ILS Z534 (Search) – Assignment 1

Yash Sumant Ketkar (yketkar@indiana.edu)

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

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

Total number of documents in the 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 cannot have any tokenization or analysis / filters applied, and will only give results for exact matches.

TextField usually have a tokenizer and text analysis attached, meaning that the indexed content is broken into separate tokens where there is no need for an exact match - each word / token can be matched separately to decide if the whole document should be included in the response.

Here, the StringField is ideal of <DocNo> As we will need an exact match while searching for the document by <DocNo>

The <TEXT> part of this data set is stored in TextField as it will be tokenized and we need to match separate tokens to check if the document should be included in response or not.

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?**
KeywordAnalyzer	No	84474	No	No	84061
SimpleAnalyzer	Yes	37330144	No	No	169981
StopAnalyzer	Yes	26216475	No	Yes	169948
StandardAnalyzer	Yes	26649680	No	Yes	233384

EXTRA QUESTION (SHINGLE 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: 5105569
- Number of documents that have at least one term for this field: 84456
- Number of tokens for this field: 62093664
- Number of postings for this field: 49038698

Output for Task 2:

```
Creating Index for: 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

Creating Index for: 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

Creating Index for: 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

Creating Index for: 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
```

Keyword Analyzer:

"Tokenizes" the entire stream as a single token. This is useful for data like zip codes, ids, and some product names.

Simple Analyzer:

An Analyzer that filters LetterTokenizer (a tokenizer that divides text at non-letters) with LowerCaseFilter (normalizes token text to lower case.).

Stop Analyzer:

Filters LetterTokenizer with LowerCaseFilter and StopFilter (removes stop words from a token stream.).

Standard Analyzer:

Filters StandardTokenizer (a grammar-based tokenizer constructed with JFlex.) with StandardFilter, LowerCaseFilter and StopFilter, using a list of English stop words.