Information Retrieval – Spring 2023
Assignment # 2

Corpus: For this assignment use Corpus 2.

Program: *score*

Write a program for ranking top 10 documents against a given query. *You have been provided with 10 documents and 4 queries in "Corpus 2".*

- **The program should show two options for ranking i.e.**
    i. **tf-idf**
    ii. **cosine**
    - The program should
        - tokenize
        - exclude all stop words (Stopwords.txt contains list of stopwords)
        - achieve stems (you can use some existing program for stemming)
        - construct tf-idf matrix for all terms and documents.
        - convert each document in unit vector (dividing by the magnitude of each document)

Program: *Test*

Write a program **test**

- Test program should read the 4 query files one by one)
- As per given query the program should rank all the documents and return them against query (display the links of the documents as provided by google search)
- For ranking use tf-idf score or cosine similarity measure which should be chosen at the start of the execution of your program.
- Compare and analyze the results of both the techniques.

Submission

You should submit a zip archive containing:

- Programs **score** and **test** (source and executable)

As part of the grading, we will run your programs against the test data and some other query.