CSE508 : Information Retrieval
Assignment 2

Instructions

- The assignment is to be attempted individually
- Language allowed: Python
- For Plagiarism, institute policy will be followed
- You need to submit ReadMe, code files and analysis.pdf

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 k documents based on tf-idf-matching-score.
2) 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 your analysis file, mention the methodology adopted. Also, mention a case where 1,2 give different results. Explain why the difference occurs.

In addition, ensure that numerical queries. Example “100 animals”, “50,000 variety of flowers”, “population of 1 billion” etc - Give special attention to the terms in the document title.