Mineração de Textos

- Os estudos em Aprendizado de Máquina normalmente trabalham com dados estruturados
- Entretanto, uma grande quantidade de informação é armazenada em textos, que são dados semi-estruturados
- Nesta apresentação é dada uma breve introdução à Mineração de Textos

José Augusto Baranauskas
Departamento de Física e Matemática – FFCLRP-USP

augusto@ffclrp.usp.br
http://dfm.ffclrp.usp.br/~augusto

Introdução

- Uma grande quantidade de toda informação disponível atualmente encontra-se sob a forma de **textos** (ou **documentos**) semi-estruturados, tais como livros, artigos, manuais, *e-mails* e a *Web*
- O termo **semi-estruturado** indica que os dados não são completamente estruturados nem completamente sem estrutura
  - Um documento pode conter alguns atributos estruturados:
    - Título, autor(es), data da publicação
  - mas também contém alguns elementos textuais sem estrutura
    - Resumo e conteúdo

Introdução

- **Mineração de Textos** (*Text Mining* - TM) tem como objetivo tratar essa informação semi-estruturada
- Em especial, a literatura biomédica é uma fonte de informação extremamente rica e um conjunto de resumos (abstracts) da base MEDLINE da *National Library of Medicine* resume esta literatura de forma compreensiva
- Apesar desta fonte de recursos ser atrativa e de fácil acesso, a extração automática de informação útil a partir dela é um desafio uma vez que os resumos estão em linguagem natural

Representação

- De maneira geral, um documento é representado por um conjunto de palavras-chave (ou termos)
- O usuário fornece um termo ou uma expressão formada por termos
  - Chá **or** café
  - Carro **and** oficina mecânica

Sinonímia & Polissemia

- Sinonímia: um termo possui vários sinônimos
  - Carro, automóvel, veículo
- Polissemia: um mesm termo tem diferentes significados, dependendo do contexto
  - Mineração (textos?), mineração (carvão?)
  - Exame (teste?), exame (médico?)

Stop List

- É possível associar uma **stop list** para um determinado conjunto de documentos
- Uma **stop list** é um conjunto de palavras que são consideradas “irrelevantes”
  - Normalmente inclui artigos, preposições, conjunções
- A **stop list** pode variar entre conjuntos de documentos (mesma área, mesma língua)

Stem

- Um grupo de diferentes termos podem compartilhar um mesmo **radical** (***stem***)
- Em geral, termos que possuem o mesmo ***stem*** correspondem a pequenas variações sintáticas uns dos outros
  - Droga, drogas, drogado, drogaria
- Com essa identificação, é possível armazenar apenas o ***stem***

Representação

- Iniciando com um conjunto de **n** documentos e **m** termos, é possível modelar cada documento como um vetor **v** no espaço **m**-dimensional
  - Os vetores podem ser binários:
    - 0 indica que o termo não ocorre no documento
    - 1 caso contrário
  - Os vetores podem ser ternários:
    - 0 indica que o termo não ocorre no documento
    - 1 que o termo ocorre uma única vez
    - 2 que o termo ocorre duas ou mais vezes no documento
  - Os vetores podem conter a freqüência absoluta de cada termo no documento (um número inteiro)
  - Os vetores podem conter a freqüência relativa de cada termo no documento (um número real), ou seja, a freqüência absoluta dividida pelo número total de ocorrências de todos os termos no documento

Matriz de Freqüência Absoluta

	t_1	t_2	t_3	t_4	t_5
d_1	321	354	15	22	74
d_2	84	91	32	143	87
d_3	0	1	167	1	85
d_4	68	56	46	203	92
d_5	1	82	289	0	25
d_6	1	0	225	0	54
d_7	430	392	1	54	121

Matriz de Freqüência Relativa

	t_1	t_2	t_3	t_4	t_5
d_1	0.41	0.45	0.02	0.03	0.09
d_2	0.19	0.21	0.07	0.33	0.20
d_3	0.00	0.00	0.66	0.00	0.33
d_4	0.15	0.12	0.10	0.44	0.20
d_5	0.00	0.21	0.73	0.00	0.06
d_6	0.00	0.00	0.80	0.00	0.19
d_7	0.43	0.39	0.00	0.05	0.12

Matriz Booleana (Binária)

	t_1	t_2	t_3	t_4	t_5
d_1	1	1	1	1	1
d_2	1	1	1	1	1
d_3	0	1	1	1	1
d_4	1	1	1	1	1
d_5	1	1	1	0	1
d_6	1	0	1	0	1
d_7	1	1	1	1	1

Matriz Ternária

	t_1	t_2	t_3	t_4	t_5
d_1	2	2	2	2	2
d_2	2	2	2	2	2
d_3	0	1	2	1	2
d_4	2	2	2	2	2
d_5	1	2	2	0	2
d_6	1	0	2	0	2
d_7	2	2	1	2	2

