INFO-533 Project 1 (100 points)

Deadline: 29th Feb, 2024 11:59 pm in Brightspace

For the project of this course, you are going to develop a basic **Information Retrieval** system. Such a system involves storing the documents by performing indexing and retrieving those documents based on a given query. You will be developing this whole system in 3 parts.

Project 1 is a team project with 2 members.

For project 1, you are required to build the positional index for provided documents. These documents are one line sentences provided in ***data.json***. The file contains an array of 220 strings which you will treat as separate documents for this project. Use this file as an input to your project.
Following are the two key points you should perform to complete this project.

1. Create terms out of the documents.
    - Clean the data by removing all the special characters and punctuations. You may perform other data cleaning steps that you deem necessary.
    - Perform case-folding (convert every term to lowercase)
    - Stemming (you may use porter's stemming)

2. Process these documents further to generate the positional indexes for all the terms (from step 1) and store them into a json file named ***postings.json*** in the following manner:
    ```
    {
       "term1":{
          "docId1":[pos1,pos2,pos3],
          "docId2":[pos1,pos2,pos3, pos4]
       },
       "term2":{
          "docId1":[pos1, pos2]
       },
       ….
    }
    ```

Bonus: The above format is using a python dictionary, which will grow as term increases. To make it more memory efficient you can store using linked list data structure. (25%)

Note: Use PYTHON as the programming language for this project. Your code should be properly commented. Your code will be tested against another input file with different data content.

All your code will be subject to check for similarity with other submissions. So you are advised not to look at other team's code.

Submission:

1. A python file (.py) or Jupyter notebook (.ipynb).

2. A PDF mentioning instructions on how to give file input and run your code. Add the Honesty Policy statement at the beginning of the PDF, signed by both team members. You can find the statement in the syllabus overview.
3. The output file, named ***postings.json*** with the correct structure as mentioned above.