**Name: Anudhriti Reddy Katanguri**

**Username: anukatan**

**Z 534 SEARCH**

1. **How many documents are there in the corpus?**

The total number of documents in the corpus: 84474.

1. **Why different fields are treated with the different kinds of the java classes? i.e., StringField and TextField are different in this example. Why?**

The major difference between the StringField and the TextField is that StringField has the filters whereas the TextField doesn’t. StringFields will give out the exact match results. Usually the fields that can be let go of the punctuation, spacing etc., before the actual index is built are used with the TextField. On the other hand, StringField never compromises, it preserves the case, punctuation and spacing.

In our assignment, DOCID uses StringField as it is unique whereas all the other fields that of TEXT, BYLINE, DATELINE, HEAD uses the TextField which contained the repetitions. Ultimately our goal is to create an index which costs us the least.

1. **Testing 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?** | **Observations** |
| StandardAnalyzer | Yes | 266496880 | No | Yes | 233384 | Removes spaces, punctuation. |
| SimpleAnalyzer | Yes | 37330144 | No | No | 169981 | It splits the text on anything that’s not a letter.  Results were very close with the stop analyzer. |
| StopAnalyzer | Yes | 26216475 | No | Yes | 169948 | Stop words were removed here. Pretty much the only difference between the simple and stop analyzer. |
| KeywordAnalyzer­­­­ | No | 84474 | No | No | 84601 | Considers all the text as a single token. |

**Task 1 OUTPUT:**

Parsing the files. This may take time.

Total number of documents in the corpus: 84474

Number of documents containing the term "new" for field "TEXT": 38604

Number of occurrences of "new" in the field"TEXT": 83642

Size of the vocabulary for this field: 233384

Number of documents that have at least one term for this field: 84456

Number of tokens for this field: 26649680

Number of postings for this field: 18049815

**Task 2 OUTPUT:**

STANDARD ANALYZER

Total number of documents in the corpus: 84474

Number of documents containing the term "new" for field "TEXT": 38604

Number of occurences of "new" in the field"TEXT": 83642

Size of the vocabulary for this field: 233384

Number of documents that have at least one term for this field: 84456

Number of tokens for this field: 26649680

Number of postings for this field: 18049815

SIMPLE ANALYZER

Total number of documents in the corpus:84474

Number of documents containing the term "new" for field "TEXT": 38618

Number of occurences of "new" in the field"TEXT": 83726

Size of the vocabulary for this field:169981

Number of documents that have at least one term for this field: 84456

Number of tokens for this field:37330144

Number of postings for this field:18973889

STOP ANALYZER

Total number of documents in the corpus:84474

Number of documents containing the term "new" for field "TEXT": 38618

Number of occurences of "new" in the field"TEXT": 83726

Size of the vocabulary for this field:169948

Number of documents that have at least one term for this field: 84456

Number of tokens for this field:26216475

Number of postings for this field:17119173

KEYWORD ANALYZER

Total number of documents in the corpus:84474

Number of documents containing the term "new" for field "TEXT": 0

Number of occurences of "new" in the field"TEXT": 0

Size of the vocabulary for this field:84061

Number of documents that have at least one term for this field: 84474

Number of tokens for this field:84474

Number of postings for this field:84474