SEARCH (Z534)
FALL 2017
Assignment I

Snehal Vartak

Task 1:

Code File: generateIndex.java

Instructions to run the code –

Update the path for the corpus and index by modifying the value of variables "indexPath" and "corpusPath" in the code.

1. Number of Documents in the Corpus – 84474
2. A String field is a field such as an email field, which you do not want to be split i.e the string field will be considered as one single term. Whereas the TextField is a field that we want to index, analyze and search. String field won't be analyzed.

Task 2:

Code File: indexComparison.java

Instructions to run the code –

Update the path for the corpus and index by modifying the value of variables "indexPath" and "corpusPath" in the code.

Comparison of Different Analyzers:

Analyzer	Tokenization Applied	How many tokens are there for the field?	Stemming applied?	Stop words removed?	How many terms are there in the dictionary?
KeywordAnalyzer	No	84474	No	No	84043
SimpleAnalyzer	Yes	37330144	No	No	169981
StopAnalyzer	Yes	26216475	No	Yes	169948
StandardAnalyzer	Yes	26649680	No	Yes	233384