

IR Assignment 3

Aakash Khadka, Gowtham Premkumar, Max Neubauer

3.a)

A picture of the output can be found on the second page. In order to solve this assignment we build a CustomTokenizer class that removes punctuation and splits the text into lowercase tokens. Then we created a BiwordFilter class. In this class we override the incrementToken method. In this method we create biwords by appending previous and current tokens until there are no tokens left. Then we append them to a list. If the previous token is null (which is the case in the first iteration) a biword cannot be created and is hence not added to the list. In our buildBiWordList method, that utilizes the incrementToken method, to add the biword tokens to a custom list, in order to print them.

3.b)

In order to solve this assignment we build biword lists for the query and a document. Then we check if all of the queries biwords appear in the documents biwords. If all of them are present we print that the result is positive. Otherwise negative. And in order to validate our solution for task 4.3 a) we verified with a true positive and a false positive solution.

```
False Positive Index:[it is, is new, new york, york there, there is, is the, the york, york university]
Query Index:[new york, york university]
Result: Positive
False Positive Index:[new york, york max]
Query Index:[it is, is new, new york, york there, there is, is the, the york, york university]
Result: Negative
```

today is
is sunny
sunny she
she is
is a
a sunny
sunny girl
girl to
to be
be or
or not
not to
to be
be she
she is
is in
in berlin
berlin today
today sunny
sunny berlin
berlin berlin
berlin is
is always
always exciting