Design Laboratory (CS69202)

Spring Semester 2024

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

Date: February 07, 2024

Deadline: February 11, 2024 11:59PM

Important Instructions:

1. Write Python code using PLY to extract the above fields. Your program should show all the possible query fields a user can ask for (from the above list items).
2. You must think correctly about what kind of errors can come in the process and try to handle them. Use the PLY package in python. PLY ref: https://www.dabeaz.com/ply/
3. You must NOT use any other parsing tools apart from PLY (ex: Beautiful Soup is a strict no or any other framework) . Should anyone not adhere to this instruction, they will be awarded ZERO marks.
4. Your code should address the objectives using PLY. Anyone found addressing the objective with no such use of PLY will be awarded ZERO marks.
5. Whatever data you extract needs to be stored in a text file. Creating text file convention is left to you as a design choice. Make note that mapper and reducer codes are to be written addressing the queries.
6. You need to design a menu-driven program to resolve user queries. We leave the design of the menu up to you, but keep in mind that your menu should be easy to use and address all queries. The user should also be able to go back to the previous menu.
7. As a bonus, you can create a GUI window satisfying the menu driven paradigm mentioned above.
8. You can write a readme file to provide any particular instructions related to program execution steps, input format, or anything that you might think is useful for the evaluator while evaluating the assignment.
9. Not adhering to these instructions can incur a penalty (worst case being 0 marks).
10. Plagiarism in any form is not allowed. Students found copying/sharing code will be awarded 0 marks.
11. All errors should be handled properly.
12. You are free to create any number of files.
13. Save this in a folder named in the format: DesLab_Project_<Group Name.>. Compress this folder to zip format, creating a compressed file DesLab_Project_Group_<Group Name>.zip. Upload this compressed file to moodle. Example: If your group belongs to a group named LYN, the folder should be DesLab_Project_LYN and the compressed file should be DesLab_Project_LYN.zip.
14. Make a group of 3 members.

15. Git repositories are to be maintained by each group. Commit should be done by all the members of the group.

Project : Lex Yacc + NoSQL

Module 1 (Crawling worldometers website) :
Link : https://www.worldometers.info/coronavirus/

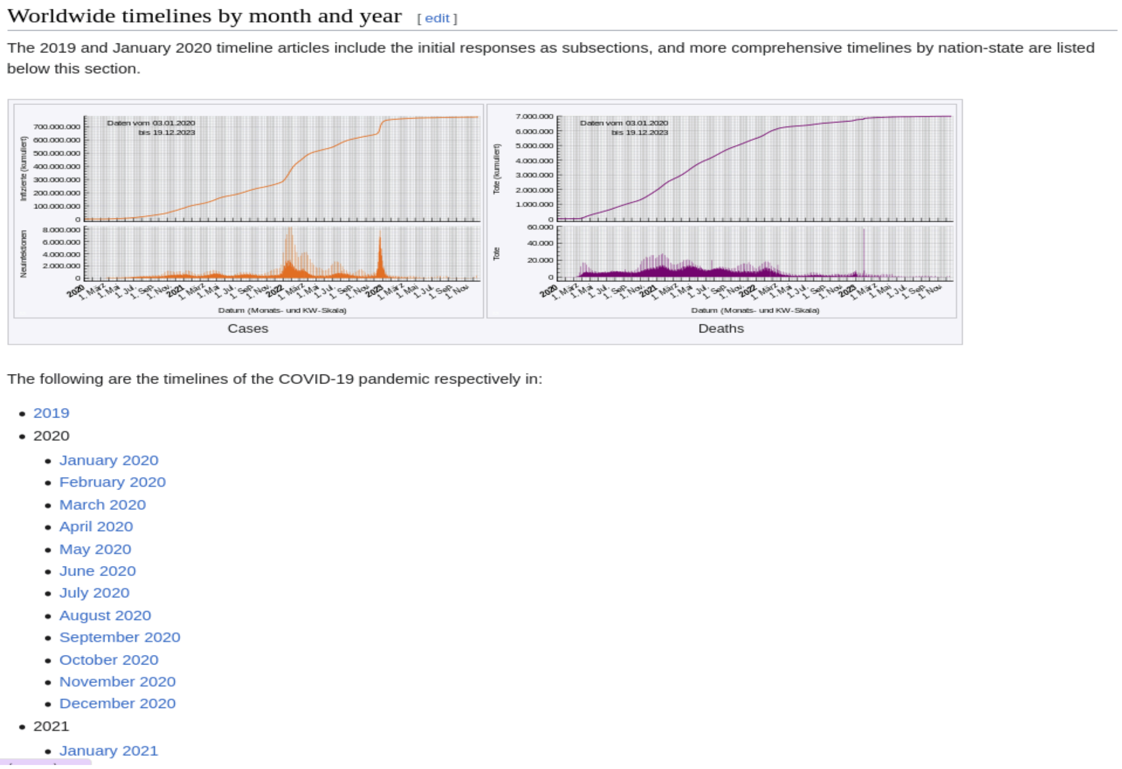
1. Worldometer is a website where you will find all the coronavirus-related statistics world/continent/country-wise, like total cases, active cases, total death, new death, total recovered, serious/critical cases, total tests are done, etc.
2. You will be provided with a file named "worldometers_countrylist.txt," which contains the continent-wise country's name.
3. Write a python code that reads the main URL & saves the page along with all the pages of countries present in the given file in HTML format.
4. Create grammar that can be used to extract the following fields for any country/continent/world based on yesterday's data and store the data in text files.
    a. Total cases
    b. Active cases
    c. Total deaths
    d. Total recovered
    e. Total tests
    f. Death/million
    g. Tests/million
    h. New case
    i. New death
    j. New recovered
5. Now for a given country, you need to extract the queries below for all the time.
    a. Active cases.
    b. Daily death
    c. New Recovered
    d. New cases
6. Write (python code using PLY) to extract the above fields. Your program should show all the possible query fields a user can ask for (from the above list items).
7. Whatever data you extract needs to be stored in a text file. Creating text file convention is left to you as a design choice.

