Operating System: Windows 10

-------------------------------------------

Check whether python is installed by using the following command in command prompt:-

python

If not installed, link for installing Python:-

https://www.python.org/download/releases/2.7.1

install prettytable using the following command in cmd:

pip install prettytable

install pickle using the following command in cmd:

pip install pickle

Install Eclipse, if not already present.

Task1

Java

IDE: Eclipse

For Windows machine:-

After the packages have been installed, run the program from this folder:

**Task1 ( java)**

Install Eclipse

The code for Task1 is present in folder IR assignment 4 in this folder.

The output file for Task1 is present in the folder Tables as scoreFile\_Lucene.txt

The output for all the 4 queries for this task is present in one file only, scoreFile\_Lucene.txt

Open eclipse and import the package “IR assignment 4” from

C:/Users/Tanvi/Desktop/HW4\_Tanvi\_Ranadive/Task1

The corpus for this task is the raw documents without any html tags, references, tables and images.

After importing the package, open the file HW3.java in src folder

Change the string “index” in main() function from:

C:/Users/Tanvi/Desktop/HW4\_Tanvi\_Ranadive/Task1/Lucene Index

to your path where you want the index file to be created

Find indexer.indexFileOrDirectory("C:/Users/Tanvi/Desktop/HW4\_Tanvi\_Ranadive/Corpus ");

And change the path to your respective path where the corpus is present

Find File file = new File("C:/Users/Tanvi/Desktop/HW4\_Tanvi\_Ranadive/Task1/scoreFile\_Lucene.txt");

And change the path to your path where you want the score table to be created

Find BufferedReader brd = new BufferedReader(new FileReader("C:/Users/Tanvi/Desktop/HW4\_Tanvi\_Ranadive/Task1/queries.txt"));

And change the path to your path where the text file containing queries are present.

Now,

Run HW3.java

The output file will be created in the path specified by you earlier

Open the output file using Notepad++ for proper formatting.

If you want to run the program again, be sure to delete the files from the “Lucene Index” folder, i.e the folder where the index was created.

Also delete the scoreFile\_Lucene.txt from the path specified by you earlier.

Then run the program again.

**Task 2 (Python):**

Operating System: Windows 10

Python:2.7.10

IDE: Sublime Text 3

Packages: PrettyTable, pickle

Task2

The ouput files for Task2 are present in Tables folder in this folder as Table\_Q2\_Query1, Table\_Q2\_Query2, Table\_Q2\_Query3, Table\_Q2\_Query4 as four separate files.

The inputs for this task are the following files already present in Task2 folder in this folder:

doc\_tokens\_dump.txt 🡪 file containing the docId s and total number of tokens in each doc

doc\_terms\_file.txt 🡪 file containing the docId and all the terms in each doc with their frequency

Dociddict.txt 🡪 file containing the document names and corresponding docId

inverted\_index\_dump.txt 🡪 file containing the inverted index

queries.txt 🡪 file containing the queries in the format:

queryId Query (queryId space thequery)

eg. 1 global warming potential

Running the program

To run the program :

Go to the function : create\_index\_table\_for\_query

Change the path to your path where the index should be created.

Change “C:/Users/Tanvi/Desktop/HW4\_Tanvi\_Ranadive/Tables/Table\_Q2\_Query” to your path.

Then run the program

Task2.py

The output files will be created in the “Tables” folder:

C:\Users\Tanvi\Desktop\HW4\_Tanvi\_Ranadive\Tables

As Table\_Q2\_Query1, Table\_Q2\_Query2, Table\_Q2\_Query3 and Table\_Q2\_Query4

**Bonus Question BM25:**

The inputs for this task are the following files already present in the same folder:

doc\_tokens\_dump.txt 🡪 file containing the docId s and total number of tokens in each doc

doc\_terms\_file.txt 🡪 file containing the docId and all the terms in each doc with their frequency

Dociddict.txt 🡪 file containing the document names and corresponding docId

inverted\_index\_dump.txt 🡪 file containing the inverted index

queries.txt 🡪 file containing the queries in the format:

Query

eg. global warming potential

Running the program

To run the program :

Open the program BM25\_FINAL.py

Open the program in sublime text 3

Go to the class BM25 at the beginning of the program

Change the path of outfile to your path where the output tables should be created.

Change “C:/Users/Tanvi/Desktop/HW4\_Tanvi\_Ranadive/bm25/BM25\_Query” to your path.

Then run the program

BM25\_FINAL.py

The output files will be created in the your specified folder:

As BM25\_Query1, BM25\_Query2, BM25\_Query3 and BM25\_Query4