Assignment 1 – Indexation – Report

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?**
→ In this example, the document number (DOCNO) was saved as StringField whereas rest of the fields HEAD, BYLINE, DATELINE, and TEXT were saved as TextField.
This approach was used because a document number is the attribute which is used to identified a particular document. So, the document number should not be broken down into token because if done so, the objective (of identifying the document) will be lost. Hence, the document number is indexed as a whole using StringField. Considering information enclosed rest of the tags, it needs to be split into tokens and indexed, and for this purpose, TextField is used.

Observations for different analyzers:

Analyzer	Tokenization Applied?	How many fields are there for this field?	Stemming applied?	Stop words removed?	How many terms are there in the dictionary?
StandardAnalyzer	YES	25410087	NO	YES	1098704
SimpleAnalyzer	YES	34848155	NO	NO	932098
StopAnalyzer	YES	25093814	NO	YES	932065
KeywordAnalyzer	NO	105383	NO	NO	102434