Apache Lucene - Lucene Sandbox

Otis Gospodentic

Table of contents

1. Lucene Sandbox

Lucene project also contains a workspace, Lucene Sandbox, that is open to all Lucene committers, as well as a few other developers. The purpose of the Sandbox is to host various third party contributions, and to serve as a place to try out new ideas and prepare them for inclusion into the core Lucene distribution.
Users are free to experiment with the components developed in the Sandbox, but Sandbox components will not necessarily be maintained, particularly in their current state.

You can access the Lucene Sandbox repository at
http://svn.apache.org/repos/asf/lucene/java/trunk/contrib/.

1.1. Snowball Stemmers for Lucene

This project provides pre-compiled versions of the Snowball stemmers for Lucene.

The repository for the Snowball contribution.

Background information on Snowball, which is a language for stemmers developed by Martin Porter.

1.2. Analyzers, Tokenizers, Filters

Contributed Analyzers, Tokenizers, and Filters for various languages.

The repository for the Analyzers contribution.

1.3. Ant

The Ant project is a useful Ant task that creates a Lucene index out of an Ant fileset. It also contains an example HTML parser that uses JTidy.

The repository for the Ant contribution.

1.4. WordNet/Synonyms

The Lucene WordNet code consists of a single class which parses a prolog file from the WordNet site that contains a list of English words and synonyms. The class builds a Lucene index from the synonyms file. Your querying code could hit this index to build up a set of synonyms for the terms in the search query.

More information on the Lucene WordNet package. WordNet is an online database of

English language words that contains synonyms, definitions, and various relationships between synonym sets.

The repository for the WordNet module.

1.5. Lucli - Lucene Command-line Interface

The Lucli application allows index manipulation from the command-line.

The repository for the Lucli contribution.

1.6. Term Highlighter

A small set of classes for highlighting matching terms in search results.

The repository for the Highlighter contribution.

1.7. Javascript Query Constructor

Javascript library to support client-side query-building. Provides support for a user interface similar to Google's Advanced Search.

The repository for the Javascript Query Constructor files.

1.8. Javascript Query Validator

Javascript library to support client-side query validation. Lucene doesn't like malformed queries and tends to throw ParseException, which are often difficult to interpret and pass on to the user. This library hopes to alleviate that problem.

The repository for the Javascript Query Validator files.

1.9. High Frequency Terms

The miscellaneous package is for classes that don't fit anywhere else. The only class in it right now determines what terms occur the most inside a Lucene index. This could be useful for analyzing which terms may need to go into a custom stop word list for better search results.

The repository for miscellaneous classes.

1.10. InstantiatedIndex

RAM-based index that enables much faster searching than RAMDirectory.

The repository for instantiated index.