INFO-533 Project 2 (100 points)

Deadline: March 31, 2024 11:59 pm in Brightspace

Please remember to include the following statement at the beginning of your submitted assignment and SIGN it. Your assignment won't be graded without the signed statement.

"*I have done this assignment completely on my own. I have not copied it, nor have I given my solution to anyone else. I understand that if I am involved in plagiarism or cheating, I will have to sign an official form that I have cheated and that this form will be stored in my official university record. I also understand that I will receive a grade of 0 for the involved assignment and my grade will be reduced by one level (e.g., from A to A- or from B+ to B) for my first offense, and that I will receive a grade of "F" for the course for any additional offense of any kind.*"

You are responsible for making your Project 1 group yourself. There can be at most two members in a group. In case you are unable to find a second member for your group, your project 2 will be considered as an individual.

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 2 is a team project with 2 members.

For project 2, you are required to retrieve the document IDs with starting positional index for given queries. You are provided with a ***postings.json*** file similar to the one you submitted in Project 1. Use the provided postings.json to find the document Ids for the following queries:

1. "One of the"
2. "The best way"
3. "Someone who knows"

Following are the key points you should perform to complete this project.

1. Preprocess the queries - Use the below function to preprocess queries since the same is used to generate the given postings.json.

```
from nltk.stem import PorterStemmer
import re
stemmer = PorterStemmer()
def preprocess(document):
    document = re.sub(r'[^a-zA-Z\s]', '', document).lower()
    terms = document.split()
    terms = [stemmer.stem(term) for term in terms]
    return terms
```

2. Find the document IDs with starting position indexes for every match for each query - Use the given positional indexing in ***positings.json*** to find the document IDs that contains the complete match of the queries.
   Example:
   Document1: "this is just a test document"
   Document2: "this is another test document"

Query: “a test document”
Output: docId 1, position 3 (“a” is at position 3, indexing from 0)

Note: Use PYTHON as the programming language for this project. Your code should be properly commented. Your code will also be tested for other queries.

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

Submission (no zipped files):

1. A python file (.py) or Jupyter notebook (.ipynb) containing output as retrieved documentIds for each query with starting positional index of the query in that document.
2. A PDF with the Honesty Policy statement at the beginning, signed by both team members. You can find the statement in the syllabus overview.