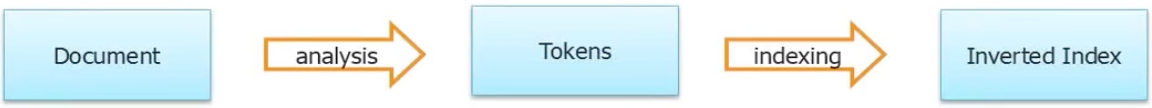
Assignment 1

1. How many Documents are there in Corpus
   - The Number of documents in corpus are 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?



   - After parsing the documents from corpus, the work of analyzer to tokenize the text of the document.
   - If a field is associated with Stringfield class the analyzer does not perform any tokenization process on it. For example, email-ID or DocID are needed as a complete string for indexing and not to be separated in different terms.
   - For the Textfield class the analyzer performs tokenization with help of a delimiter and separates it into different terms.

3. 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 dictionary
Keyword Analyzer	No	84474	No	No	83320
Simple Analyzer	Yes	32303248	No	No	889017
Stop Analyzer	Yes	23239767	No	Yes	888984
Standard Analyzer	Yes	23522724	No	Yes	1043576

Git Repository: https://github.iu.edu/sambavan/ambavanekar-information-retrieval.git