ConceptsAndDefinitions

Lucene Concepts and Definitions

This page contains concepts and definitions related to Lucene. It is not a substitute for knowledge in InformationRetrieval.

Definitions

Please keep in alphabetical order when editing.

Analyzer - Lucene class used for preparing text for indexing. Most applications can use the StandardAnalyzer for English and latin based languages.

Payloads - A payload is an array of bytes stored at one or more term positions

Snowball Stemmers - The Snowball Stemmers are third party implementation of several stemmers that have been hooked into Lucene to help with indexing. See the Snowball website for more info.

Stemmer - From Wikipedia Stemmer: "A stemming algorithm, or stemmer, is a computer program or algorithm for reducing inflected (or sometimes derived) words to their stem, base or root form — generally a written word form." Stemmers are often used to reduce the search space and index size. Often times a user searching for "widgets" is interested in documents that contain the term "widget".

Core Classes

Document

A Lucene Document is a record in the index. A Document has a list of fields; each field has a name and a textual value.

Term

A Term is Lucene's unit of indexing. In western languages, a Term is often a word.

TermEnum

TermEnum is used to enumerate all terms in the index for a given field, regardless of which documents the terms occur in (or where they occur).

Some query subclasses are implemented by enumerating terms that match a pattern, and building a large OR query from the enumeration. E.g. WildcardQuery, PrefixQuery, RangeQuery.

See LuceneFAQ, *How do I retrieve all the values of a particular field that exists within an index, across all documents?* which also includes sample code.

TermDocs

Unlike TermEnum (see above), TermDocs is used to identify which documents contain a given Term. TermDocs also gives the frequency of the term in the document.

TermFreqVector

A TermFreqVector (aka Term Frequency Vector or just Term Vector) is a data structure containing a given Document's term and frequency information and can be retrieved from the IndexReader only when Term Vectors are stored during indexing.

Directory

IndexReader

IndexSearcher