Z534 SEARCH

# ASSIGNMENT 1

# Indexation

## Introduction

Indexation is the first stage in information retrieval and search engine design. Indexing documents facilitates faster document retrieval. In this project we use Lucene to generate indexes and later compare different analysers and their approach to indexing, i.e if they use tokenization, stemming, stop word removal etc.

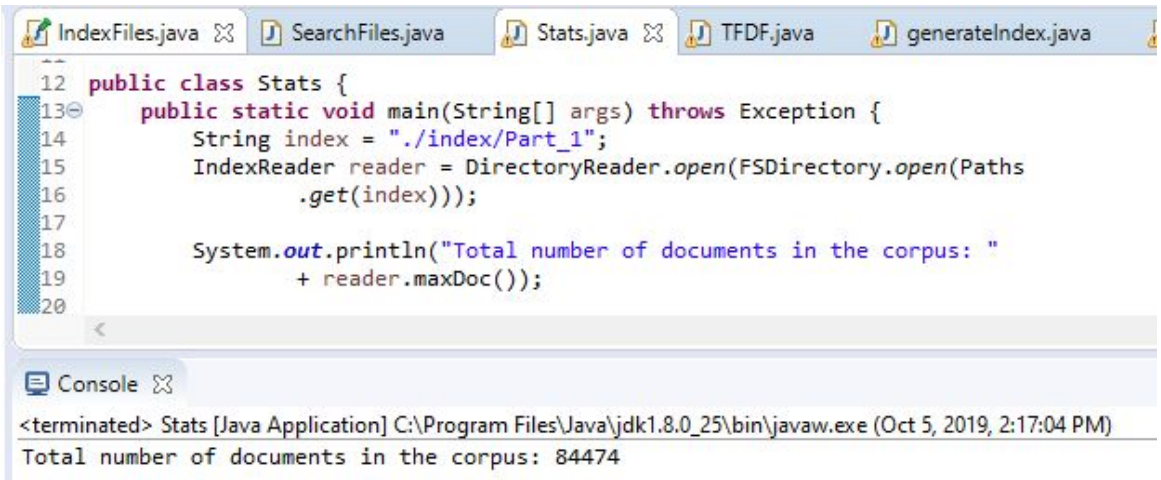
### Task 1 : Generating Lucene Index for Experiment Corpus (AP89)

In this experiment, we parse through the AP89 corpus to generate a lucene index on the following tags - 1.DOCNO, 2. HEAD, 3. BYLINE, 4. DATELINE, and 5. TEXT.

#### Questions?

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

To find the number of documents in the corpus, we can use IndexReader to read the indexes and use maxDoc() to get count.



In AP89 experiment corpus there are 84474 documents.

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?***

In Lucene, different fields are treated with different kinds of java classes so that analyzers can treat these differently. This makes a StringField different from a TextField, in the way they are handled while indexing. For example, StringFields are not tokenized and are handled as a single unit while TextFields are not. This enables StringFields to be used when we want to store Ids, emails, urls etc.

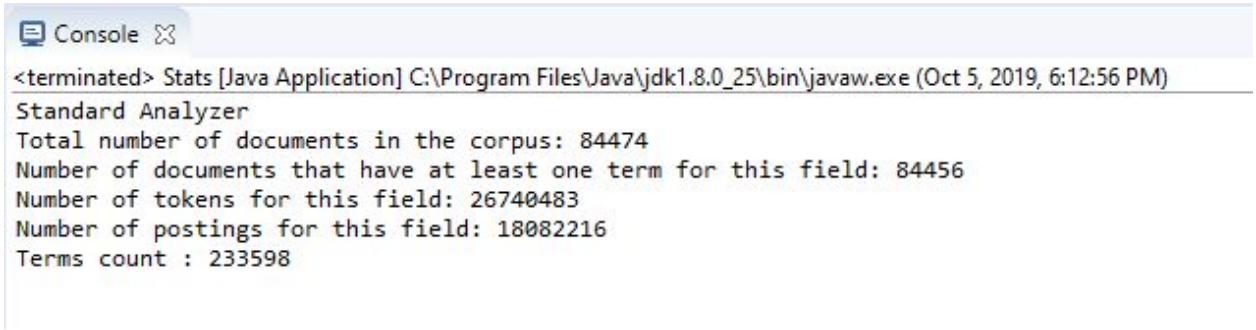
## Task 2: Test different analyzers

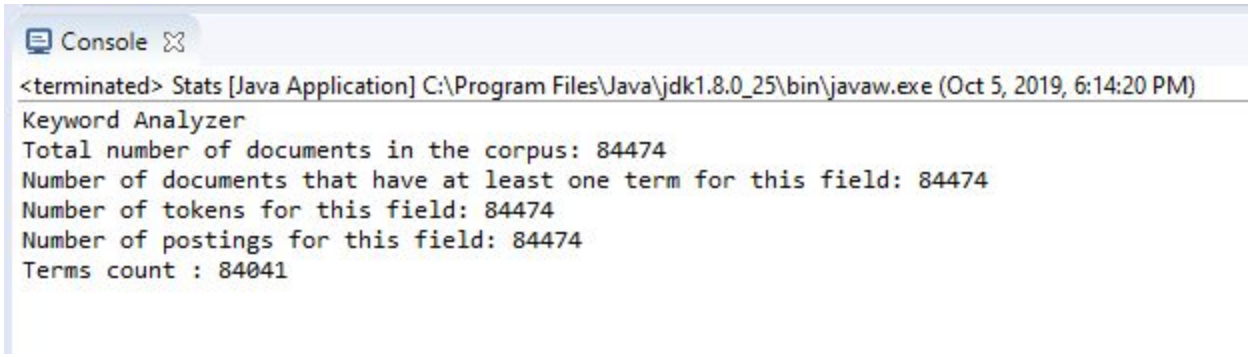
In this experiment we generate a lucene index for AP89 using four analysers - Keyword, Simple, Stop and Standard analyzers. The index is created with <TEXT> field.

