Status Report

-We used Python 3.7 for the project 1 and implemented in PyCharm(2019). We obtained correct results.

-We inserted correct values into document.txt and posting.txt

-There are 514 lines in document.txt.

-This is how we implemented:

1)We opened and read the "200\_title.txt". We converted all letter to lowercase and splitted each word. In line 25, we removed all punctuations. We gave all words to document id and inserted “word,documentid” pairs to terms list. In line 31, sorted both words and document-ids.

2)Lines 33 through 43, we we calculated how many times each word is used. if there is one more same word, we have increased df(document frequency) to one. We append word and df to terms1.

3)In terms3, we we have sorted elements it as follows: term, df, doc-id, tf(term frequency).

4) In lines from 66 to 70, first txt file is written which is posting.

4) Lines 71 through 87, we compined all “term, df, doc-id, tf” pairs. We wrote down the document id and term frequency of all the wrods in one line and added it to the terms4.

5)Lines from 90 to 103, second txt file is written which is dictionary one.