Design-Lab-Lex-Yacc-and-NoSQL Project

Project : Lex Yacc and NoSQL - crawling Covid Statistics and News

Repository Link : https://github.com/aakashg339/Design-Lab-Lex-Yacc-and-NoSQL-crawling-Covid-Statistics-and-News

Distribution of work :

- Aakash Gupta(23CS60R22) : Module 2 (2nd part) and Module 3.2
- Rajdeep Ghosh(23CS60R10) : Module 1 and Module 3.1 and Module 4
- Satya Prakash Nayak(23CS60R05) : Module 2 (1st part) and Module 3.2

How to run the code:

The entire project has a single entry point which is the file 'driver.py' to run the whole project.

```
cd Design-Lab-Lex-Yacc-and-NoSQL-crawling-Covid-Statistics-and-News
python3 driver.py
```

Project Description:

The project is divided into 4 modules.

- Module 1 : Crawling Covid-19 statistics from the website https://www.worldometers.info/coronavirus/ (for all countries and individual countries over a period of time)
- Module 2 : Crawling Covid-19 news wiki page https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic containing date wise text data of the Covid-19 news and also for multiple countries.
    - Module 2.1 and 2.2 : Extracting the worldwide news and responses for all times.
    - Module 2.3 : Extracting the news and responses for a particular country over a period of time.
- Module 3 :
    - Module 3.1 : Using NOSQL database to give the results of the statistics and news in a structured format for the queries related to MODULE 1
    - Module 3.2 : Using NOSQL database to give the results of the statistics and news in a structured format for the queries related to MODULE 2
- Module 4: Combining the entire project and creating a user interface for the user to interact with the database and get the required information.

Work Done:

Module 1

- All Countries table -> Extract the data from the mentioned fields for all the countries and store it in a NoSQL database (output.txt). There is a driver.py to run the code. The grammars are mentioned in the file 'extract.py' and the output is stored in 'output.txt'. LEXX and YACC are used to parse the data.

- Individual Country -> Extract the data for the 4 fields mentioned for a particular country mentioned in worldmeters_countrylist.txt . The 'driver.py' is used to run the code. The grammars are mentioned in the file 'extract_activecases.py' , 'extract_newcases.py' , 'extract_newdeaths.py' and 'extract_newrecoveries.py' . After that the outputs are merged into a single file present in "merged_files" as 'mfile_countryname.txt'. LEXX and YACC are used to parse the data.

NOTE All countries mentioned in worldmeters_countrylist.txt are do not have all of the 4 fields. Hence I modified the list to include only those countries which have all the 4 fields.

Module 2

Module 2.1 and 2.2

This part of the module crawls the Wikipedia link: https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic and then extracts worldwide news and responses for all times.

- To obtain all the required Wikipedia webpages, we first download the main wiki page using Mainpage_download.py.
- From the main wiki page, we extract all the timelines and response pages present on that page using links_extract.py.
- Finally, using extract_info.py, we extract all the response pages one by one, categorized by their respective month names, which are then organized inside their respective years. For this process, the extracted parts contain H2 headers like "Reactions and measures in Europe", H3 headers like "1 April", and all the list and paragraph contents under them.
- For this part, we have utilized modules such as ply.lex, ply.yacc, os, and re.

Module 2.3

Completed part 2.3 for India. Currently it works only for India and for few Australia page.

- HTML file of each country is of different format.
- Within each file also the structure can change.
- Therefore used two files 'IndiaParser1.py' and 'IndiaParser2.py' to parse the data and a seperate one for Australia.

First extracted the required webpages and then extracting the data.

Module 3.1

There are broadly two parts to this module.

- Part 1 : display the statistics of a particular country for a set of fields mentioned. The user is asked about the country name and set of fields . These are stored in files and feed into the mapper along with the raw data we obtained from module 1. The mapper passes the data to the reducer which inturn preprocess and filters and send it to the reducer for the final output. The output is stored in a file named 'output_1.txt' .

- Part 2 : display the details of 4 fields of a particular country. The user is asked about the name of the country as well as start and end date. The country, start date and end date are written in a file which is to be feed into the mapper. Also all country names are written in a file which is to be feed into the mapper along with date. The mapper passes the data to the reducer which inturn preprocess and filters and send it to the reducer for the final output.The logic of closest country in respect to percentage change is also done int the reducer. The output is stored in a file named 'output_2.txt' .

Outputs can also be displayed from the menu.

Module 3.2

There are 2 Parts present in this module.

- Part 1 : Displays all the worldwide responses given a time range. The user is asked about (start month start year end month end year). All the responces were extracted and saved in module 2. Given the time range, goes through all the subfolders and extracts all txt files which are inbetween those time range and puts all text in a single txt file. This is then fed to mapper. The mapper passes the data to the reducer which preprocess the data and sends it to the reducer for the final output. The output is displayed on terminal.

