Assignment-1

ILS-Z 534: Search

By, Supreeth Keragodu Suryaprakash (skeragod)

Task 1:

1. How many documents are there in this corpus? 2. Why different fields are treated with different kinds of java class? i.e., StringField and TextField are used for different fields in this example, why?

   Solution:

   1. Total Number of documents in the corpus: 84474.
   2. We have used String Fields DOCNO. String field doesn't tokenize but just indexes the given string. This will be mostly used to represent unique fields such as Phone, Email ID or zip code.
   3. We have used TextField for text areas. Text field will both tokenize and indexes the given string. Since the text field contains text in the trecText files, Tokenization and Indexation are both important.

Task 2:

Analyzer	Tokenization Applied?	How many tokens are there in the filed?	Stemming applied?	Stop Words removed?	How many terms in the dictionary?
Keyword	No	84474	No	No	84061
Simple	Yes	37330144	No	No	169981
Stop	Yes	26216475	No	Yes	169948
Standard	Yes	26649680	No	Yes	233384

Bonus Question:

Yes, It will group commonly occurring n-grams which are frequent and will then index it.

References and Discussions:

- http://wiki.apache.org/lucene-java/FrontPage?action=show&redirect=FrontPageEN
- http://stackoverflow.com/questions/17527741/what-is-the-default-list-of-stopwords-used-in-lucenes-stopfilter
- Had a high level discussion about concepts in IR with Suhas, Raghuveer and Yash.