INFORMATION RETRIEVAL

Assignment 2: Gimme Results!

Instructions:

- Make a report in PDF format, clearly mentioning question/ part number against your answers.
- Plagiarism of any sort is not acceptable and will be severely punished.
- Late submission will not be accepted.
- **Submit your report, my_output.txt, performance.txt and code as one compressed file (.zip, .rar) on Google Classroom. The name of file should be your roll numbers i.e. <Roll#1_Roll#2>.zip**
- Deadline to submit this assignment is: **Sunday, 22nd March, 2020, 11:59 pm.**

Question:

In this assignment, you will use **the positional index you created in Assignment 1** to extract relevant documents. PFA two text files with this assignment document.

query.txt contains:

<query_id> <query>

output.txt contains:

<query_id> <document name>

You've to implement query processing to extract relevant documents from the corpus for the queries given in query.txt. You don't need to rank the documents, just extract all the relevant documents. Create a text file my_output.txt to store results of your query processing. The format of my_output.txt should be same as output.txt. You can match your results with the output.txt file. For every query find accuracy, precision and recall. You'll need to count true positives, true negatives, false positives and false negatives in the result of each query. Save these results in a text file performance.txt. The format of performance.txt should be following:

<query_id> <accuracy> <precision> <recall>

After calculating accuracy, precision and recall for each query separately, calculate average of these three metrics for your search engine.

You should process the text of the query in exactly the same way that you processed the text of a document in assignment 1. That is:

1. Split the query into tokens.
2. Apply case folding.

3. Apply stop-wording to the query using the same list you used in assignment 1.
4. Apply the same stemming algorithm to the query which you used in your indexer.

Report:

Prepare a report. List the steps that you followed for query processing to extract relevant documents. Report your results (accuracy, precision, recall for each query and the average values of these metrics).

Submission:

You have to submit:

- code for query processing
- code for accuracy, precision, recall
- my_output.txt file
- performance.txt file
- Report