CSC790: Information Retrieval
Spring 2024, Programming Assignment #3
Date Assigned: Friday, March 08, 2024
Due Date: Monday, March 25, 2024 at 11:59 pm
(on Brightspace) (150 + 30 points)

Objectives:

- Implementing and using vector space model.
- Implementing similarity between documents.

Tasks
Write a python code that uses the vector space model to compare the documents provided in previous homework. Your code should ask the user to type a number K and retrieve the top k closed documents to each other. Your code should use the following measures as vector elements:

1. Only term frequency.
2. The tf-idf measure ($tf\text{-}idf_{t;d} = tf_{t;d} \times idf_t$).
3. The sublinear tf scaling:

$$wf_{t,d} = \begin{cases} 1 + log_{10}(tf_{t,d}) & \text{if} \quad tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$$

The sublinear tf scaling is defined as:

$$wf\text{-}idf_{t,d} = wf_{t,d} \times idf_t$$

Your code should display the results as follows(this is just an example, it is not necessarily correct).

```
========== CSC790–IR Homework 01 ==========
First Name: your first name
Last Name : your last name
============================================================
The number of unique words is: 2454
The top 10 most frequent words are :
    1. university.
    2. search.
    3. ....
```

```
    .
    .
    .
==========================================================
The top k closest documents are:
    1. Using tf
        file2 , file25 with similarity of 0.90
        file17 , files 34 with similarity of 0.87
        .
        .
        .
    2. Using tf_idf:
        file12 , file245 with similarity of 0.99
        file157 , files734 with similarity of 0.95
            .
            .
            .
    3. Using wf_idf:
        file202 , file745 with similarity of 0.78
        file157 , files734 with similarity of 0.72
            .
            .
            .
```

Submission

1. Write your own code. Use as many functions as you can.
2. Make sure you writing you name and assignment number on all files you submit.
3. Your python code and the instructions on how to run it.
4. Enclose all your files in a folder named **HW03_yourlastname.zip**.
5. Submit the zip file using blackboard.