CSE508 : Information Retrieval
Assignment 2

Deadline : 10th Feb'18, 2359 hrs

Total: 100 marks + 10 marks bonus

Instructions

- Assignment is to be attempted individually. Please keep the discussions on an abstract level
- Language allowed : Python
- For Plagiarism, institute policy will be followed
- You need to submit ReadMe, code files and analysis.pdf
- You folder should be renamed in the NameRollNo_HW2 format before zipping

Document Retrieval

Download http://archives.textfiles.com/stories.zip dataset.

You need to implement a CLI tool for

1) Tf-Idf based document retrieval
    - For each query, your system will output top 5 documents based on tfxidf-matching-score.
2) Tf-Idf based vector space document retrieval
    - For each query, your system will output top 5 documents based on cosine similarity between query and document vector

Data preprocessing will be done as similar to the previous assignment. In addition to that, you need to implement [you may use libraries for this]

- Spelling correction in query
- Consider numerical queries. Example "100 animals", "50,000 variety of flowers", "population of 1 billion" etc
- Give special attention to the terms in document title

Bonus: Implement query caching [upto last 20 queries]