Module 2 (Crawling Wikipedia Covid-19 timeline)

Link : https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic

1. Crawl the Wikipedia Covid-19 timeline page, do the following as below
   a. Extract all the worldwide news for all the times.
   b. Extract all the worldwide responses for all the times.

You will find news in individual timeline pages as shown below and need to extract that :



Worldwide timelines by month and year [edit]

The 2019 and January 2020 timeline articles include the initial responses as subsections, and more comprehensive timelines by nation-state are listed below this section.

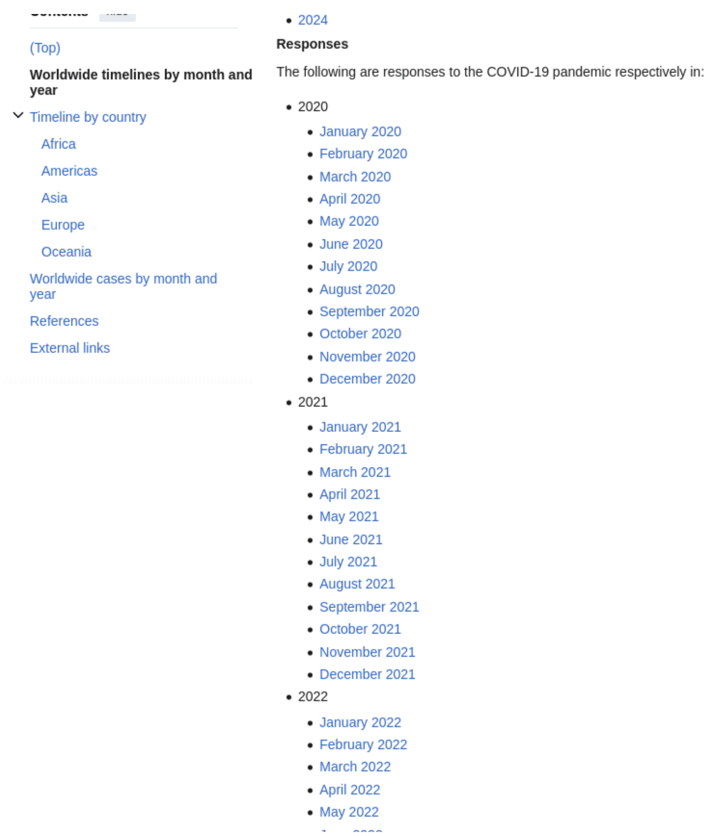
Cases

Deaths

The following are the timelines of the COVID-19 pandemic respectively in:

