Information Retrieval

Exercise

Christian Spurk
cspurk@dfki.de

German Research Center for Artificial Intelligence (DFKI)
Language Technology Lab

Language Technology 1
2012-11-07

Exercise Solutions

Exam 1 – Computation of TF-IDF

$\text{idf}_{car} = 0.74, \text{idf}_{auto} = 0.51, \text{idf}_{best} = 0.74$

		Doc1	Doc2	Doc3	Doc4	Doc5
car	tf	0.38	0	0	0.42	1
	tf-idf	0.28	0	0	0.31	0.74
auto	tf	1	1	0	1	0
	tf-idf	0.51	0.51	0	0.51	0
best	tf	0	0.17	1	0	0.08
	tf-idf	0	0.13	0.71	0	0.06

		Doc6	Doc7	Doc8	Doc9	Doc10
car	tf	0	0	1	1	0.1
	tf-idf	0	0	0.74	0.74	0.07
auto	tf	0	0.75	0.5	0.38	1
	tf-idf	0	0.38	0.26	0.19	0.51
best	tf	1	1	0	0.25	0
	tf-idf	0.74	0.74	0	0.19	0

Exam 2 – Inverted Lists

Index Term	df			
1990	1	→ Doc4, 1		
2006	2	→ Doc1, 1	Doc3, 1	
champion	1	→ Doc1, 1		
germany	3	→ Doc2, 1	Doc3, 1	Doc4, 1
italy	3	→ Doc1, 1	Doc2, 1	Doc4, 1
played	1	→ Doc2, 1		
semifinal	3	→ Doc2, 1	Doc3, 1	Doc4, 1
won	1	→ Doc4, 1		
world	1	→ Doc1, 1		

Information Retrieval in Practice: Apache Lucene/Solr

What is Apache Lucene?

- a fast, reliable and feature-rich text search engine
- a well-documented, cross-platform Java library
- free and open-source software (Apache License 2.0)
- widely-used – often indirectly through Apache Solr
- actively developed under the umbrella of the Apache Software Foundation

→ `http://lucene.apache.org/`

Lucene's Basic Features

- scalable, high-performance indexing (over 95 GB/hour on modern hardware)
- powerful, accurate and efficient search algorithms
- fielded searching
- many different text analyzers for improved searches, for example:
  - shallow string normalization such as lower-casing, removal of certain characters, etc.
  - tokenization
  - phonetic analysis
  - morphological analysis
  - stemming
  - stop-word removal
  - synonym expansion
- different query types, e.g., phrase queries, wildcard queries, proximity queries, range queries

What is Apache Solr?

- a fast, reliable and feature-rich text search platform
- a Java web application wrapper around Apache Lucene
- free and open-source software (Apache License 2.0)
- widely-used (we'll come back to this in a minute)
- actively developed under the umbrella of the Apache Software Foundation

→ `http://lucene.apache.org/solr/`

Solr vs. Lucene

Solr adds commonly needed “convenience” features around the bare-bones Lucene search engine:

- a full-text search server available as a web service
- powerful external XML configuration

→ easy to use from virtually any programming language

- hit highlighting
- faceted search and filtering
- rich document handling (e.g., MS Word, PDF)
- geospatial search
- spelling suggestions for user queries
- “More Like This” suggestions
- plus many other (enterprise) features such as distributed search and index replication, dynamic clustering, multiple search indices, highly configurable and user extensible caching, etc.

Websites and Companies Using Solr and Lucene




. . . and many more:

`http://wiki.apache.org/solr/PublicServers`

Tf-idf in Lucene

In the default scoring implementation of Lucene:

- the Boolean Model (BM) is combined with the Vector Space Model (VSM): documents “approved” by the BM are scored by the VSM
- the weights of the vectors in the VSM are tf-idf values
- Cosine Similarity is used for scoring indexed documents relative to a search query
- document and field boosts can be applied at index time
- query, subquery and query term boosts can be applied at query time

More details:
`http://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html`

Apache Lucene/Solr Demo