Information Retrieval course mini project

Topic : Historical information related search based on Wikipedia

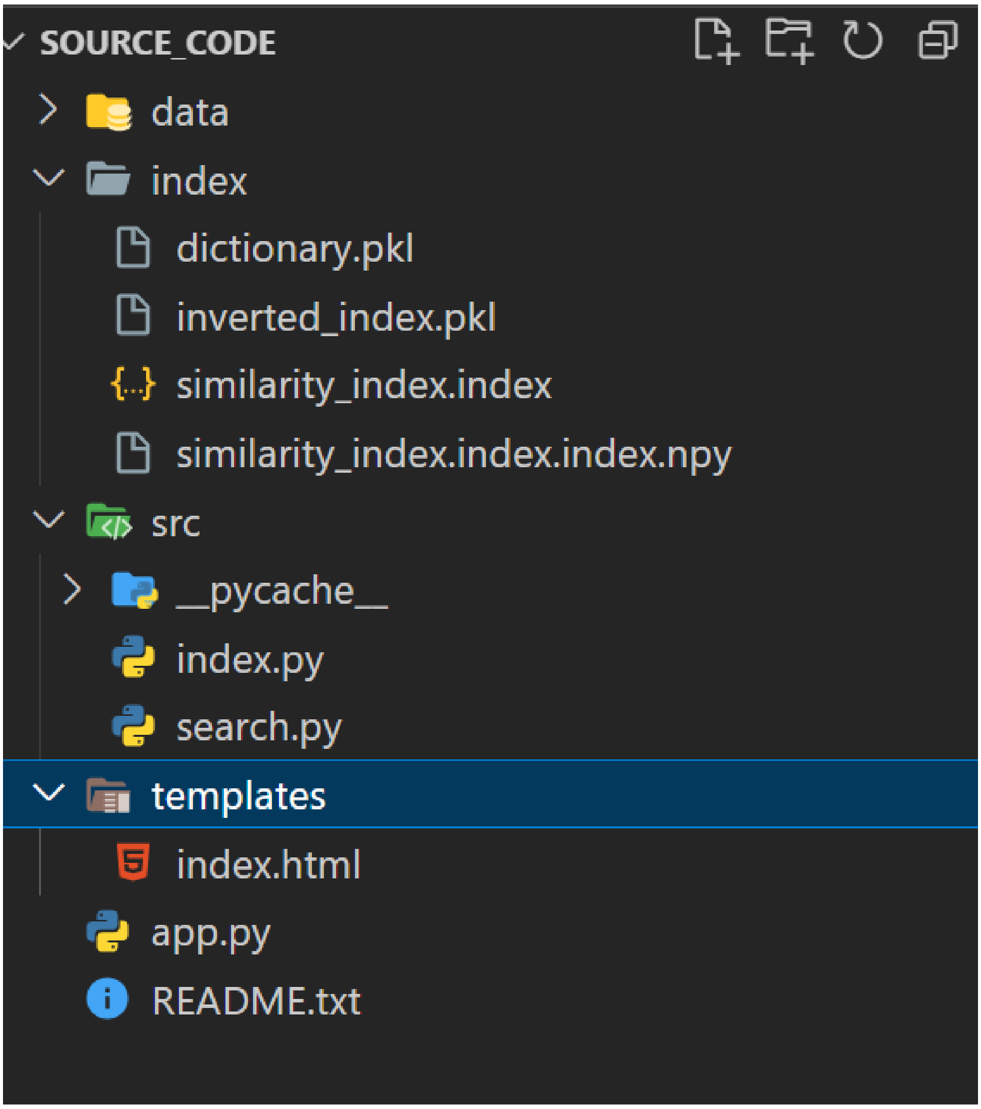
Topic id: 58

Name : DEVADA KUMARA SWAMY

Roll no: S20210010060

PROJECT OVERVIEW (TASKS) :

1. **Project Directory structure**



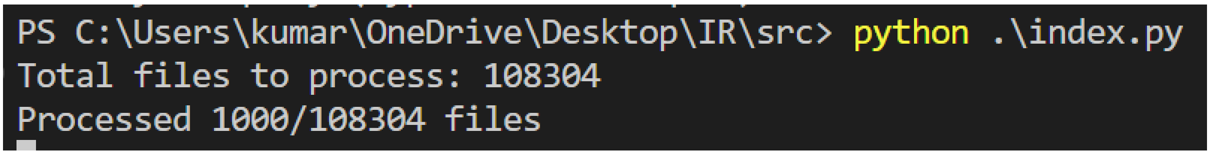
```
SOURCE_CODE
  > data
  v index
      dictionary.pkl
      inverted_index.pkl
      similarity_index.index
      similarity_index.index.index.npy
  v src
    > __pycache__
      index.py
      search.py
  v templates
      index.html
    app.py
    README.txt
```

