Manikandan Murugesan
murugesm@indiana.edu
September 29, 2016

A1 - Indexation

ILS Z534 - Search

Task 1) Generating Lucene Index for Experiment Corpus (AP89)

Qn 1) How many documents are there in this corpus?

Answer:

There are a total of **84837** documents in this corpus.

Qn 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 helps us preserve punctuations, case and spacing. While TextField helps us in Text Analysis and Tokenizing. In this example, DOCNO is represented in StringField since we run exact matching queries on that.

```
Total number of documents in the corpus: 84837
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) Test 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	84837	No	No	84062
Simple Analyzer	Yes	37330144	No	Yes	169981
Stop Analyzer	Yes	26216475	No	Yes	169948
Standard Analyzer	Yes	26649680	No	Yes	233384

```
Indexing file - keyWord
.....................................................................................
keyWord indexing completed successfully
keyWord- Total number of documents in the corpus: 84837
keyWord- Number of documents containing the term "new" for field "TEXT": 0
keyWord- Number of occurrences of "new" in the field "TEXT": 0
keyWord- Size of the vocabulary for this field: 84062
keyWord- Number of documents that have at least one term for this field: 84837
keyWord- Number of tokens for this field: 84837
keyWord- Number of postings for this field: 84837
```

```
Indexing file - simple
.....................................................................................
simple indexing completed successfully
simple- Total number of documents in the corpus: 84837
simple- Number of documents containing the term "new" for field "TEXT": 38618
simple- Number of occurrences of "new" in the field "TEXT": 83726
simple- Size of the vocabulary for this field: 169981
simple- Number of documents that have at least one term for this field: 84456
simple- Number of tokens for this field: 37330144
simple- Number of postings for this field: 18973889
```

```
Indexing file - stop
.....................................................................................
stop indexing completed successfully
stop- Total number of documents in the corpus: 84837
stop- Number of documents containing the term "new" for field "TEXT": 38618
stop- Number of occurrences of "new" in the field "TEXT": 83726
stop- Size of the vocabulary for this field: 169948
stop- Number of documents that have at least one term for this field: 84456
stop- Number of tokens for this field: 26216475
stop- Number of postings for this field: 17119173
```

```
Indexing file - standard
.....................................................................................
standard indexing completed successfully
standard- Total number of documents in the corpus: 84837
standard- Number of documents containing the term "new" for field "TEXT": 38604
standard- Number of occurrences of "new" in the field "TEXT": 83642
standard- Size of the vocabulary for this field: 233384
standard- Number of documents that have at least one term for this field: 84456
standard- Number of tokens for this field: 26649680
standard- Number of postings for this field: 18049815
```