Vipul Munot (vipmunot)

Search Assignment 1

1. How many documents are there in this corpus?

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?

StringField = A field that is indexed but not tokenized: the entire String value is indexed as a single token. For example this might be used for a 'country' field or an 'id' field, or any field that you intend to use for sorting or access through the field cache.

In our scenario, we use StringField for DOCNO tag.

TextField = A field that is indexed and tokenized, without term vectors. For example this would be used on a 'body' field that contains the bulk of a document's text.

In our scenario, we use StringField for BYLINE, DATELINE, HEAD and TEXT tags.

Output of first task:

```
Output:
 Indexing to directory 'C:\Users\Ganesh\Desktop\Vipul\index_folder_true_0
Total Number of documents in Corpus is = 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 TEXT field:233384
Number of documents that have at least one term for TEXT field: 84456
Number of tokens for TEXT field:26649680
Number of postings for TEXT field:18049815
```

Analyzer	Tokenization applied?	How many tokens are there for this field?	Stemming applied?	Stop words removed?	How many terms are there in the dictionary?
KeywordAnalyzer	No	168948	No	No	84049
SimpleAnalyzer	Yes	74660288	No	No	169981
StopAnalyzer	Yes	26216475	No	Yes	169948
StandardAnalyzer	Yes	26649680	No	Yes	233384

Output of Second task:

Enter a query string

movies

Analyzing stats for analyzer: org.apache.lucene.analysis.standard.StandardAnalyzer@119d7047

Indexing to directory 'E:\IUB\Search\Assignment 1\index_folder_false_0

Hits: 1194

Total Number of documents in Corpus is = 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 TEXT field:233384

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

Number of tokens for TEXT field:26649680

Number of postings for TEXT field:18049815

Analyzing stats for analyzer: org.apache.lucene.analysis.core.SimpleAnalyzer@74a10858

Indexing to directory 'E:\IUB\Search\Assignment 1\index_folder_false_1

Hits: 2388

Total Number of documents in Corpus is = 168948

Number of documents containing the term"new" for field"TEXT": 77236

Number of occurences of"new" in the field"TEXT": 167452

Size of the vocabulary for TEXT field:169981

Number of documents that have at least one term for TEXT field: 168912

Number of tokens for TEXT field:74660288

Number of postings for TEXT field:37947778

Analyzing stats for analyzer: org.apache.lucene.analysis.core.StopAnalyzer@41629346

Indexing to directory 'E:\IUB\Search\Assignment 1\index_folder_false_0

Hits: 1194

Total Number of documents in Corpus is = 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 TEXT field:169948

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

Number of tokens for TEXT field:26216475

Number of postings for TEXT field:17119173

Analyzing stats for analyzer: org.apache.lucene.analysis.core.KeywordAnalyzer@1d371b2d

Indexing to directory 'E:\IUB\Search\Assignment 1\index_folder_false_1

Hits: 0

Total Number of documents in Corpus is = 168948

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 TEXT field:84049

Number of documents that have at least one term for TEXT field: 168948

Number of tokens for TEXT field:168948

Number of postings for TEXT field:168948