ILS-Z 534 Search
Assignment 1

Marshal Patel
marshalp@indiana.edu

10-05-2015

1. (a) There are **84474** documents in total.

   (b) StringField is used are used for the fields we **do not** wish to tokenize. For eg. DOCNO in our case. On the other hand, TextField is used for the fields for which we wish to generate tokens. Like TEXT field in this case.

   Therefore, if we need exact match to the query we index the text as StringField, else we tokenize the text and index the terms using TextField.

2. Following statistics were obtained after indexing with 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?**
KeywordAnalyzer	NO	84474	NO	NO	84049
SimpleAnalyzer	YES	37330144	NO	NO	169981
StopAnalyzer	YES	26216475	NO	YES	169948
StandardAnalyzer	YES	26649680	NO	YES	233384