Z534 Search Assignment 1: Indexation

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?
Fields treated with java class TextField while indexing, will have the field text tokenized (split into a stream of tokens) by Lucene. We use TextField when we want the field text to be tokenized. No tokenization is applied to fields treated with java class StringField, the field values are passed as they are while indexing. We use StringField for fields whose values such as a phone number or ID shouldn't be split but passed as single identifier.

Analyzer	Tokenization applied?	How many tokens for this field?	Stemming applied?	Stop words removed?	Number of terms in dictionary?
Keyword analyzer	No	84474	No	No	84054
Simple analyzer	Yes	34843730	No	No	932081
Stop analyzer	Yes	25089642	No	Yes	932048
Standard analyzer	Yes	25405918	No	Yes	1098687

Executed output

Total number of documents in the corpus with Standard Analyzer:84474
Total number of documents in the corpus with Keyword Analyzer:84474
Total number of documents in the corpus with Simple Analyzer:84474
Total number of documents in the corpus with Stop Analyzer:84474

Number of documents containing the term "new" for field "TEXT": 35533
Number of occurrences of "new" in the field "TEXT": 71657

Number of terms in the dictionary with Standard Analyzer:1098687
Number of terms in the dictionary with Keyword Analyzer:84054
Number of terms in the dictionary with Simple Analyzer:932081
Number of terms in the dictionary with Stop Analyzer:932048

Number of tokens for this field with Standard Analyzer:25405918

Number of tokens for this field with Keyword Analyzer: 84474
Number of tokens for this field with Simple Analyzer:34843730
Number of tokens for this field with Stop Analyzer:25089642

Number of postings for this field with Standard Analyzer: 18497488
Number of postings for this field with Keyword Analyzer: 84474
Number of postings for this field with Simple Analyzer: 19403550
Number of postings for this field with Stop Analyzer: 17612237

Number of documents that have at least one term for this field with Standard Analyzer: 84456
Number of documents that have at least one term for this field with Keyword Analyzer: 84474
Number of documents that have at least one term for this field with Simple Analyzer: 84456
Number of documents that have at least one term for this field with Stop Analyzer: 84456