CSE508 : Information Retrieval

Assignment 2

Max Marks: 70

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 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: **A2_<roll_no>.zip**

Question 1: Download http://archives.textfiles.com/stories.zip dataset **(60 Marks)**

You need to implement a CLI tool for:

1) **Jaccard Coefficient based document retrieval:** For each query, your system will output top k documents based on jaccard score.

2) **Tf-Idf based document retrieval:** For each query, your system will output top k documents based on tf-idf-matching-score. Implement different versions of Tf-Idf based document retrieval then compare and analyze which performs better and why.

3) **Tf-Idf based vector space document retrieval:** For each query, your system will output top k documents based on a cosine similarity between query and document vector.

In addition, ensure that numerical queries work. Example “100 animals”, “50,000 variety of flowers”, “population of 1 billion” etc.

Give special attention to the terms in the document title and analyze the change in result with and without attention to terms in title.

Compare and state pros and cons for all the techniques.

Question 2: Download the dictionary from http://www.gwicks.net/dictionaries.htm (UK ENGLISH - 65,000 words) **(10 Marks)**

Take a sentence as input from user. For each non dictionary words present in the sentence suggest top k words on the basis of minimum edit distance. Cost of operations is defined as:
Insert: 2
Delete: 1
Replace: 3