- Part 2: Displays the news and responses for a particular country over a period of time. The user is asked about the name of the country as well as start and end date. The country, start date and end date are written in a file which is to be feed into the mapper. The inputs are feed into the mapper along with date. The mapper passes the data to the reducer which inturn preprocess and filters and send it to the reducer for the final output. The logic of closest country in respect to percentage change is also done int the reducer. The output is stored in a file named 'output_2.txt' .

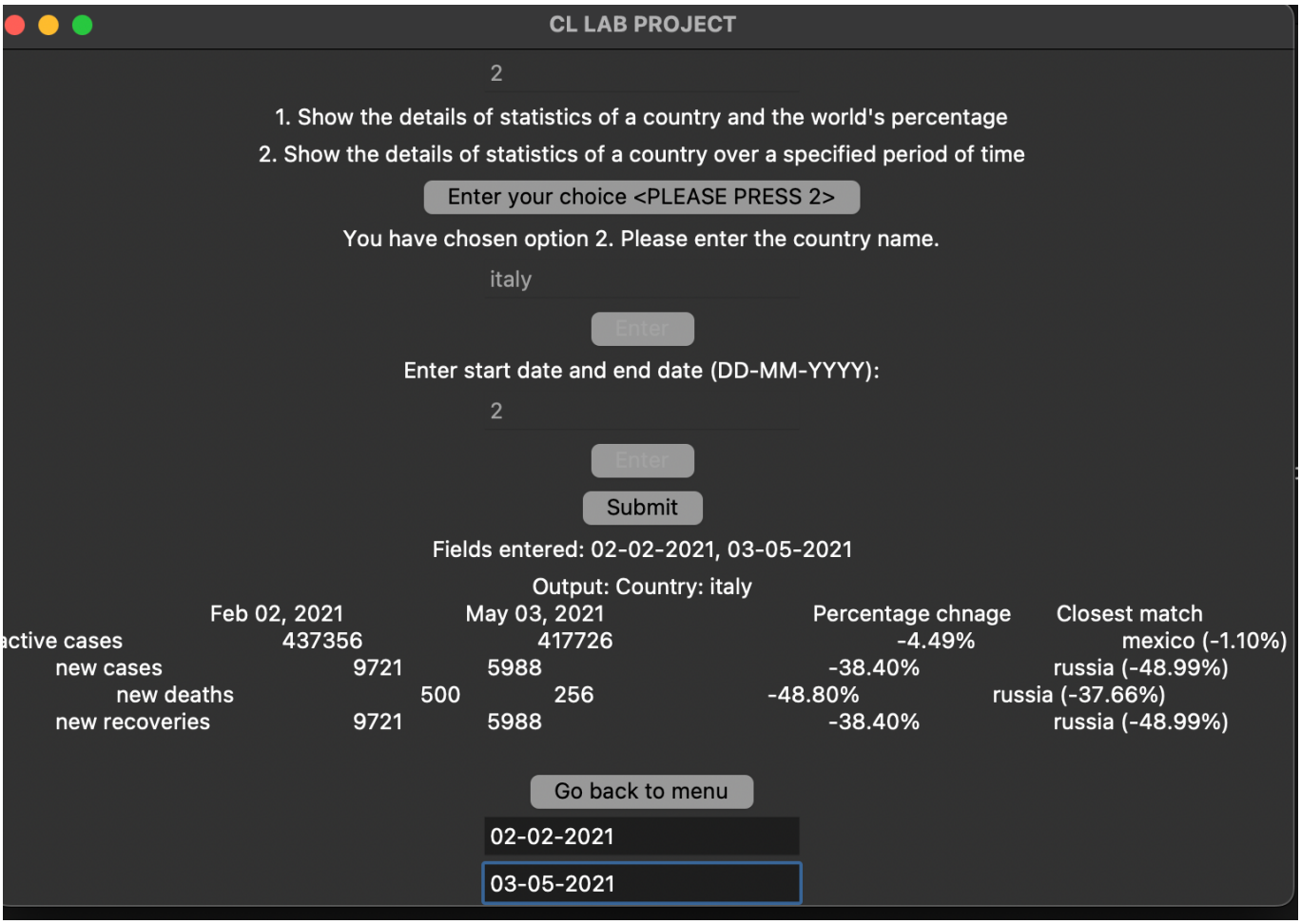
Module 4

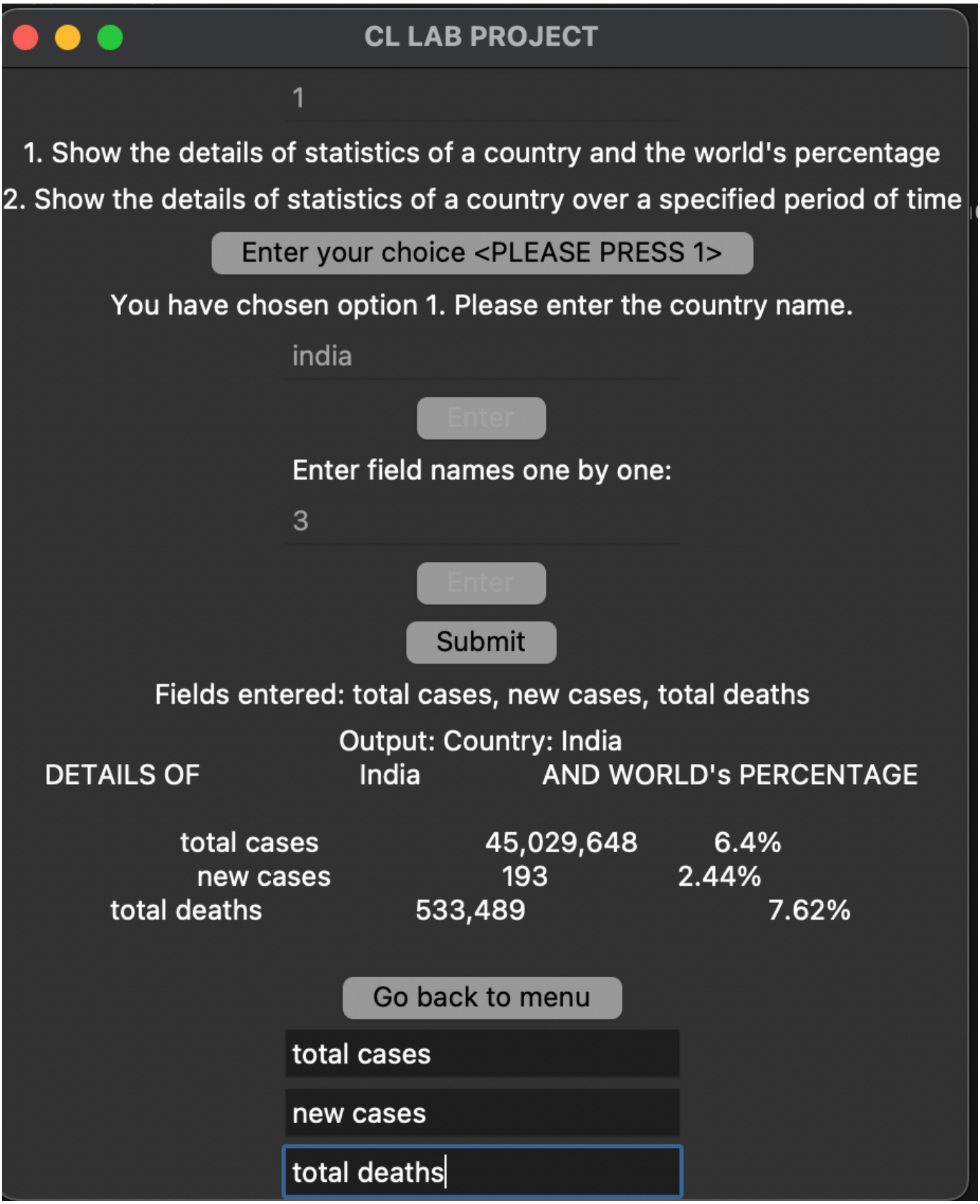
A menu is created to interact with the user. The user can select the module and the required query. The entire project can be accessed from the menu. The user can also see the output of the queries from the menu. GUI part has been done only for Module 3.1 for displaying the outputs. But is not fully functional.Hence scope for improvement is there. GUI is done using tkinter library in python.

