Information Retrieval WS22/23

Programming Task P01

Task (a)

- The task was to implement a standard tokenizer and a whitespace tokenizer
- The key difference between a standard tokenizer and a whitespace tokenizer is the whitespace tokenizer tokenizes the document at every instance of a whitespace occurrence. The punctuations are kept as it is, for example, 'Berlin!' using a whitespace tokenizer remains as 'Berlin!' but the standard tokenizer gives 'Berlin' as an output.
- The task was implemented by using standardTokenizer and whitespaceTokenizer from Apache Lucene Library

Task (b)

- The task was to implement the standard tokenizer with case sensitive stopword filter as well as non- case sensitive stop word filter. For Example, if the filter is case sensitive, the stopword 'To' will remain in the document as 'T' is capitalized in 'To' whereas if the filter is not case sensitive 'To' will be removed irrespective of whether 'T' is capitalized or not
- The task was implemented by using standardTokenizer and StopFilter from Apache Lucene Library

Task (c)

- The task was to create a custom analyzer with standard tokenizer, lowercase filter, stopword filter and porter stemmer filter.
- The task was implemented with the help of apache lucene's custom analyzer and used all the above mentioned filters.

The output is as follows,

```
Input document : Today is sunny. She is a sunny girl. To be or not to be. She is in Berlin today. Sunny
Berlin! Berlin is always exciting!

Standard tokenizer output: [Today, is, sunny, She, is, a, sunny, girl, To, be, or, not, to, be, She, is, in,
Berlin, today, Sunny, Berlin, Berlin, is, always, exciting]

Standard tokenizer with stop word filter and ignoreCase=false (case sensitive): [Today, sunny, She, a,
sunny, girl, To, or, not, She, Berlin, today, Sunny, Berlin, Berlin, always, exciting]

Standard tokenizer with stop word filter and ignoreCase=true: [Today, sunny, She, a, sunny, girl, or,
not, She, Berlin, today, Sunny, Berlin, Berlin, always, exciting]

Whitespace tokenizer output : [Today, is, sunny., She, is, a, sunny, girl., To, be, or, not, to, be., She,
is, in, Berlin, today., Sunny, Berlin!, Berlin, is, always, exciting!]

Analyser output : [todai, sunni, she, a, sunni, girl, or, not, she, berlin, todai, sunni, berlin, berlin,
alwai, excit]
```