- 2019
- 2020
  - January 2020
  - February 2020
  - March 2020
  - April 2020
  - May 2020
  - June 2020
  - July 2020
  - August 2020
  - September 2020
  - October 2020
  - November 2020
  - December 2020
- 2021
  - January 2021

You will find news in individual response pages as shown below and need to extract that :



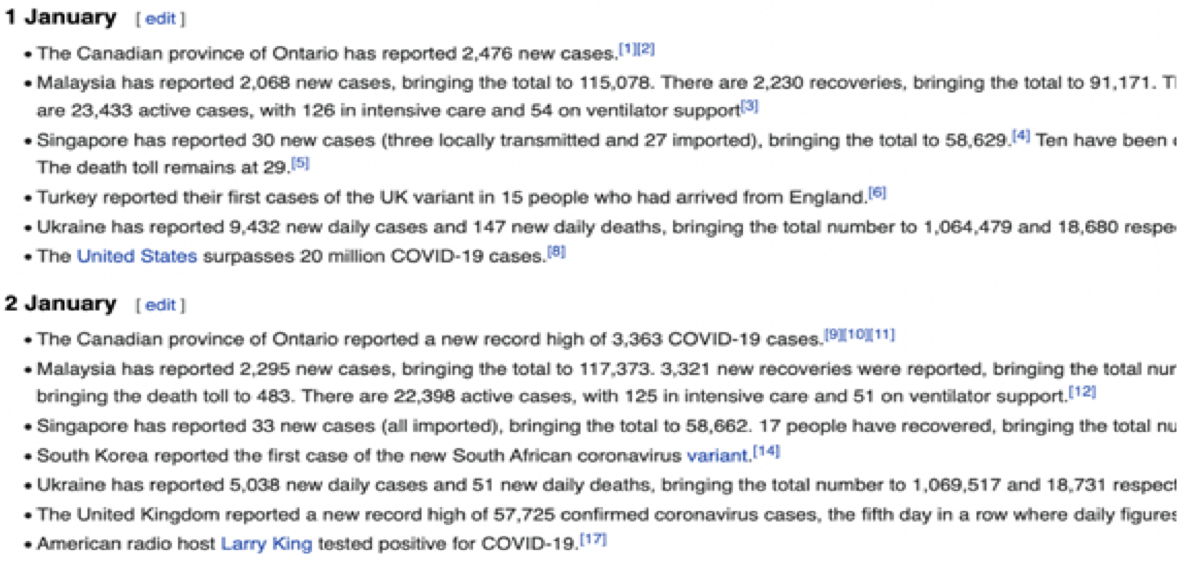
- 2024

Responses

The following are responses to the COVID-19 pandemic respectively in:

- 2020
  - January 2020
  - February 2020
  - March 2020
  - April 2020
  - May 2020
  - June 2020
  - July 2020
  - August 2020
  - September 2020
  - October 2020
  - November 2020
  - December 2020
- 2021
  - January 2021
  - February 2021
  - March 2021
  - April 2021
  - May 2021
  - June 2021
  - July 2021
  - August 2021
  - September 2021
  - October 2021
  - November 2021
  - December 2021
- 2022
  - January 2022
  - February 2022
  - March 2022
  - April 2022
  - May 2022

The document will look something alike as given below:



