CSE508 : Information Retrieval Assignment 4

Max Marks: 80

Instructions

- The assignment is to be attempted individually
- Language allowed: Python
- For Plagiarism, institute policy will be followed
- You need to submit README.pdf, Code files (it should include both .py files and .ipynb files), and Analysis.pdf
- You are allowed to use libraries such as NLTK for data preprocessing.
- Mention methodology, preprocessing steps and assumptions you may have in README.pdf.
- Mention your outputs, analysis done (if any) in Analysis.pdf
- Submit code, readme and analysis files in ZIP format with the following name: **A4_<roll_no>.zip**
- Save all your precomputed indexes and tables which may take time to compute.
- ***Note: Due to the Covid-19 outbreak and lockdown, it may so happen that assignment demos cannot be taken, hence you are advised to prepare a well documented IPYNB file and report/analysis with all the justifications that may be necessary.***

Download 20newsgroup dataset. You need to pick documents of **comp.graphics, sci.med, talk.politics.misc, rec.sport.hockey, sci.space.**

1. Implement a cosine similarity measure with tf-idf weighting. Your index should contain the information that you will need to calculate the cosine similarity measure such as tf and idf values.
2. Implement the Rocchio Algorithm (with query refinement).
   a. You have to display top k (k should be at least 100) docs for the initial query.
   b. To provide feedback, you have to mark p% of k docs to be relevant. The p% selected documents would be from the folder which is assumed to be the Actual relevant set (Ground truth).
   For e.g., if Ground Truth relevant are docs of the folder **sci.med** and k = 100 then,
   for p = 10%, you have to mark 10% of 100 i.e. **10 docs of sci.med** as relevant from the top of the retrieved list of 100 docs.

   c. Show the revised top k results after performing relevance feedback. Mark the documents as * which were judged as relevant during the relevant feedback phase.
   d. Use α= 1, β= 0.75, and γ=0.25 as parameters for the Rocchio's algorithm. For each iteration of the relevance feedback, you have to show top k docs and provide feedback in the same way as stated above.
   e. Consider the following queries & all documents inside the given folder as a relevant set (ground truth).

Query 1: Pretty good opinions on biochemistry machines
Relevant set 1: Documents inside folder **sci.med**
Query 2: Scientific tools for preserving rights and body
Relevant set 2: Documents inside folder **talk.politics.misc**
Query 3: Frequently asked questions on State-of-the-art visualisation tools
Relevant set 3: Documents inside folder **sci.med**

Report the following:

1. PR curve plot for each of the queries. You have to plot the PR curve after each relevance feedback iteration. (Do around 3 to 4 feedback iterations)
2. MAP for the above-mentioned query set after each relevance feedback iteration.

Note: This set of queries are just for reference, keep your code generalized to be able to accept different queries and mark whatever documents to be relevant as per our wish.

f. In the report, show how the query vector changes (for this particular set of queries) after applying each iteration of the Rocchio Algorithm. Justify or explain the results after each iteration. Show a 2D TSNE plot of the vectors to demonstrate the difference. You can use Sklearn's inbuilt functions to make this plot.