Data Management : Lab 1

Antoine Friant

11 octobre 2018

1 Understanding the Lucene API

1.1 Does the command line demo use stopword removal ?

Yes, because the query "a an or and at query response" returned the exact same results as "query response". If it weren't using stopword removal, we'd get less results when using stopwords.

1.2 "Does the command line demo use stemming ?"

No, because the queries "query" and "queries" don't give the same number of results.

1.3 "Is the search of the command line demo case insensitive ?"

Yes, because "QUERY" and "query" give the same results.

1.4 "Does it matter whether stemming occurs before or after stopword removal ?"

It only matters if stopwords find themselves grouped with non stopwords after the stemming step. If such a thing is possible, then we need to remove stopwords before stemming. If we don't, words from the query could be ignored or stopwords could change the result of the query.

2 Indexing and Searching the CACM collection

2.1 Explain which field type can be used for id, title, summary and author

We need an integer for "id" so LongPoint or IntPoint. The others need to be TextField to perform full-text searches.