1 January [edit]

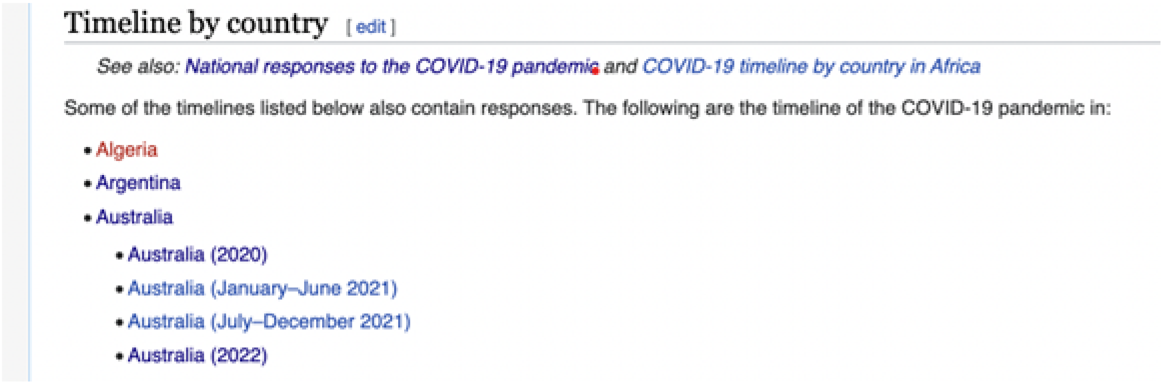
- The Canadian province of Ontario has reported 2,476 new cases.[1][2]
- Malaysia has reported 2,068 new cases, bringing the total to 115,078. There are 2,230 recoveries, bringing the total to 91,171. T are 23,433 active cases, with 126 in intensive care and 54 on ventilator support[3]
- Singapore has reported 30 new cases (three locally transmitted and 27 imported), bringing the total to 58,629.[4] Ten have been The death toll remains at 29.[5]
- Turkey reported their first cases of the UK variant in 15 people who had arrived from England.[6]
- Ukraine has reported 9,432 new daily cases and 147 new daily deaths, bringing the total number to 1,064,479 and 18,680 respe
- The United States surpasses 20 million COVID-19 cases.[8]

2 January [edit]

- The Canadian province of Ontario reported a new record high of 3,363 COVID-19 cases.[9][10][11]
- Malaysia has reported 2,295 new cases, bringing the total to 117,373. 3,321 new recoveries were reported, bringing the total nur bringing the death toll to 483. There are 22,398 active cases, with 125 in intensive care and 51 on ventilator support.[12]
- Singapore has reported 33 new cases (all imported), bringing the total to 58,662. 17 people have recovered, bringing the total nu
- South Korea reported the first case of the new South African coronavirus variant.[14]
- Ukraine has reported 5,038 new daily cases and 51 new daily deaths, bringing the total number to 1,069,517 and 18,731 respect
- The United Kingdom reported a new record high of 57,725 confirmed coronavirus cases, the fifth day in a row where daily figures
- American radio host Larry King tested positive for COVID-19.[17]

2. Store the above results in suitable text files of small sizes.
   Example : Store the responses of 2021 in 12 text files where each text file contains information of a particular month of 2021 and so on for other queries.
3. For the next part of extraction, consider India (excluding Kerala), Australia, Malaysia, England and Singapore as countries for which information is to be extracted.

4. Given a country name, extract the news information that is available for that country. Also store the dates for which the information was given. Please follow the below screenshots for clarification

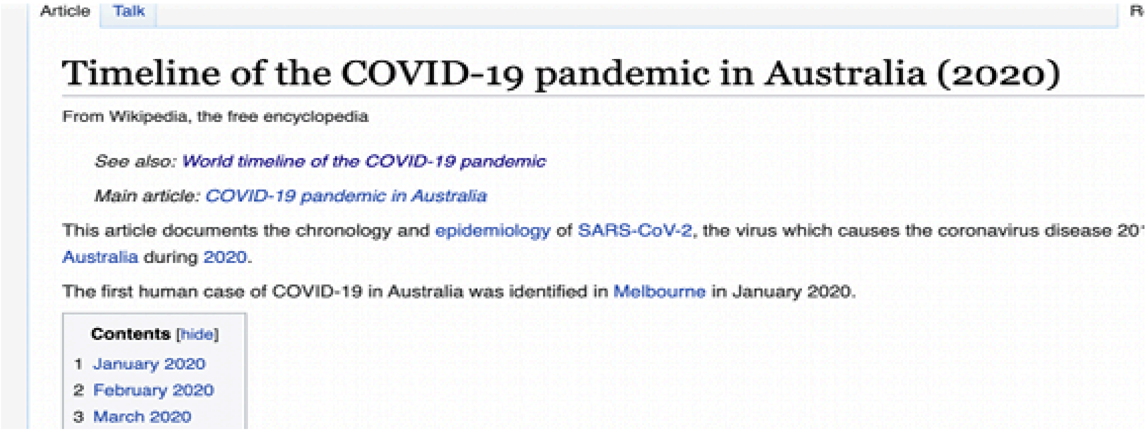


Timeline by country [edit]

