Indexação e Tratamento de Dados Heterogêneos: Variedade

Mecanismos de Busca

Aula 5

Análise de sentimento com NLTK

Informações sobre o Trabalho em grupo

Temas para o exercício 5

Análise de sentimento

Como saber, programaticamente, o sentimento que foi expresso através de um determinado texto?

- Sentimento positivo, negativo, neutro, misturado

Processamento de linguagem natural

- Apesar de grandes avanços na área ainda falta bastante para alcançar o mesmo nível humano de entendimento

Análise de sentimento

Uma área complicada ainda para ciência da computação

- Aquela manga está suja (manga da camisa ou a fruta?)
  - Reconhecimento de entidades
- Esse filme é tão bom, minha nota pra ele, de zero a cinco, é 1!
  - Ironia

Análise de sentimento

Métricas

- Praticamente as mesmas de information retrieval

Aplicações

- Previsão de comportamento do mercado baseado em mídias sociais
- Sentimento dos consumidores
- Previsão de resultado de eleição

Como aplicar análise de sentimento

Classificar um texto de acordo com o sentimento expressado:

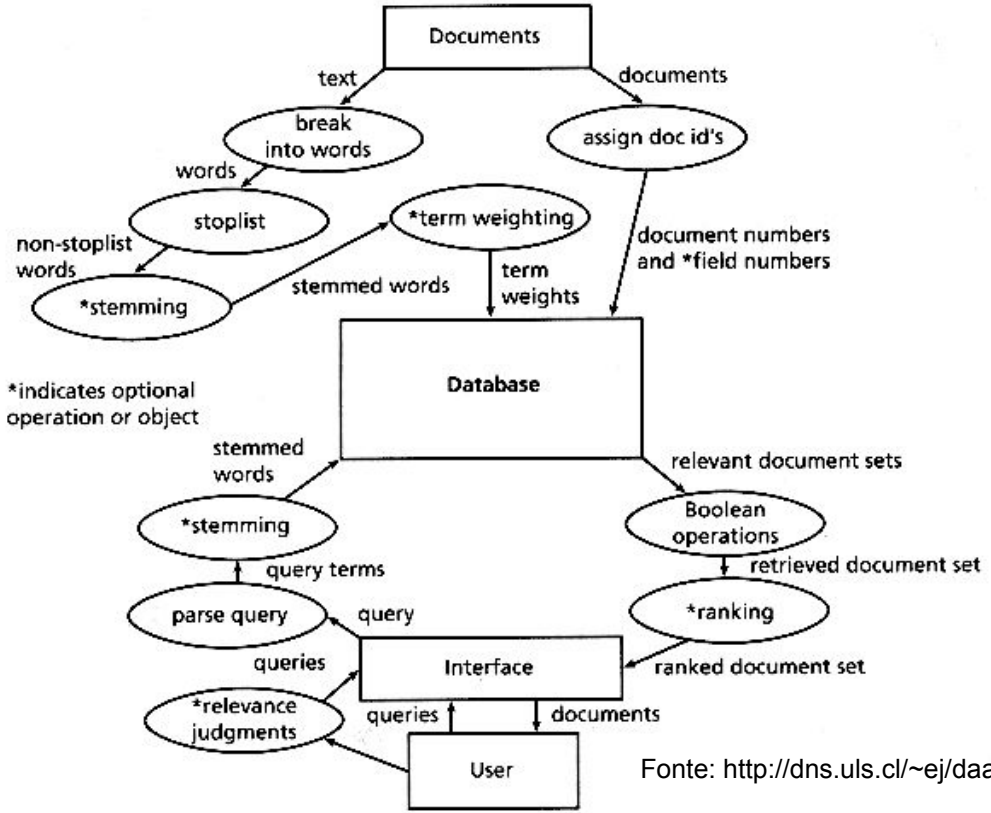
- Ler o texto
- Tokenizar e aplicar pré-processamento aprendido em sala
- Identificar os sentimentos a serem classificados (negativo, positivo, ...)
- Separar base de dados em treino e teste (80 / 20 ou 70 / 30)
- Classificar com algum algoritmo de aprendizado (Ex: Naive bayes)
- Medir a acurácia do acerto

Como aplicar análise de sentimento

Análise de sentimento com NLTK

- http://www.nltk.org/howto/sentiment.html
- https://pythonprogramming.net/tokenizing-words-sentences-nltk-tutorial/
- naive bayes: http://scikit-learn.org/stable/modules/naive_bayes.html

Recuperação da Informação



Fonte: http://dns.uls.cl/~ej/daa_08/Algoritmos/books/book5/chap01.htm

Trabalho em grupo

Ler a base de dados de fibrose cística:

- https://github.com/raulsenaferreira/Systems-Engineering/tree/master/BRI/Work_3/data

Ler o conjunto de queries e rodar no buscador construído durante as aulas do curso

- https://github.com/raulsenaferreira/Systems-Engineering/blob/master/BRI/Work_3/queries.csv

Trabalho em grupo

Comparar resultado com o resultado já gerado pelo Lucene

- https://github.com/raulsenaferreira/Systems-Engineering/tree/master/BRI/Work_3/ResultsLucene

Plotar o gráfico de 11 pontos de precision e recall. Ex:

- https://github.com/raulsenaferreira/Systems-Engineering/blob/master/BRI/Work_3/MetricResults/interpolated-precision-recall-11pts.pdf

Trabalho em grupo

Métricas restantes para o trabalho em grupo:

- Métrica de similaridade
  - Cosseno
- Métrica de acurácia
  - Interpolação de 11 pontos de recall e precision

Métodos para ler xml, csv, e exemplo de implementação das métricas podem ser encontradas no repositório

- https://github.com/raulsenaferreira/Systems-Engineering/tree/master/BRI/Work_3

Temas para o exercício 5

Mineração de texto

Solr

Elasticsearch

Aprendizado de máquina

Exercício

Enviar arquivo contendo 2 páginas falando sobre o tema escolhido

- raul.ferreira@prof.infnet.edu.br
- Colocar no assunto do email “Trabalho 5 - Turma de quinta-feira”
- Colocar identificação no corpo do email (nome e sobrenome)

Trabalho individual. Trabalhos copiados = zero

Próxima aula

Maquina virtual com alguns programas já instalados

- Lucene
- Apache Solr
- Elasticsearch

Exercício prático com as 3 ferramentas