Identificando Documentos Similares

- Uma vez obtida a matriz de freqüência (binária, ternária, absoluta ou relativa) é possível aplicar qualquer métrica de distância, uma vez que é esperado que documentos similares tenham freqüências similares
- É possível medir a similaridade entre um conjunto de documentos ou entre um documento e uma *query* (consulta), freqüentemente definida por meio de um conjunto de termos

Identificando Documentos Similares

- Após obter a freqüência de um termo em um documento é possível modificá-la de forma a considerar a **importância percebida** daquele termo
- A formulação *tf-idf* é utilizada para computar pesos ou *scores* para os termos
- Os valores permanecem positivos de forma a capturar a presença ou ausência do termo no documento

Métrica tf-idf

- O peso associado ao termo *j* é a freqüência do termo (tf = *term frequency*) modificado por um fator de escala para a importância do termo
- O fator de escala é chamado de freqüência inversa do termo *j* em todos documentos (idf = *inverse document frequency*)
  - Ele simplesmente verifica o número de documentos que contêm o termo *j* (df = *document frequency*) e inverte a escala
  - *n* é o número total de documentos

$$\text{tf} - \text{idf}(j) = \text{tf}(j) \times \text{idf}(j)$$

$$\text{idf}(j) = \log_2 \left(\frac{n}{\text{df}(j)} \right)$$

Métrica tf-idf

- Quando o termo aparece em muitos documentos ele é considerado irrelevante e o fator de escala é diminuído, tendendo a zero
- Quando o termo é relativamente único e aparece em poucos documentos o fator de escala aumenta uma vez que ele parece ser importante
- Existem métricas alternativas à formulação *td-idf* mas a motivação geral é a mesma
- O resultado desse processo é um *score* positivo que substitui a freqüência em uma célula em nossa tabela
- Quanto maior o *score* mais importante seu valor esperado para o método de aprendizado

Identificando Documentos Similares

- Qualquer medida de similaridade/distância pode ser utilizada
- Uma métrica de similaridade comumente utilizada é o co-seno entre os vetores
- Sejam *u* e *v* dois vetores de documentos; a métrica de similaridade de co-seno é definida como

$$\cos(u,v) = \frac{u \cdot v}{|u||v|}$$

- onde

$$u \cdot v = \sum_{j=1}^{m} u_j v_j \qquad |v| = \sqrt{v \cdot v}$$

- Quanto mais próximo de zero o valor encontrado, mais próximos estão os documentos
  - -1 <= cos(α) <= 1; cos(0) = 1; cos(π/2) = 0; cos(π) = -1
  - Como *u* e *v* assumem somente valores positivos, 0 <= cos(*u*,*v*) <= 1

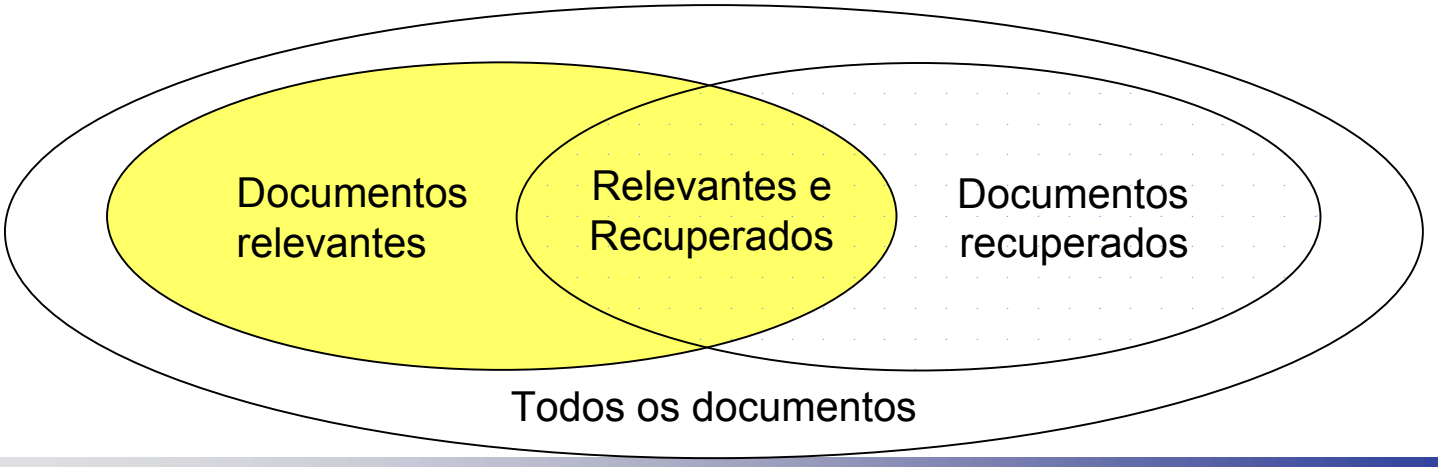
Identificando Documentos Similares

	t_1	t_2	t_3	t_4	t_5
d_1	321	354	15	22	74
d_2	84	91	32	143	87
d_3	31	71	167	72	85
d_4	68	56	46	203	92
d_5	72	82	289	31	25
d_6	15	6	225	15	54
d_7	430	392	17	54	121

