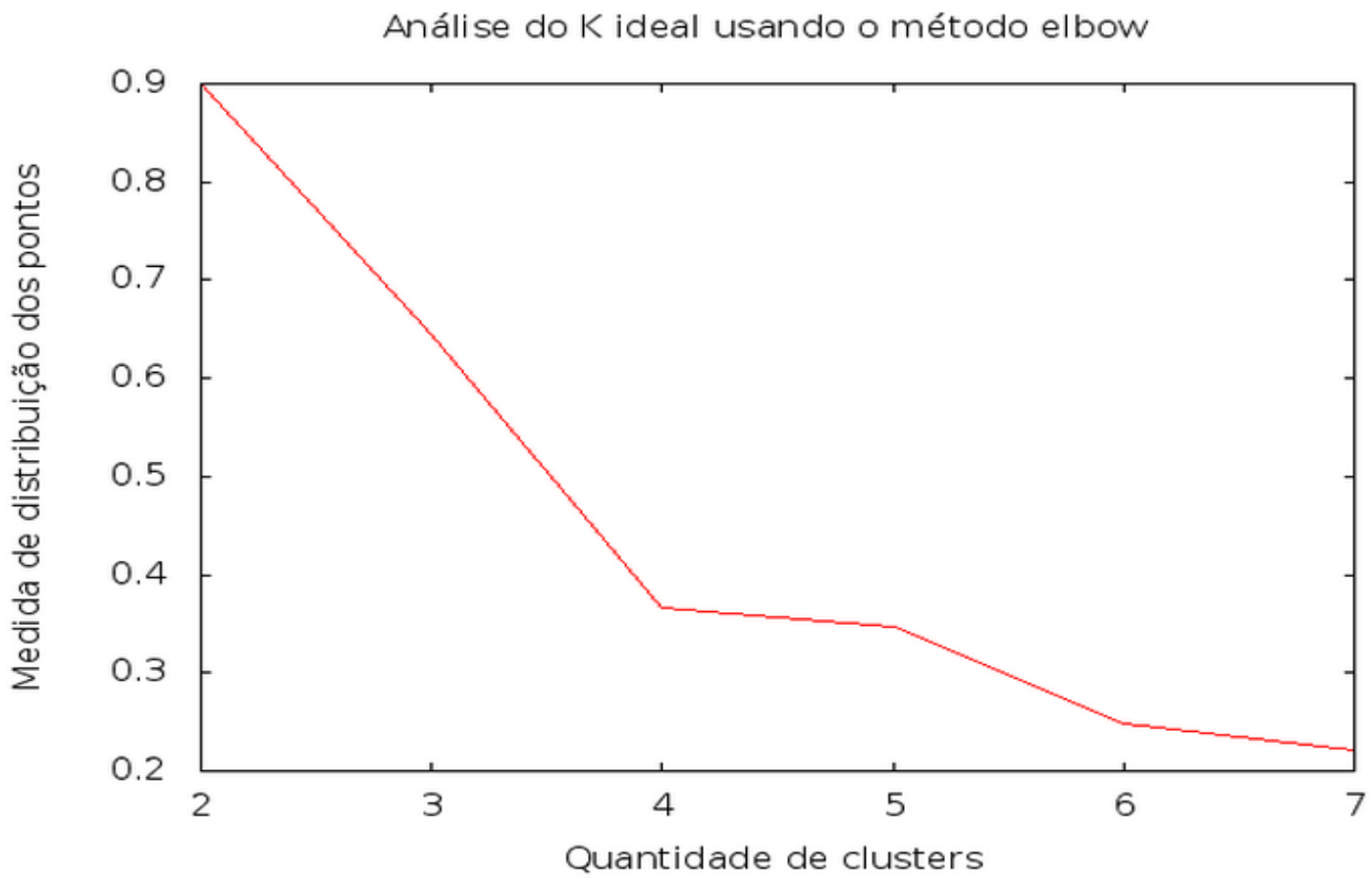
$$| \vec{x} - \vec{y} | = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

- a função para computar o ponto central pode ser:

$$\vec{\mu} = \frac{1}{M} \sum_{\vec{x} \in C} \vec{x} \tag{3}$$

onde M é igual ao número de pontos no agrupamento C.

Como determinar o melhor k?



A medida de distribuição dos pontos normalmente empregada é *sum of squared errors*.

Agrupamento de mensagens do twitter com o R

http://rpubs.com/fbarth/agrupamentosTwitter

Referências

- RDataMining.com: Text Mining. http://www.rdatamining.com/examples/text-mining. Acessado em 14 de junho de 2013.
- Ingo Feinerer. Introduction to the tm Package: Text Mining in R. http://cran.r-project.org/web/packages/tm/vignettes/tm.pdf. Acessado em 14 de junho de 2013.

- BARTH, F. J. Ferramentas para a detecção de grupos em Wikis. In: VII Simpósio Brasileiro de Sistemas Colaborativos, 2010, Belo Horizonte. Anais do VII Simpósio Brasileiro de Sistemas Colaborativos. IEEE Computer Society, 2010. v.II. p.8 - 11.
- BARTH, F. J. ; BELDERRAIN, M. C. R. ; QUADROS, N. L. P. ; FERREIRA, L. L. ; TIMOSZCZUK, A. P. . Recuperação e mineração de informações para a área criminal. In: VI Encontro Nacional de Inteligência Artificial, 2007, Rio de Janeiro. Anais do XXVII Congresso da SBC, 2007.