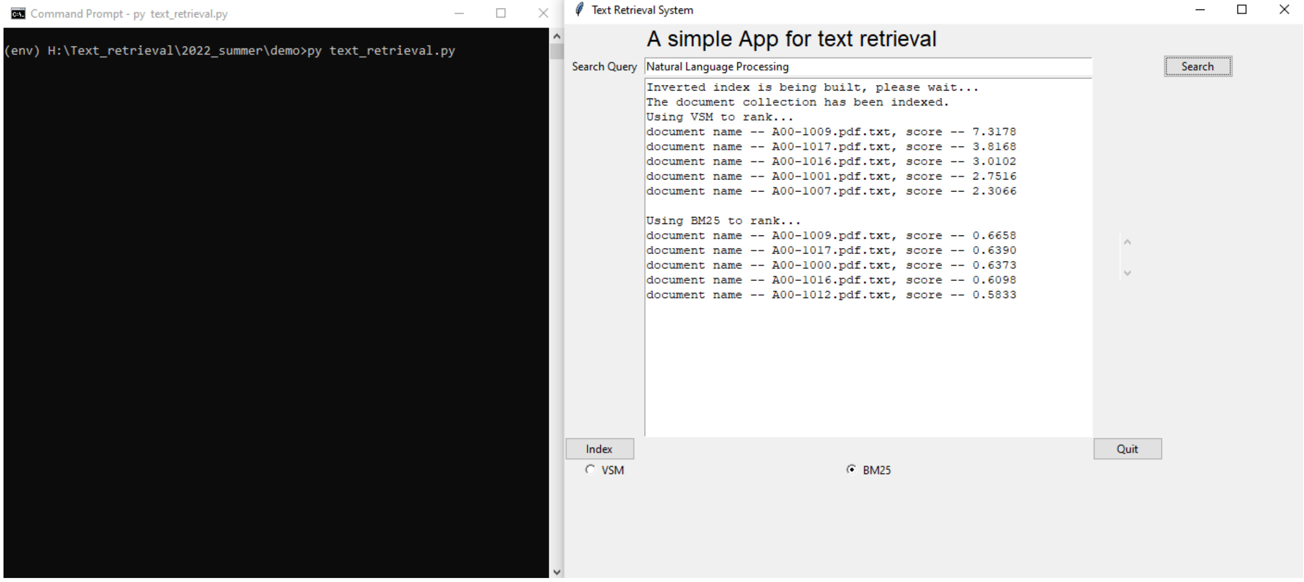
Wilfrid Laurier University, Dept. Physics and Computer Science

CP423 B Text Retrieval & Search Engine
Coding Assignment

Due on July 27, 2022, at 7:00 pm

You shall use Python to finish this coding assignment. You need to implement a simple App for text retrieval using the vector space retrieval model and the BM25 retrieval model. When you are successful to finish all coding tasks, you can use your computer console to launch the App. You can use PyCharm, the IDE designed for Python develop to create your Python project. For the tutorial to install Python and PyCharm to your computer (Windows machine), you can refer to my Python tutorial slides. After the Python project is created, you should copy all the .py files, the requirements.txt file, and the "documents" folder to your project folder.

1 When you finished the program assignment, you can launch the Python App like below:



Command Prompt - py text_retrieval.py

```
(env) H:\Text_retrieval\2022_summer\demo>py text_retrieval.py
```

Text Retrieval System
A simple App for text retrieval
Search Query
Search
Index
Quit

Search Query: Natural Language Processing

```
Inverted index is being built, please wait...
The document collection has been indexed.
Using VSM to rank...
document name -- A00-1009.pdf.txt, score -- 7.3178
document name -- A00-1017.pdf.txt, score -- 3.8168
document name -- A00-1016.pdf.txt, score -- 3.0102
document name -- A00-1001.pdf.txt, score -- 2.7516
document name -- A00-1007.pdf.txt, score -- 2.3066

Using BM25 to rank...
document name -- A00-1009.pdf.txt, score -- 0.6658
document name -- A00-1017.pdf.txt, score -- 0.6390
document name -- A00-1000.pdf.txt, score -- 0.6373
document name -- A00-1016.pdf.txt, score -- 0.6098
document name -- A00-1012.pdf.txt, score -- 0.5833
```

○ VSM ● BM25

The App can be operated by the following procedure:

[1] In the GUI, first click "index", then in the "Select Folder" window, select the "documents" folder, the GUI will display "Inverted index is being built, please wait...". After a few seconds, after the inverted index tables are created in the RAM, the GUI will display "The document collection has been indexed."

[2] In the Search Query box, enter your query, such as "Natural Language Processing", or "Machine learning". Click the Radio button to select either VSM or BM25 as the ranking function (default is VSM), then click "Search". The top 5 document names with the ranking scores will display on the text box in the center of the GUI.

[3] You can switch between the two ranking functions to see the difference of the ranking results.

[4] Click "Quit" to exit the App.

2. Instructions

There are Four .py files in the folder:

- text_retrieval.py: this is the python file to launch the App.
- get_index.py: this is the python file to index the text documents and containing all utility functions
- search.py: this is the python file to rank the text documents
- config.py: this is the python file to download the NLTK popular packages, run this script before your coding.

The main Python script is the first one, i.e., text_retrieval.py. Please finish the code in get_index.py and in search.py before starting to work on text_retrieval.py.

The project is ready to run if you use Pycharm and properly create the project, create the virtual environment, but the buttons are not responsible when being clicked.

Please finished all the functions and implement the functionality of the GUI components. You can refer to you python exercise notebooks of the lectures, particularly the “Vector_Space_Model(VSM).IPYNB” in Week 5 and the “BM25_probabilitic_model.IPYNB” in Week 6. If you study the code in these two exercise, this assignment will be very easy.

When the project is finished, you can run the App after activating the virtual environment in console by typing “py text_retrieval.py” or “python text_retrieval.py”.

You can also build an executable file to launch the App by pyinstaller. After you finish all coding tasks and test it. After activating the virtual environment, run the following command:

```
pyinstaller text_retrieval.py --onefile –windowed
```

The process will take about 2 minutes. After the App is built, you can find an executable file in the newly created “dist” folder. In Windows machine, the name of the executable file is text_retrieval.exe; in a Mac, the name of the executable file is text_retrieval.app. Since everything is bundled into the executable, you can run the App by this file only.

DO NOT submit the executable file as part of the assignment, it is too large.

This is the end of the instruction. Happy coding!