SEARCH Assignment #1

Introduction:

Indexing is the most important step for Information Retrieval.

In this assignment, we use Apache Lucene 6.2 to create an index, and run different analyzers such as:

1) StandardAnalyzer
2) SimpleAnalyzer
3) StopAnalyzer
4) KeywordAnalyzer

We have been given a dataset name CORPUS: **AP89**

The Code has below components:

- **IndexComparison.java:** This java file is for Task 2 which creates indexes using different analyzers such as StandardAnalyzer, SimpleAnalyzer, StopAnalyzer, KeywordAnalyzer and shows some statistics about them such as Number of terms and tokens.
- **Generateindex.java:** This file is for Task 1 where an index is generated with "DOCNO", "HEAD", "BYLINE", "DATELINE", "TEXT" fields, given an input directory containing *.trectext files on which indexing is needed.
- **FileReader.java:** I have created this additional java file to make the code readable and reduce the complexity. From one of the input parameter to the application, this reads *.trectext files.

Answers to the assignment question:

Task 1: Generating Lucene Index for Experiment Corpus (AP89)

1) **Question: How many documents are there in this corpus?**

   Answer: 84474

2) **Question: 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?**

   Answer: Text is content, article, post, document and anything read by human. TextField needs to be indexed for search and retrieval. For this purpose, it is analyzed, indexed and optionally stored. For example, this would be used on a 'body' field, that contains the bulk of a document's text. Whereas in String Field the field is indexed but not tokenized i.e. the entire String value is indexed as a single token. For example, this might be used for a 'country' field or an 'id' field, or any field that we intend to use for sorting or access through the field cache.

Task 2: Test different analyzers

Analyzer	Tokenization applied?	How many tokens are there for this field?	Stemming applied?	Stop words removed?	How many terms are there in the dictionary?
Keyword Analyzer	**No**	**84474**	**No**	**No**	**84061**
Simple Analyzer	**Yes**	**37316074**	**No**	**No**	**169981**
Stop Analyzer	**Yes**	**26202405**	**No**	**Yes**	**169948**
Standard Analyzer	**Yes**	**26635610**	**No**	**Yes**	**233384**

Issues:

The analyzers such as SimpleAnalyzer, StopAnalyzer, KeywordAnalyzer has been moved from Lucene Core jar to Lucene Common Analyzers jar.

This required import of Lucene Common Analyzers.jar in the Package.

References:

XML DOM Parser was implemented using https://www.journaldev.com/898/read-xml-file-java-dom-parser as a reference.