2. Downloading the data set

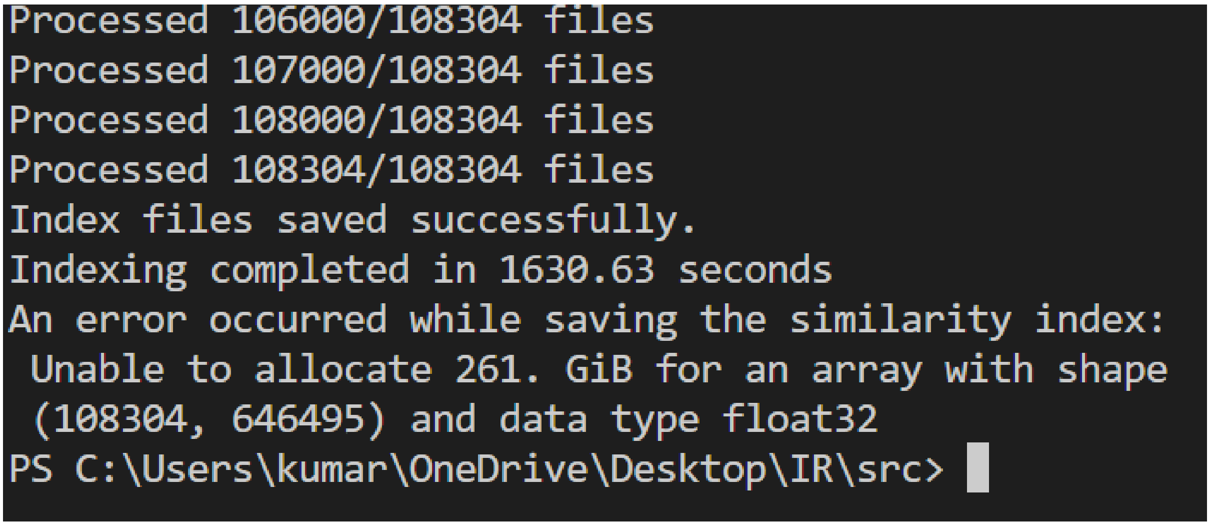
I downloaded around 1,00,000 documents with their names as their id numbers. These txt documents are stored in the folder namely data.

3. Creating inverted index.

I used gensim library in python to make TF-IDF based inverted index.
After executing the file index.py the inverted_index file is created in folder namely index and further used to process and search the documents for a given query.



```
PS C:\Users\kumar\OneDrive\Desktop\IR\src> python .\index.py
Total files to process: 108304
Processed 1000/108304 files
```

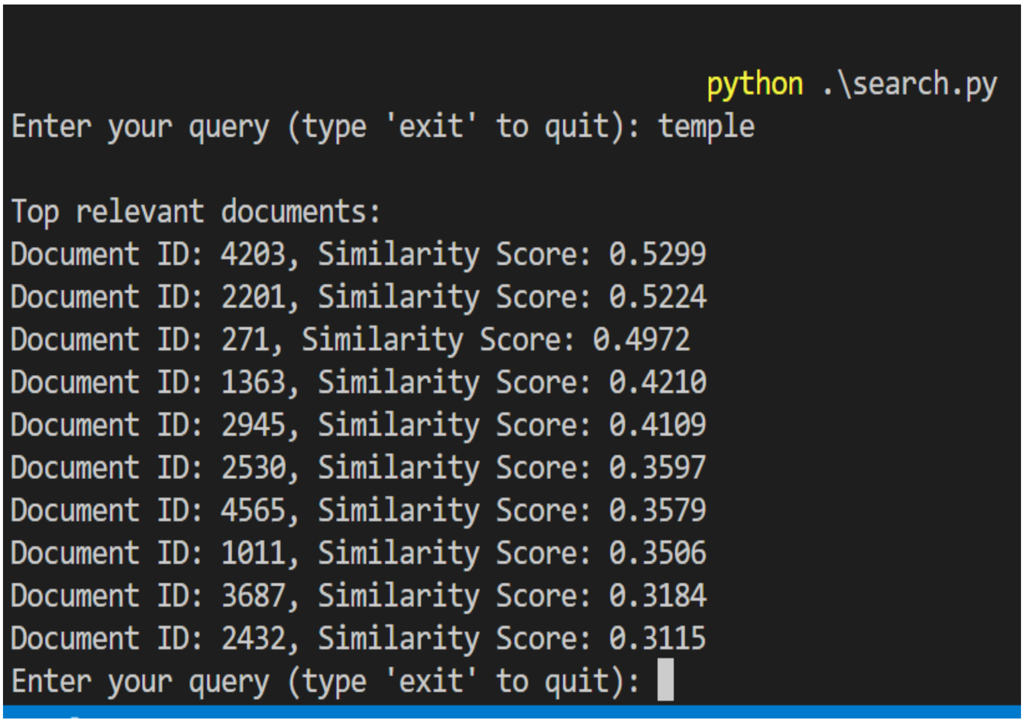


```
Processed 106000/108304 files
Processed 107000/108304 files
Processed 108000/108304 files
Processed 108304/108304 files
Index files saved successfully.
Indexing completed in 1630.63 seconds
An error occurred while saving the similarity index:
 Unable to allocate 261. GiB for an array with shape
 (108304, 646495) and data type float32
PS C:\Users\kumar\OneDrive\Desktop\IR\src>
```