See also: National responses to the COVID-19 pandemic and COVID-19 timeline by country in Africa

Some of the timelines listed below also contain responses. The following are the timeline of the COVID-19 pandemic in:

- Algeria
- Argentina
- Australia
  - Australia (2020)
  - Australia (January–June 2021)
  - Australia (July–December 2021)
  - Australia (2022)



Timeline of the COVID-19 pandemic in Australia (2020)

From Wikipedia, the free encyclopedia

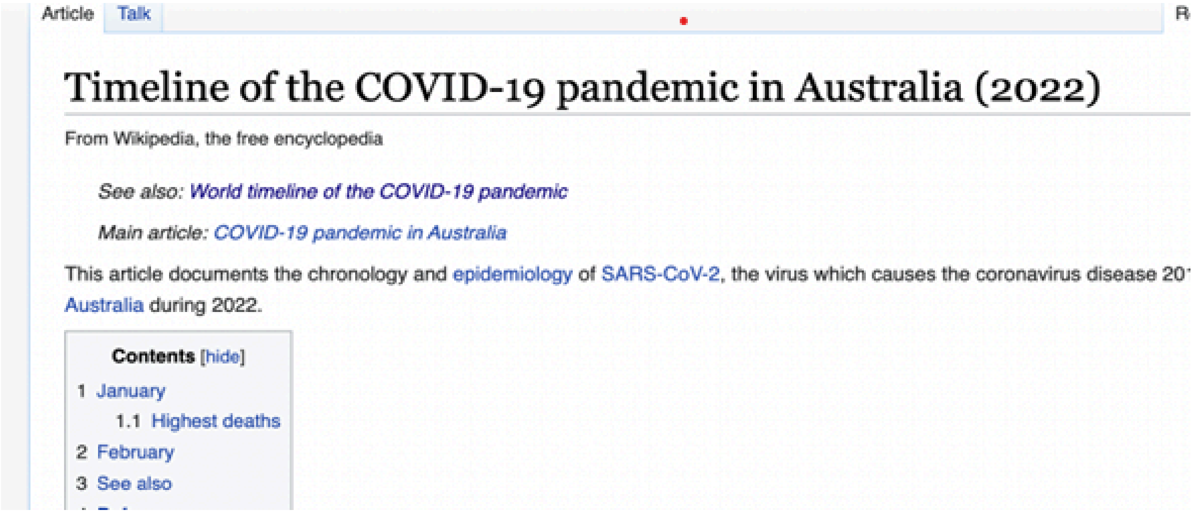
See also: World timeline of the COVID-19 pandemic

Main article: COVID-19 pandemic in Australia

This article documents the chronology and epidemiology of SARS-CoV-2, the virus which causes the coronavirus disease 20 Australia during 2020.

The first human case of COVID-19 in Australia was identified in Melbourne in January 2020.

Contents [hide]



Timeline of the COVID-19 pandemic in Australia (2022)

From Wikipedia, the free encyclopedia

See also: World timeline of the COVID-19 pandemic

Main article: COVID-19 pandemic in Australia

This article documents the chronology and epidemiology of SARS-CoV-2, the virus which causes the coronavirus disease 20 Australia during 2022.

Contents [hide]

5. For Queries 1-4, store related information in text files appropriately. Make sure your data is chunked into text files of small sizes.
Example :

Store the data of Australia (2020), Australia (Jan-Jun 21), Australia(Jul-Dec 21) and Australia (2022) into 4 separate text files respectively. (Check above image for reference headlined ‘Timeline By Country’).

