Programming Assignment 1- Information Retrieval (CS F469)
Deadline: Sep 29, 2020 11:59 PM, Max Marks: 10

To begin with the assignment, students can download the attached dataset. It consists of *20 folders* enumerated as your group IDs and a *document.txt* file. Each folder contains two files *query.txt* and *relevance_assessment.txt*. *query.txt* contains a list of 10 queries. *relevance_assessment.txt* contains the list of document IDs relevant to a query. Each group must use the folder with the same name as their group ID. Parse the file *documents.txt* and extract the content for the document IDs present in their respective *relevance_assessment.txt* file; only these document IDs shall be used for creating the index.

Task 1 **[2 Points]** Inverted Index Construction

(a) Construct a full-text inverted index I_{full} and display the size of vocabulary.

(b) Plot the dictionary terms in the decreasing order of their frequency in I_{full}. Identify the stopwords in the corpora (if any) based on the size of the postings (not the standard lexicon of stopwords in nltk/spacy/online sources).

(c) Compute Precision and Recall for all 10 queries using I_{full}. X axis shows the query ID and Y axis shows the performance score. Legends show the Precision and Recall scores.

NOTE: All queries are written in uppercase letters, you need to convert them to lowercase before processing. White spaces are treated as logical OR operation.

Task 2 **[2 Points]** Select one or more linguistic models (text operations) and re-construct your inverted index; I_P: to increase the precision and I_R: to increase the recall.

(a) Report the changes in vocabulary size of I_P and I_R.

(b) Run the same set of queries used in **Task 1** on the new revised inverted indices (I_P and I_R) and report the precision and recall for each query. Display the results in form of a grouped bar plot for each query:

i. Precision results of I_{full} and I_P

ii. Recall results of I_{full} and I_R

iii. Precision and Recall results of I_P and I_R

NOTE: If you are using more than one linguistic model/pre-processing steps then you must show results for individual and pipeline of the steps.

Task 3 **[2 Marks]** Give inferences and justification for the followings:

(a) the models selected in **Task 2**. If more than one linguistic models are used then why? and why only this pipeline should be used?

(b) the changes in the results of i, ii, and iii in **Task 2**b

Be creative and do not write definitions in the justification.

Task 4 **[4 Points]** Generate a bi-gram index on I_{full}, I_P, and I_R. Convert at least three words in each query to the following wildcard patterns: *X, X*, and X*Y. Example: `This is a sample sentence` can be converted to `*is is a sa*ple sente*`. You are allowed to do this manually for each query. Now, compute the precision and recall. Justify the k-gram index results against I_{full}, I_P, and I_R.

Assignment submission instructions:

1. Only one member from the group shall submit the assignment. Each group has to submit only one file i.e. Python notebook.
2. The notebook itself should contain source code, results, plots, and justifications.
   (a) First cell of the notebook file should have names and roll numbers of the students in the group.
   (b) Task number should be clearly mentioned in the solution.
   (c) any libraries that might be needed !pip install
   (d) Structure your source code and cells well- avoid repetitive cells or the ones with small snippets.
3. File name should be "PA1_groupID". **Submissions made over email will not be evaluated.**