Task 1:

1. How many documents are there in this corpus?

Answer: ***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?

Answer:

StringField are used for content we do not want tokenized. StringField is usually used to store IDs, which in our case is the <DOCNO> tag.

We want the DOCNO to be as a whole, not be tokenized and split into terms. On the other hand, TextField is used for content we want to be tokenized. For e.g. <TEXT> tag in our data has huge text which we to tokenize and split into terms and then apply the index. TextField is never used to store IDs.

*Reference: Please find generateIndex.java*

Task 2:

Observations 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?** |
| **Keyword Analyzer** | No | 84474 | No | No | 84043 |
| **Simple Analyzer** | Yes | 37330144 | No | No | 169981 |
| **Stop Analyzer** | Yes | 26216475 | No | Yes | 169948 |
| **Standard Analyzer** | Yes | 26649680 | No | Yes | 233384 |

*Reference: Please find indexComparision.java*