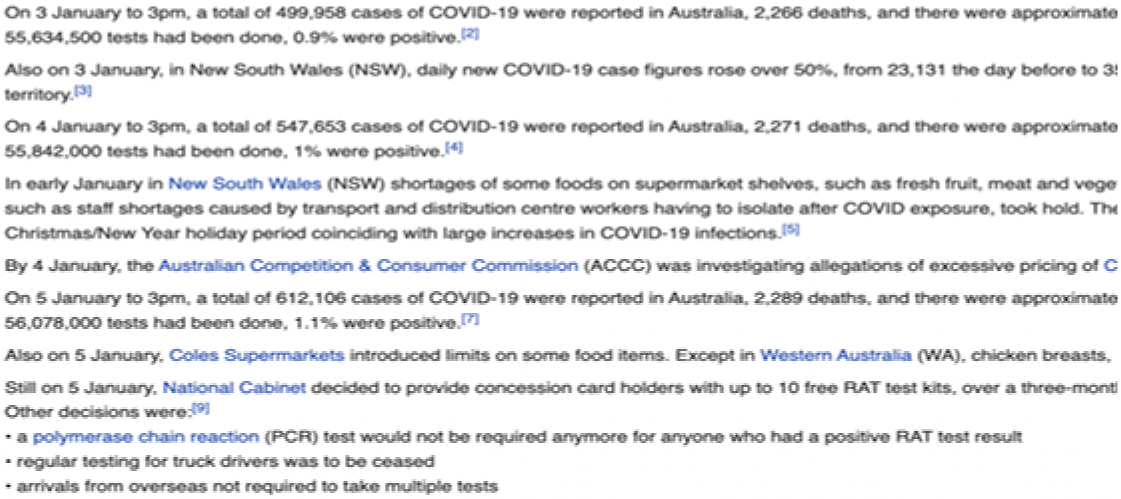
Module 3.1(Addressing Queries of Worldometer Covid Statistics)

1. Use the data extracted in Module 1 to address the queries. The queries are to be retrieved from text files using the MapCombineReduce paradigm of NoSQL.You have to integrate the paradigm inside the python code and retrieve the output query.
2. Use yesterday’s data to answer the queries given below for each country whose statistics are extracted (321 countries in tabular data as discussed in class). Also, provide the percent of total world cases for each of the queries. You can ignore other fields except the below queries.
    a. Total cases
    b. Active cases
    c. Total deaths
    d. Total recovered
    e. Total tests
    f. Death/million
    g. Tests/million
    h. New case
    i. New death
    j. New recovered
3. Now for a given country in “worldometers_countrylist.txt”, you need to answer the queries below-[given time range]
    a. Change in active cases in %
    b. Change in daily death in %
    c. Change in new recovered in %
    d. Change in new cases in %
    e. Closest country similar to any query between a-d. Example : Input is India, and out of all the countries in the text file mentioned, find the country whose value is such that |Value of specified country - Value of country x| is close to 0. x-> closest country.

    Ask the user for the start and end date. [dd-mm-yyyy format]
4. So basically you need to design a menu-driven program resolving user queries. We leave the design of the menu up to you, but keep in mind that your menu should be easy to use and address all the queries. The user should also be able to go back to the previous menu.

Module 3.2(Addressing Queries of Wikipedia Covid News)

1. Use the data extracted in Module 2 to address the queries. The queries are to be retrieved from text files using the MapCombineReduce paradigm of NoSQL. You have to integrate the paradigm inside the python code and retrieve the output query.
2. Answer the user queries as below (given a time range as input, including start & end date)
    a. Show all the worldwide news between the time range.
    b. Show all the worldwide responses between the time range.
3. Given a country name mentioned in Module 2, show the date range for which news information is available for that country. Suppose the user enters a country name as Australia, then your program should output the range as January,2020-May,2022.
4. Given a country name and date range, extract all the news between the time duration. Suppose the user enters the country name as Australia and date range as [02-01-2022,05-01-2022], then your output for the news should include the sentences from the below :



On 3 January to 3pm, a total of 499,958 cases of COVID-19 were reported in Australia, 2,266 deaths, and there were approximate
55,634,500 tests had been done, 0.9% were positive.[2]

Also on 3 January, in New South Wales (NSW), daily new COVID-19 case figures rose over 50%, from 23,131 the day before to 3!
territory.[3]

On 4 January to 3pm, a total of 547,653 cases of COVID-19 were reported in Australia, 2,271 deaths, and there were approximate
55,842,000 tests had been done, 1% were positive.[4]

In early January in New South Wales (NSW) shortages of some foods on supermarket shelves, such as fresh fruit, meat and vege
such as staff shortages caused by transport and distribution centre workers having to isolate after COVID exposure, took hold. The
Christmas/New Year holiday period coinciding with large increases in COVID-19 infections.[5]

