Indexação e Tratamento de Dados Heterogêneos: Variedade

Mecanismos de Busca

Conteúdo do Bloco

Indexação

Recuperação de Informação

Solr

Elasticsearch

Aula 3

Conhecendo o Lucene

Conhecendo o Apache Solr

Conhecendo o Elasticsearch

Relembrando

Métricas de similaridade

- Cosseno
- Coeficiente de Jaccard
- Coeficiente de Pearson
- Distância Euclidiana

Relembrando

Métricas de acurácia

- Precision and recall
- Interpolação de 11 pontos de recall e precision
- Precision @ K
- Mean Average Precision (MAP)
- NDCG
- F1 score

Relembrando

Search engine ou motor de busca

- Recebe um conjunto de palavras-chave e realiza a busca por documentos que possuam coerência com tal conjunto
- Apresentar resultados dessas buscas de forma rápida e eficiente
- Composto basicamente de um sistema de indexação e armazenamento de documentos e um sistema de busca por tais documentos

Search engine

Existem motores de busca open-source para ajudar a montar o seu próprio buscador

Lucene: Um motor de busca famoso e de código aberto feito para criar toda a parte de lista invertida

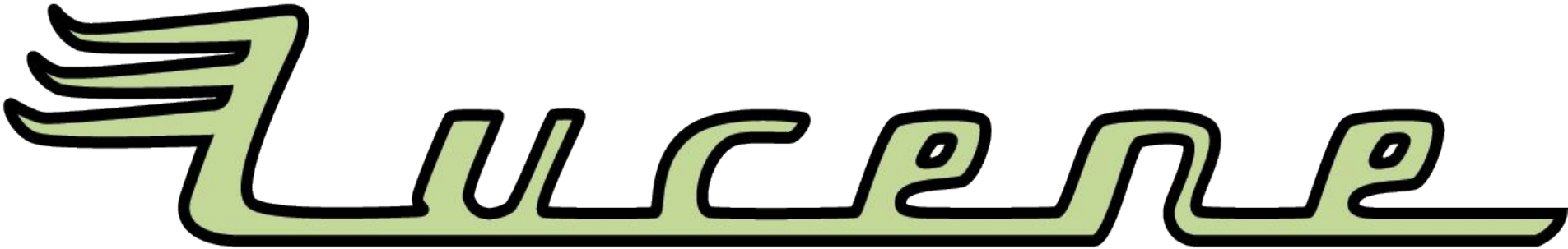
Lucene é usado dentro de sistemas muito conhecidos entre eles o Solr e o Elasticsearch

Lucene

Basicamente converte tudo para texto (dentro do possível), indexa e proporciona a busca de conteúdo dentro do córpus gerado

Com poucas linhas de código é possível indexar um texto

Com algumas dezenas de código é possível construir um buscador completo



Lucene

Está disponível em várias linguagens onde as duas formas mais comuns para se ter um rápido contato com o Lucene são:

- Versão em java
  - http://lucene.apache.org/core/
- Versão em python
  - Pylucene: http://lucene.apache.org/pylucene/

Versão preferida para nossas aulas será em python mas o aluno pode escolher qualquer uma das duas opções acima ou as demais existentes

Lucene

Instalando PyLucene: http://lucene.apache.org/pylucene/install.html

Testando no console com python2.7

>>> import lucene

>>> lucene.initVM()

Se nenhum erro aparecer então o lucene foi corretamente instalado!

Primeiros passos: http://www.lucenetutorial.com/lucene-in-5-minutes.html

Apache Solr

Servidor web pré-configurado que serve para fazer consultas

Usa o Lucene “por baixo dos panos”

http://lucene.apache.org/solr

Apache Solr

Instalação e primeiros passos com o Solr

- http://lucene.apache.org/solr/quickstart.html
- http://www.solrtutorial.com/
- https://wiki.apache.org/solr

Elasticsearch

Parecido com o Solr porém com outras funções como: Analytics

Também usa o Lucene como indexador

https://www.elastic.co/

Elasticsearch

Instalação e primeiros passos com o Elastic

- http://www.elasticsearchtutorial.com/
- https://www.elastic.co/guide/index.html

Solr x Elasticsearch

Solr é mais maduro

Elastic é mais simples

Solr teve preferência em ser usado por grandes empresas de big data:

