Z534 Search Assignment 1 Report

Task 1:

1. How many documents are there in this corpus?

Total number of documents in the corpus: 84474.

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?
   1) There are two kinds of fields in used in this assignment: StringField and TextField.
   2) StringField treats the input as an immutable string making it perfect for fields such as document ID (DOCNO in our case), so we don't need to worry about the change of this field by the tokenizer or stemmer. The search for such field is exact match or not match.
   3) TextField treats the input as text expecting to work with proper tokenizer, stemmer and analyzer. In our case, TEXT is stored in TextField so that the whole input will be cut into words which can be individually treaded during the index and search phase.

Task 2:

Analyzer	Tokenization applied?	How many tokens are there for this field?	Stemming applied?	Stop words removed?	How many terms are there in the dictionary?
KeywordAnalyzer	No	84474	No	No	84061
SimpleAnalyzer	Yes	37330144	No	No	169981
StopAnalyzer	Yes	26216475	No	Yes	169948
StandardAnalyzer	Yes	26649680	No	Yes	233384

In the second task, I tried different Analyzers: KeywordAnalyzer, SimpleAnalyzer, StopAnalyzer and StandardAnalyzer.

The KeywordAnalyzer simply treats the whole input as a single token, so there is no tokenization, stemming or stop words removal applied. For the rest of the analyzers, tokenization is surely applied and we can test if stop word removal is applied easily.

To test if the stop words are removed, simply search for words like “the”, “and” and “here”. If no result is returned, then the stops words are removed.

The SimpleAnalyzer uses LetterTokenizer with LowerCaseFilter doing some simple tokenization and text process work.

The StopAnalyzer add StopFilter compared to SimpleAnalyzer.

The StandardAnalyzer implements StandardTokenizer with StandardFilter, LowerCaseFilter and StopFilter along with a list of English stop words doing a much fancier processing work.

No of the analyzers applies stemming.

Appendix:

Program output on statistics of different analyzers (indexComparison.java):

```
Begin to generate index for Keyword Analyzer
Total number of documents in the corpus: 84474
Number of documents containing the term "new" for field "TEXT": 0
Number of occurrences of "new" in the field "TEXT": 0
Size of the vocabulary for this field: 84061
Number of documents that have at least one term for this field: 84474
Number of tokens for this field: 84474
Number of postings for this field: 84474

Begin to generate index for Simple Analyzer
Total number of documents in the corpus: 84474
Number of documents containing the term "new" for field "TEXT": 38618
Number of occurrences of "new" in the field "TEXT": 83726
```

```
Size of the vocabulary for this field: 169981
Number of documents that have at least one term for this field: 84456
Number of tokens for this field: 37330144
Number of postings for this field: 18973889

Begin to generate index for Stop Analyzer
Total number of documents in the corpus: 84474
Number of documents containing the term "new" for field "TEXT": 38618
Number of occurrences of "new" in the field "TEXT": 83726
Size of the vocabulary for this field: 169948
Number of documents that have at least one term for this field: 84456
Number of tokens for this field: 26216475
Number of postings for this field: 17119173

Begin to generate index for Standard Analyzer
Total number of documents in the corpus: 84474
Number of documents containing the term "new" for field "TEXT": 38604
Number of occurrences of "new" in the field "TEXT": 83642
Size of the vocabulary for this field: 233384
Number of documents that have at least one term for this field: 84456
Number of tokens for this field: 26649680
Number of postings for this field: 18049815
```