ILS-Z534 - Assignment 1: Indexing

Prahasan Gadugu

Q1) How many documents are there in this corpus?

The number of documents in the AP89 Corpus is 84474.

Q2) 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?

Text and String may seem similar from a general perspective. But while dealing with Information Retrieval especially the text in document is considered as Text. Text may be an article, information in a website etc., something that needs to be analyzed.

String is different from text, it may be the URL of the website, identity number etc., We generally do not analyze it.

That's the reason in Lucene, the one needs to be analyzed and searched for is indexed as Text by abstracting it under Text Field, the one that is not indexed like DOCNO, is being added as StringField. As TEXT, HEAD,BYLINE and DATELINE we indexed them by abstracting them under TextField.

Task 2 observations:

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	84061
Simple Analyzer	Yes	37330144	No	No	169981
Stop Analyzer	Yes	26216475	No	Yes	169948
Standard Analyzer	Yes	26649680	No	Yes	233384

Output:

generateIndex.java output,

```
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
```

indexComparison.java output,

```
Keyword Analyzer
------------------------------------------------------------------------------
Total number of documents in the corpus:84474
Number of documents containing the term "new" for field "TEXT": 0
Number of occurrences 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

Simple Analyzer
------------------------------------------------------------------------------
Total number of documents in the corpus:84474
Number of documents containing the term "new" for field "TEXT": 38618
Number of occurrences 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 occurrences 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

Standard Analyzer
------------------------------------------------------------------------------
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
```