- Cloudera, Hortonworks, MapR

Elastic é mais completo em termos de monitoramento e métricas

Solr é mais forte em problemas “text-search based”

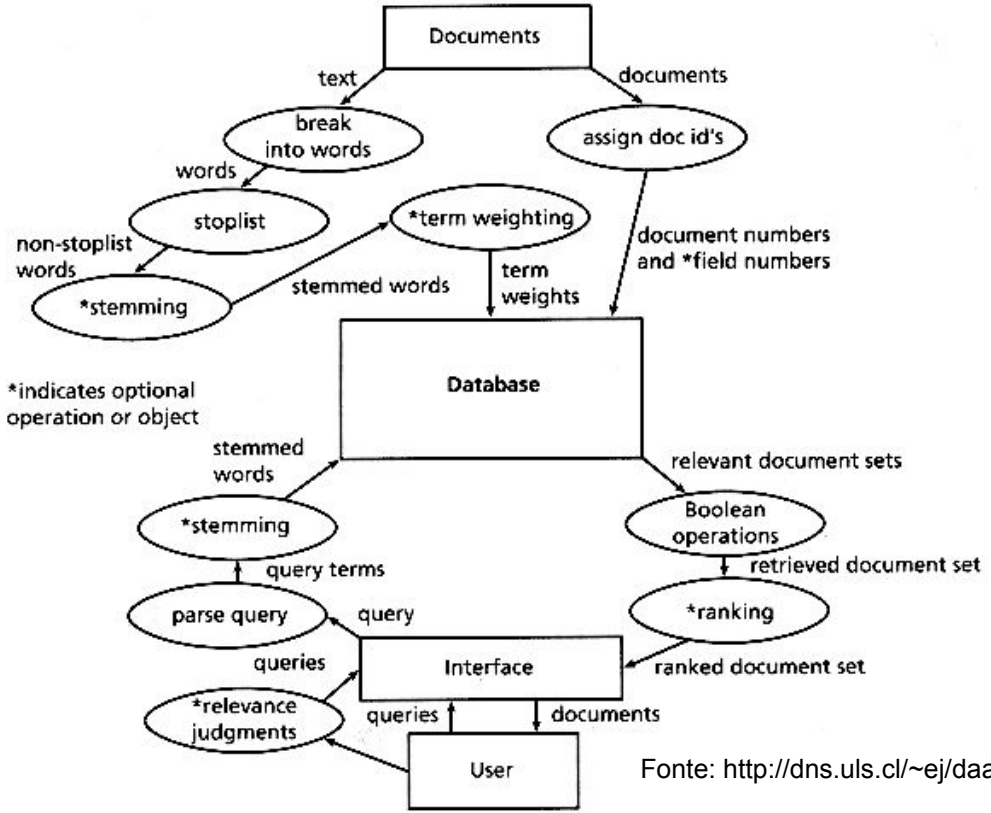
Elastic é mais forte em “analytics solutions based”

Solr x Elasticsearch

http://solr-vs-elasticsearch.com/

https://www.datanami.com/2015/01/22/solr-elasticsearch-question/

Recuperação da Informação



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

Exercício

Criar um indexador utilizando pesos (tf-idf):

- Ler arquivo que contém a lista invertida (Exercício 2)
- Criar um método de cálculo do tf-idf e atribuir o peso aos termos da lista invertida lidos do arquivo
    - Consultar aula anterior para ver como é o cálculo do TF-IDF e como esse valor é atribuído na lista invertida
- Salvar resultado final em um arquivo
- O resultado pode ser salvo da mesma forma como foi salvo no exercício anterior, por exemplo:
    - termo;[doc1: 2.98, doc2: 1.74]

Exercício

Sugestão para estrutura a ser salva em arquivo

- Processador de query tem a mesma lógica do exercício anterior

Enviar arquivo (zipado) com código do exercício

- raul.ferreira@prof.infnet.edu.br
- Colocar no assunto do email “Trabalho 3 - Turma de quinta - Infnet”
- Colocar identificação no corpo do email (nome e sobrenome)

Trabalho individual. Trabalhos copiados = zero

Próxima aula

Brincando com o Lucene

Aprofundando um pouco mais no Solr e no Elastic

Continuar a construção do nosso buscador