CSC790: Information Retrieval and Web Search
Spring 2024, Assignment #1
Date Assigned: Monday, January 29, 2024
Due Date: Monday, February 05, 2024 at 11:59 pm (on Brightspace) (130 points)

Objectives:

- Text processing.
- Build the inverted index from a list of documents.

Tasks
Download the dataset documents.zip from blackboard. Write the python code that:

1. Read and process text dataset.
    - Read text.
    - Tokenize.
    - Stem.
    - Remove stop words (you must use the provided list)
2. Build the inverted index. Write your own code
3. Displays the size of the index (in byte and MB).
4. Display the top n frequent terms.
5. Save the index in a file/load the saved index file.

Important

- Your code should have at least four function. One function should display course and student information as follows:
  ========== CSC790-IR Homework 01 ==========
  First Name: your first name
  Last Name : your last name
  ==========
- Document your code: write comments to explain the role each function and block of code; what are the input parameters and the output/return parameters.
- You must use NLTK.

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 **HW01_yourlastname.zip**.
5. Submit the zip file using Brightspace.