CSC790: Information Retrieval and Web Search
Spring 2025, Homework Assignment 01

Date Assigned	Monday, January 27, 2025
Due Date	Wednesday, Febrary 05, 2025 at 11:59 pm
Submission	Brightspace
Total points	120

Objectives:

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

Tasks (100 points)
For this assignment, you need to create an environment with python 3.13.1 and nltk 3.9.1
Download the dataset documents.zip from Brightspace. Write the python code that:

1. Read and process text dataset.
    - Read text.
    - Tokenize.
    - Remove stop words (you must use the provided list)
    - Normalize (lower case, remove punctuation, stem, ...)
2. Write your own code to build the inverted index.
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 (20 points)

- 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.
- Code must be well organized and easy to use.

- 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.