4. Query processing and searching.

Two methods are implemented for viewing output/results.

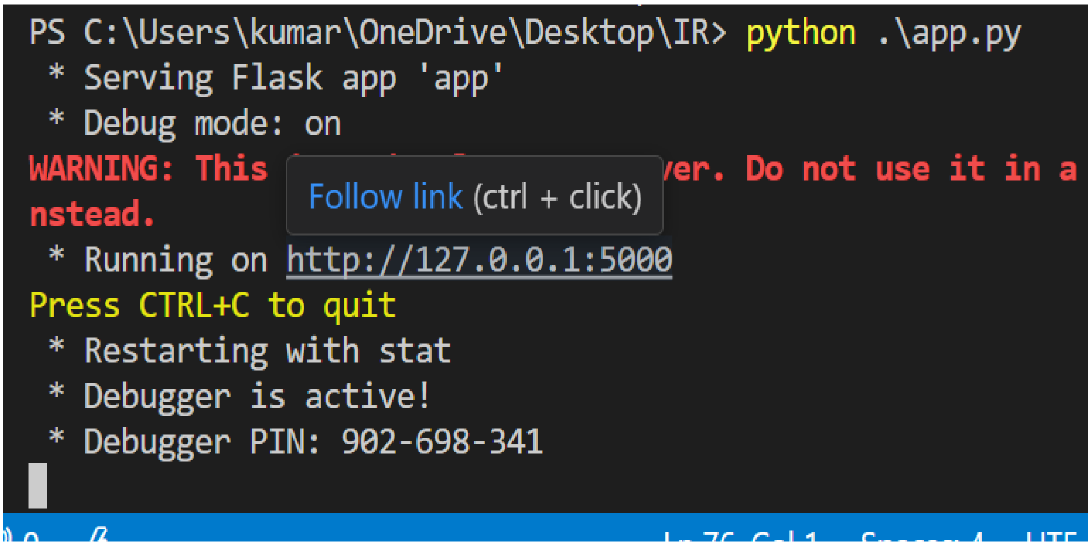
1. Command line interface



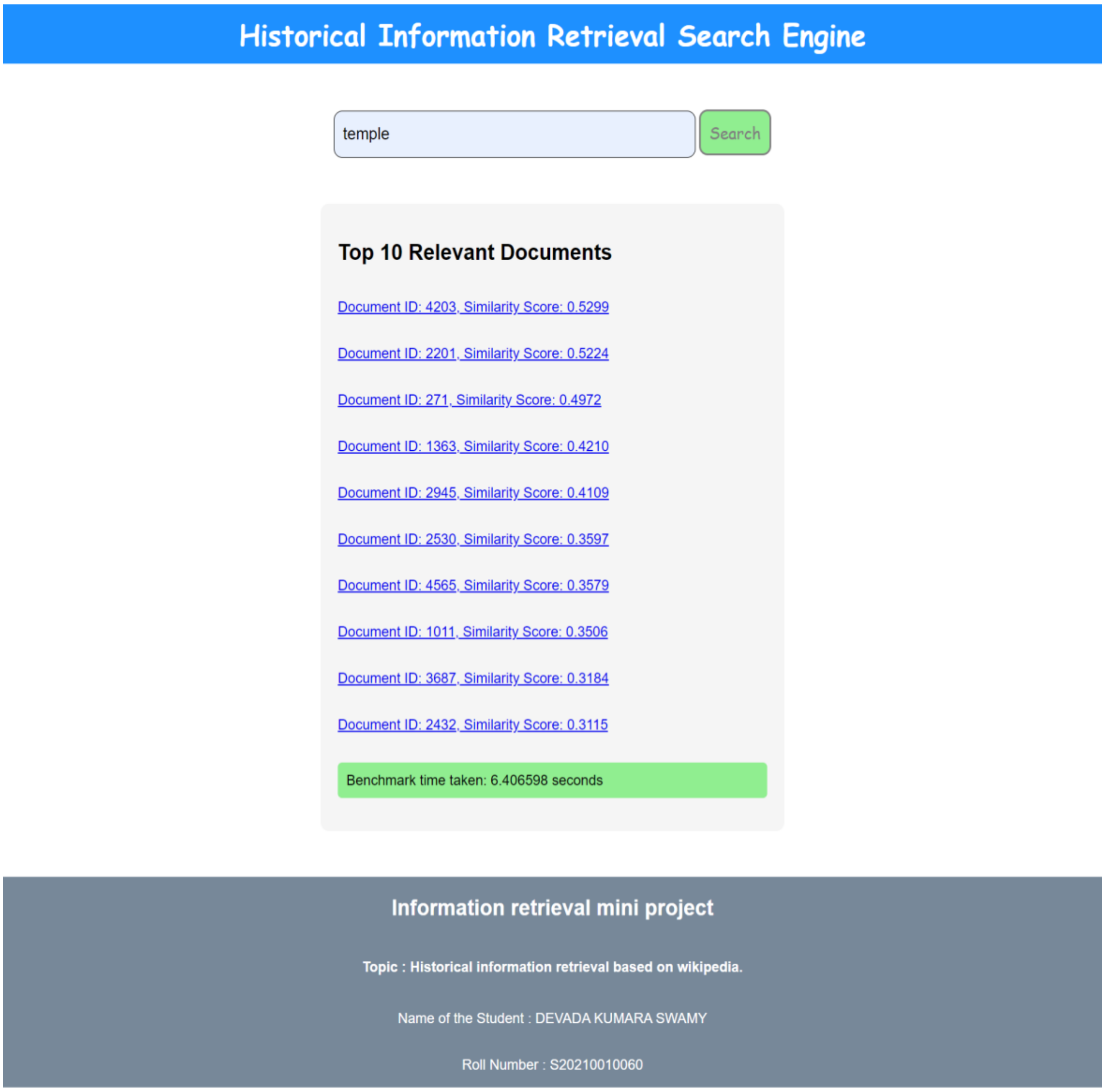
```
                                          python .\search.py
Enter your query (type 'exit' to quit): temple

Top relevant documents:
Document ID: 4203, Similarity Score: 0.5299
Document ID: 2201, Similarity Score: 0.5224
Document ID: 271, Similarity Score: 0.4972
Document ID: 1363, Similarity Score: 0.4210
Document ID: 2945, Similarity Score: 0.4109
Document ID: 2530, Similarity Score: 0.3597
Document ID: 4565, Similarity Score: 0.3579
Document ID: 1011, Similarity Score: 0.3506
Document ID: 3687, Similarity Score: 0.3184
Document ID: 2432, Similarity Score: 0.3115
Enter your query (type 'exit' to quit):
```

2. Simple HTML webpage UI
   I used Flask library in pyhton to make a server which will serve the response based on the request that was made to it using HTML page.



```
PS C:\Users\kumar\OneDrive\Desktop\IR> python .\app.py
 * Serving Flask app 'app'
 * Debug mode: on
WARNING: This                   ver. Do not use it in a
nstead.
 * Running on http://127.0.0.1:5000
Press CTRL+C to quit
 * Restarting with stat
 * Debugger is active!
 * Debugger PIN: 902-698-341
```



Historical Information Retrieval Search Engine

Top 10 Relevant Documents

Document ID: 4203, Similarity Score: 0.5299

Document ID: 2201, Similarity Score: 0.5224

Document ID: 271, Similarity Score: 0.4972

Document ID: 1363, Similarity Score: 0.4210

Document ID: 2945, Similarity Score: 0.4109

Document ID: 2530, Similarity Score: 0.3597

Document ID: 4565, Similarity Score: 0.3579

Document ID: 1011, Similarity Score: 0.3506

Document ID: 3687, Similarity Score: 0.3184

Document ID: 2432, Similarity Score: 0.3115

Benchmark time taken: 6.406598 seconds

Information retrieval mini project

Topic : Historical information retrieval based on wikipedia.

Name of the Student : DEVADA KUMARA SWAMY

Roll Number : S20210010060

-Thank you-