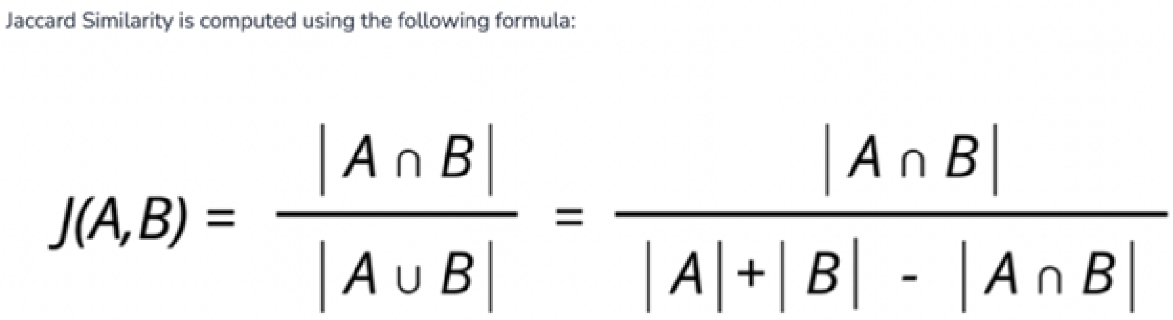
By 4 January, the Australian Competition & Consumer Commission (ACCC) was investigating allegations of excessive pricing of C

On 5 January to 3pm, a total of 612,106 cases of COVID-19 were reported in Australia, 2,289 deaths, and there were approximate
56,078,000 tests had been done, 1.1% were positive.[7]

Also on 5 January, Coles Supermarkets introduced limits on some food items. Except in Western Australia (WA), chicken breasts,

Still on 5 January, National Cabinet decided to provide concession card holders with up to 10 free RAT test kits, over a three-mont
Other decisions were:[9]
• a polymerase chain reaction (PCR) test would not be required anymore for anyone who had a positive RAT test result
• regular testing for truck drivers was to be ceased
• arrivals from overseas not required to take multiple tests

5. Provide the name of the closest country according to the Jaccard similarity of the extracted news. (ignore stopwords while calculating the Jaccard similarity) For the running example, Australia as the country and [02-01-2022,05-01-2022] as the date range, you will extract the news for all the countries that are on the country list file. Then you need to calculate the Jaccard similarity of the extracted news (think of them as a set of words) between Australia and all the other countries, and report the one according to the score.



Jaccard Similarity is computed using the following formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Ref : https://www.geeksforgeeks.org/find-the-jaccard-index-and-jaccard-distance-between-the-two-given-sets/

6. So basically you need to design a menu-driven program resolving user queries. We leave the design of the menu up to you, but keep in mind that your menu should be easy to use and address all the queries. The user should also be able to go back to the previous menu.

Module 4(Combining Modules 3.1 and 3.2)

1. Combine Modules 3.1 and 3.2 and create a menu driven method to access either of the modules. We leave the design of the menu up to you, but keep in mind that your menu should be easy to use and address all the queries. The user should also be able to go back to the previous menu.
2. In the combining phase, mention the date as to when the extraction was done. In your menu driven method, print the date when the data was last extracted.
3. You are left to choose as to how to implement the GUI [Bonus]. Also, the queries generated are dynamic and it is your design decision as to how to implement it.
4. For information extraction w.r.t. Modules 1 and 2, store the information in text files as per your design decision.
5. For the queries mentioned w.r.t. Modules 3.1 and 3.2. The queries are to be extracted from the text files using Map Combine Reduce Paradigm and display the results.
6. Create a Readme File describing your approach. Also mention the

Module 5(Maintaining Project in Github Repository)

1. Each group should maintain a Github repository as mentioned. The tutorial for using Git is https://www.youtube.com/watch?v=RGOj5yH7evk
2. Each member of the Group should commit your project in the repository.
3. Do the needful in the Github repository (Uploading codes, readme etc.). Update the readme file by sharing your Github repository. Also mention your Github ids against your collaborators in the readme file for identification. Make it private so that others cannot see your work.

4. Make sure that all of the members of the group commit at least once. The commit log of the Github repository will be checked for confirmation.