Assignment 5
Web Crawling and Extracting Information-Part2

Computing Lab-II

9th Feb 2022

Add the below extra functionalities with your previous menu-driven program. As earlier, you need to write grammar rules using PLY or Lex/Yacc to extract the query results.

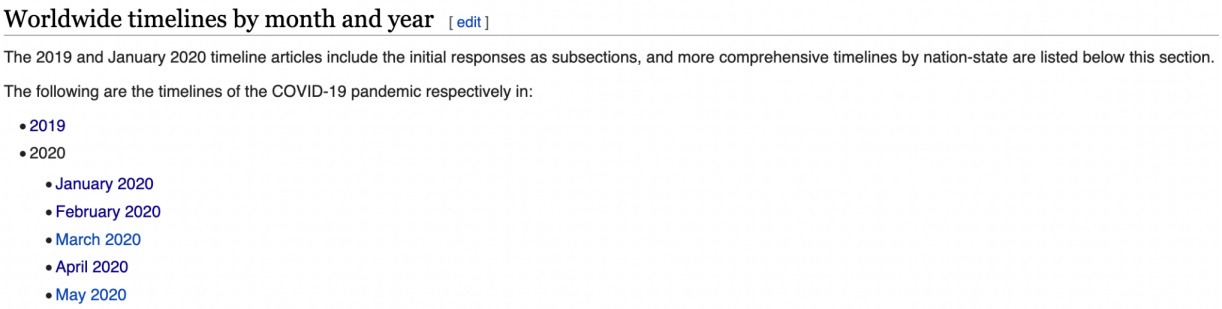
Task (Crawling Wikipedia Covid-19 timeline→*https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic*)

1. Crawl the Wikipedia Covid-19 timeline page, and 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.

    For both the above cases, plot a word cloud.

    Suppose one wants to see all the news between January 2, 2021, to March 31, 2021, then he/she will enter the input as [02-01-2021,31-03-2021].

    You will find the news from the section shown below, then you need to crawl and extract through the individual timeline URLs -



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.

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

- 2019
- 2020
    - January 2020
    - February 2020
    - March 2020
    - April 2020
    - May 2020

You will find the response information from the section shown below-

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

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

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. Tl 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. Given two non-overlapping time ranges, answer the below queries-
    a. Plot two different word clouds for all the common words (ignore stopwords) and only covid related common words.
    b. Print the percentage of covid related words in common words (ignore stopwords).
    c. Print the top-20 common words (ignore stopwords) and covid related words.

Common words contain all the words in the corpus (for your case it is the collection of sentences in the extracted news) that you are dealing with excluding the stopwords. A stop word is a commonly used word (such as “the”, “a”, “an”, “in”), they don’t carry much information about the concerned topic.

Stop-word tutorial: https://www.geeksforgeeks.org/removing-stop-words-nltk-python/

We will provide you with a file containing all the covid related words.

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites. [Ref-geeksforgeeks]

Suppose you want to visualize important words from the popular TV show Game of Thrones, the word cloud for it would look like below -

Ref- 1. https://www.geeksforgeeks.org/generating-word-cloud-python/
2. https://python-course.eu/applications-python/python-wordcloud-tutorial.php
3. https://www.datacamp.com/community/tutorials/wordcloud-python

3. Given a country name, show the date range for which news information is available for that country.
Suppose the user enters a country name as Austalia, then your program should output the range as January,2020-February,2021. 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)

Article Talk

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]

Article Talk

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]

4. Given a country name and date range, extract all the news between the time duration, plot a word cloud. (ignore stopwords while plotting the word cloud)
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 names of the top-3 closest countries 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, report the top-3 according to the score.

   Jaccard Similarity is computed using the following formula:

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

   Jaccard similarity (ref.) - https://studymachinelearning.com/jaccard-similarity-text-similarity-metric-in-nlp/

6. Provide names of the top-3 closest countries according to Jaccard similarity of covid words match.
   As described in the previous point, here also you need to do the same but here your word list should include only the words from the covid words file.
7. Here also your program should have the capacity to go back in the menu.
8. You need to merge the solutions of both assignments somehow so that the user can access all the queries from the previous and current assignments from the same menu.
9. 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. It should be a lightweight one-stop destination for all covid related queries.
10. 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.

We will provide the country and covid word list.

Deliverables:

1. Codes for the tasks and readme file if any.
2. Save this in a folder named in the format: <Roll No.>_CL2_A5. Compress this folder to zip format, creating a compressed file <Roll No.>_CL2_A5.zip. Upload this compressed file to moodle. Example: If your roll no. is 21CS60R05, the folder should be 20CS60R05_CL2_A4, and the compressed file should be 21CS60R05_CL2_A5.zip.
3. Not adhering to these instructions can incur a penalty.

Evaluation Scheme

Main Task: 80 marks(all the fields grammar + correct output)
Merge two assignments: 10 marks
Error handling: 5 marks
Coding Style: 5 marks

Important Instructions

1. Plagiarism Rule: If your code matches (more than 50%) with another student's code, all those students whose codes match will be awarded zero marks without any evaluation. Therefore, it is your responsibility to ensure that neither you copy anyone's code nor anyone can copy yours.
2. Code error: If your code doesn't run or gives an error while running, you will be awarded zero marks.