Создание поисковой системы

Дмитрий Дзема
Google Commerce Search

$50 000
IBM OmniFind

$96,800
Apache Lucene
Apache Solr
Sphinx
Xapian
Crawling
Indexing
Searching
Ranging
Indexing
Tokenizing
Stop words
Normalization
Lemmatization
Stemming
am, is, are → be

car, cars, car’s → car
colour

color
7/10/2010

July 10, 2010

October 7, 2010
Index
Inverted document index
Searching
Boolean query
Relevance models
Ranging
Vector space model

$$\text{coord}(q, d) \ast qn(q) \sum_{t \in q} [tf(t \in d) \ast idf(t)^2 \ast \text{norm}(t, d)]$$
PageRank
Java
REST

HTTP + XML/JSON
Documents
Fields
Schema
Clustering
Highlighting
More Like This
Spellchecker
Auto-suggesting
Top results configurations

Faceting