The first part (a) of the programming assignment is to implement Biword Tokenization and tokenize the following:

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!

Result after analyzing with biword tokenizer, standard tokenizer and lowercase filter, is given below:

[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]

The second part (b) of the assignment is to implement assignment 4.3(a), which will analyze the given query (here 'New York University') and return a false positive result / document. This was done by creating a BiwordFilterFactory class which extends lucene's 'TokenFilterFactory'.

Given query: New York University---> Analyzed with biword tokenizer, standard tokenizer and lowercase filter, is given below:

[new york, york university]

Given query string: New York University

Found 1 matches.

1. Most students from New York who apply to York University in Toronto are accepted for their high grades in academics and extra-curricular activities.

False Positive Case Identified