Hui Chen Z534: Search
Due Date: September 28, 2016

Assignment 1: Indexation

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

Please answer the following questions:

1. How many documents are there in this 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?

A text field is treated as sequence of terms that has been tokenized while a string field is treated as a single term and punctuation and spacing is ignored for this field. So we will use TextField for a field that needs to be tokenized, lowercased, stemmed, and even stop words removed, such as the text of a document. And use StringField for a field that does not require tokenization, such as document ID.

Task 2: Test different analyzers

In this task, please generate Lucene index for AP89 with the four analyzers listed in the table below. Let's only work with the <TEXT> field for this question. Fill in the empty cells with your observation.

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	84,474	NO	NO	84,052
SimpleAnalyzer	YES	37,330,144	NO	NO	169,981
StopAnalyzer	YES	26,216,475	NO	YES	169,948
StandardAnalyzer	YES	26,649,680	NO	YES	233,384