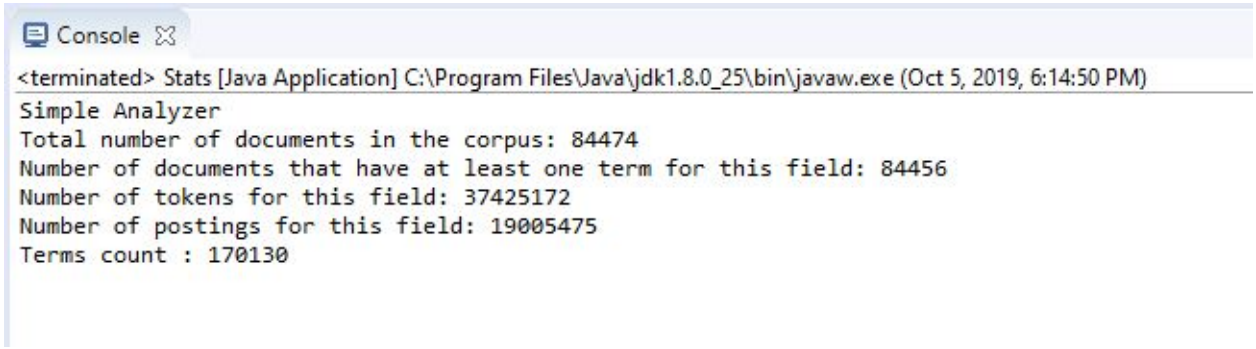
### Observations

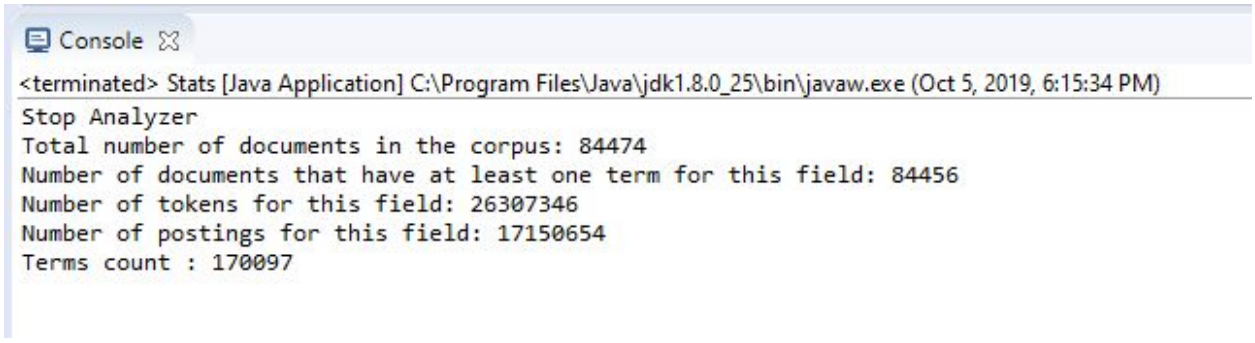
Following are the observations -

| Analyser | 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 | 84474 | No | No | 84041 |
| **SimpleAnalyzer** | Yes | 37425172 | No | No | 170130 |
| **StopAnalyzer** | Yes | 26307346 | No | Yes | 170097 |
| **StandardAnalyzer** | Yes | 26740483 | No | Yes | 233598 |


Console
<terminated> Stats [Java Application] C:\Program Files\Java\jdk1.8.0_25\bin\javaw.exe (Oct 5, 2019, 6:12:56 PM)
Standard Analyzer
Total number of documents in the corpus: 84474
Number of documents that have at least one term for this field: 84456
Number of tokens for this field: 26740483
Number of postings for this field: 18082216
Terms count : 233598


Console
<terminated> Stats [Java Application] C:\Program Files\Java\jdk1.8.0_25\bin\javaw.exe (Oct 5, 2019, 6:14:20 PM)
Keyword Analyzer
Total number of documents in the corpus: 84474
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
Terms count : 84041


Console
<terminated> Stats [Java Application] C:\Program Files\Java\jdk1.8.0_25\bin\javaw.exe (Oct 5, 2019, 6:14:50 PM)
Simple Analyzer
Total number of documents in the corpus: 84474
Number of documents that have at least one term for this field: 84456
Number of tokens for this field: 37425172
Number of postings for this field: 19005475
Terms count : 170130


Console
<terminated> Stats [Java Application] C:\Program Files\Java\jdk1.8.0_25\bin\javaw.exe (Oct 5, 2019, 6:15:34 PM)
Stop Analyzer
Total number of documents in the corpus: 84474
Number of documents that have at least one term for this field: 84456
Number of tokens for this field: 26307346
Number of postings for this field: 17150654
Terms count : 170097