SCREENSHOTS of the project:

```
(base) ➜  Design-Lab-Lex-Yacc-and-NoSQL-crawling-Covid-Statistics-and-News git:(master) ✗ python3 driver.py
WELCOME TO LEX-YACC BASED COVID-19 DATA ANALYSIS PROJECT
HERE's WHAT WE HAVE FOR YOU:
Module 1: Extracting data from WORLDOMETER website [a. Tabular data for all countries, b. Individual country data from scripts]
Module 2: Extracting data from WIKIPEDIA website
Module 3: USING MAPPER_COMBINER-REDUCER to show the result of queries

LESSS GO !!!!!!!!!!!!!!!!

PRESS 1 for WORLDOMETER DATA EXTRACTION & QUERIES
PRESS 2 for WIKIPEDIA DATA EXTRACTION & QUERIES
PRESS e/E to exit
```

```
WELCOME TO MODULE 3.1 SHOWING DETAILS OF COVID DATA
1. Show the details of statistics of a country and the world's percentage
2. Show the details of statistics of a country over a specified period of time
Press b/B to go back to previous menu
Enter your choice: 1

1. Run the queries on PRE-EXISTING data
2. EXTRACT data from web and then run the queries
Enter your choice: 1
Enter the country name: France
press X to see all fields, else press enter to continue:

Enter the number of fields you want to see: 2
enter field name:  total cases
enter field name:  total deaths
chmod +x ./mapper.py ./reducer.py ./combiner.py
python3 ./mapper.py ../../Module1/allcountries_table/output.txt country.txt fields.txt | python3 combiner.py | python3 reducer.py > output_1.txt
The details are shown in the file ./A/output.txt

Do you want to see the output? (y/n): y
Country: France
DETAILS OF                    France              AND WORLD's PERCENTAGE

total cases                    40,138,560          5.7%
total deaths                   167,642                                 2.39%

Do you want to continue? (y/n):
```




CL LAB PROJECT
1
1. Show the details of statistics of a country and the world's percentage
2. Show the details of statistics of a country over a specified period of time
Enter your choice <PLEASE PRESS 1>
You have chosen option 1. Please enter the country name.
india
Enter
Enter field names one by one:
3
Enter
Submit
Fields entered: total cases, new cases, total deaths
Output: Country: India
DETAILS OF India AND WORLD's PERCENTAGE
total cases 45,029,648 6.4%
new cases 193 2.44%
total deaths 533,489 7.62%
Go back to menu
total cases
new cases
total deaths

```
PRESS 1 for WORLDOMETER DATA EXTRACTION & QUERIES
PRESS 2 for WIKIPEDIA DATA EXTRACTION & QUERIES
PRESS e/E to exit
2
1. PRESS 1 for WORLDWIDE NEWS (global) data
2. PRESS 2 for COUNTRYWISE NEWS data for specific country
PRESS b/B to go back to main menu
PRESS e/E to exit
1
WELCOME TO MODULE 3.2 SHOWING RESPONSES BETWEEN A TIME RANGE

Press 1 Enter time range
Press b/B to Go back to previous menu

Enter your choice: 1
Enter starting month(1-12)
3
Enter starting year
2021
Enter ending month(1-12)
5
Enter ending year
2022
2021_3_march.txt
2021_4_april.txt
2021_5_may.txt
2021_6_june.txt
2021_7_july.txt
2021_8_august.txt
2021_9_september.txt
2021_10_october.txt
2021_11_november.txt
2021_12_december.txt
2022_1_january.txt
2022_2_february.txt
2022_3_march.txt
2022_4_april.txt
2022_5_may.txt
Response stored in output.txt
Output is saved in output.txt. Do you want to see the output? (y/n)
```

```
PRESS 1 for WORLDOMETER DATA EXTRACTION & QUERIES
PRESS 2 for WIKIPEDIA DATA EXTRACTION & QUERIES
PRESS e/E to exit
2
1. PRESS 1 for WORLDWIDE NEWS (global) data
2. PRESS 2 for COUNTRYWISE NEWS data for specific country
PRESS b/B to go back to main menu
PRESS e/E to exit
2

Press 1 to see country list
Press 2 to see date range of a country
Press 3 to enter country and date, and see the information
PRESS b/B to go back to main menu

Enter your choice: 1
India
Australia
Malaysia
England

Press 1 to see country list
Press 2 to see date range of a country
Press 3 to enter country and date, and see the information
PRESS b/B to go back to main menu

Enter your choice: 2

Enter country: india
Start Date: 02-Feb-2020 | End Date: 26-May-2021

Press 1 to see country list
Press 2 to see date range of a country
Press 3 to enter country and date, and see the information
PRESS b/B to go back to main menu

Enter your choice: 3

Enter country: india
Example date format: 01-April-2021
Enter start date (dd-Month-YYYY): 10-February-2021
Enter end date (dd-Month-YYYY): 04-July-2021
Output is saved in output.txt. Do you want to see the output? (y/n)
y
01-Apr-2021| Vaccinations were made available to all Indians over the age of 45.9110
02-Apr-2021| India reported 89,129 new cases, the most in more than six months.9111
02-May-2021| India reported a record 3,689 new deaths and 392,488 new cases.9157
04-May-2021| India surpassed 20 million total cases since the start of the pandemic.9159
```

Team Members:

Aakash Gupta

📖

Rajdeep Ghosh

📖

Satya Prakash Nayak

📖