$\cos(d_1,d_1) = 1.0000$ $\cos(d_1,d_2) = 0.6787$ $\cos(d_1,d_3)=0.4363$

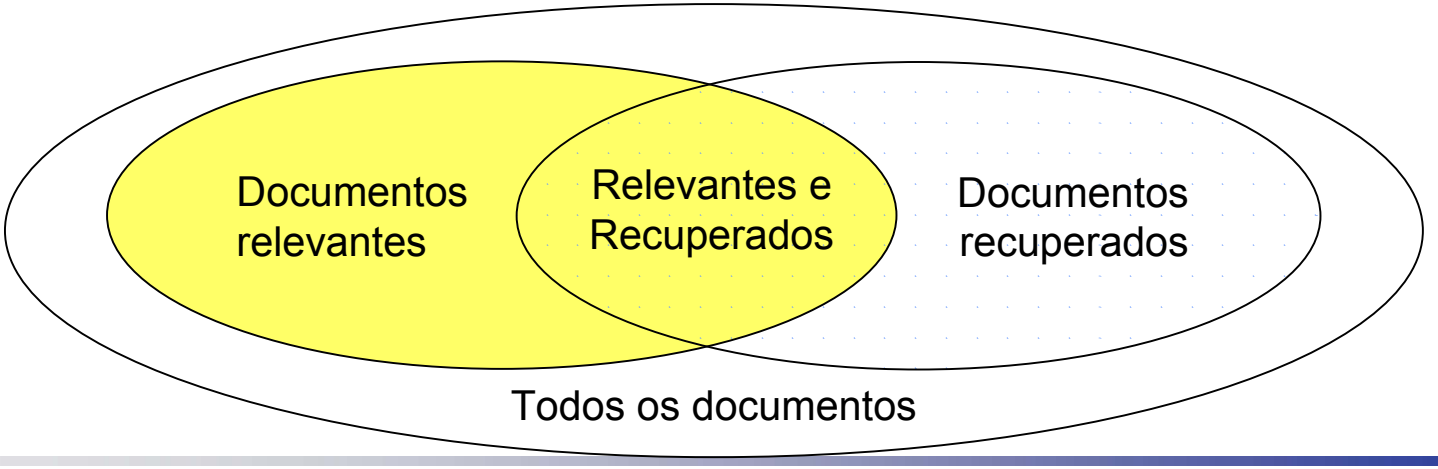
Métricas Básicas

- As duas métricas usualmente utilizadas para avaliar o desempenho são
  - **Precisão**: porcentagem de documentos recuperados que de fato são relevantes
  - ***Recall***: porcentagem de documentos que são relevantes e foram, de fato, recuperados



Métricas Básicas

- As duas métricas usualmente utilizadas para avaliar o desempenho são
  - **Precisão** = |Relevantes ∩ Recuperados| / |Recuperados|
  - ***Recall***: |Relevantes ∩ Recuperados| / |Relevantes|



Métricas Básicas

Relevantes Recuperados Tp	Relevantes Não Recuperados Fn
Não Relevantes Recuperados Fp	Não Relevantes Não Recuperados Tn

Métricas Básicas

- As duas métricas usualmente utilizadas para avaliar o desempenho são
  - **Precisão** = Tp/(Tp+Fp) = confiabilidade positiva (prel)
  - ***Recall*** = Tp/(Tp+Fn) = sensitividade (sens)
- Adicionalmente, a métrica F-measure também é utilizada:
  - **F-measure** = 2 / (1/prel + 1/sens)

Métricas Básicas

- Considere um conjunto de documentos rotulados
- Focalizando em um rótulo específico, por exemplo, *saúde*; assuma um classificador que rotula documentos como sendo sobre *saúde* ou não e vamos utilizá-lo para recuperar todos os documentos que ele rotula
  - **Precisão** a porcentagem de documentos que o classificador corretamente rotula como sendo sobre *saúde*
  - ***Recall*** é a porcentagem de todos os documentos sobre *saúde* que foram recuperados
  - **F-measure** é definida como a média harmônica de **precisão** e ***recall*** e é frequentemente utilizada para medir o desempenho de um sistema quando um único número é desejado