https://github.com/Niharika-20-MAI-01/Winter-2021-DWDM-LabCSE5021-Niharika-20MAI01/blob/main/20MAI0001\_DWDM-Lab-CSE-5021 Lab-Assignment-6%20Implementation%20of%20Web%20Scrapping%20on%20WikipediA%20Webpage%20for%20%20COVID-19%20%20in%20West%20Bengal%20%20%20.ipynb https://github.com/Niharika-20-MAI-01/Winter-2021-DWDM-LabCSE5021-Niharika-20MAI01 In [543... | #Importing necessary Libraries import pandas as pd import numpy as np import matplotlib.pyplot as plt import seaborn as sns %matplotlib inline import re import time from datetime import datetime import matplotlib.dates as mdates import matplotlib.ticker as ticker Importing 'requests' Python library to download the desired (Input) WikipediA WebPage containing the detailed record of the effect and situation of the COVID-19 pandemic in West Bengal, for implementing Web Scrapping using the requests.get() method: from urllib.request import urlopen import requests page = requests.get("https://en.wikipedia.org/wiki/COVID-19\_pandemic\_in\_West\_Bengal") #pg = requests.get("https://en.wikipedia.org/wiki/COVID-19\_pandemic\_in\_West\_Bengal") page Out[544... < Response [200] > After running requests.get(), a Response object is obtained. This object has a status\_code property, which indicates if the page was downloaded successfully or not: In [545... page.status\_code Out[545... 200 To Import and use the 'BeautifulSoup' python library to parse the HTML document, corresponding to the WikipediA Webpage being scrapped, downloaded/obtained through using 'requests' and to extract the text from the div tag oh the HTML document: from bs4 import BeautifulSoup In [546... To create an instance of the BeautifulSoup class: 'soup\_wb' to parse the HTML document corresponding to the WikipediA Webpage being Scraped: soup\_wb = BeautifulSoup(page.content, 'html.parser') In [547.. #list(soup\_wb.children) In [548... To check the type of each element in the list: In [549... [type(item) for item in list(soup\_wb.children)] [bs4.element.Doctype, bs4.element.NavigableString, bs4.element.Tag] In [ ]: Extracting /Pulling important information using :Beautiful Soup's find() and find all() methods, from the 3 tables: A) "Details of Testing Lab " B) "Statistics (Samples Tested)" C) "COVID-19 pandemic in West Bengal by district" of the WikipediA WebPage for COVID-19 in West Bengal: In [550... table\_tags = soup\_wb.find\_all('table') #table\_tags  $\#Element\ table2 = table.get(1)$ In [551... table = soup\_wb.findAll('table', {'class':'wikitable'}) #table Extracting /Pulling important information from the A) "Details of Testing Lab" table of the WikipediA WebPage for COVID-19 in West Bengal:

DWDM-Lab-CSE-5021\_Lab-Assignment-6:- Implementation of

Web Scrapping for information extraction on the 'WikipediA'

pandemic in West Bengal and using PANDAS DataFrame to

https://github.com/Niharika-20-MAI-01/Winter-2021-DWDM-LabCSE5021-Niharika-20MAI01/blob/main/20MAI0001\_DWDM-Lab-

https://github.com/Niharika-20-MAI-01/Winter-2021-DWDM-LabCSE5021-Niharika-20MAI01/blob/main/20MAI0001 DWDM-Lab-

6%20Implementation%20of%20Web%20Scrapping%20on%20WikipediA%20Webpage%20for%20%20COVID-

6%20Implementation%20of%20Web%20Scrapping%20on%20WikipediA%20Webpage%20for%20%20COVID-

create csv files from out of the useful important information

Webpage about the effect and situation of the COVID-19

scraped and extracted from (the tables) of the WikipediA

Webpage

Git Repo link:

CSE-5021 Lab-Assignment-

CSE-5021\_Lab-Assignment-

Submitted By: 20MAI0001 - NIHARIKA MAITRA

To find the Input data and the .ipynb file used for this Assignment please visit:

Link for Input data / Web Page Link on which Web Scrapping is implented :

https://en.wikipedia.org/wiki/COVID-19\_pandemic\_in\_West\_Bengal

https://en.wikipedia.org/wiki/COVID-19\_pandemic\_in\_West\_Bengal

19%20%20in%20West%20Bengal%20%20%20.ipynb

19%20%20in%20West%20Bengal%20%20%20.ipynb

In [552.. In [553...

**for** tab **in** table:

data2 = []

import warnings

df2.head()

#data2

In [ ]:

In [554...

In [555...

Out[555...

In [556...

In [558...

In [ ]:

In [559...

In [560...

Out [560...

In [ ]:

In [561...

In [ ]:

In [562...

In [563...

In [ ]:

In [564...

In [ ]:

In [565...

Out [565...

In [ ]:

In [566..

-1.csv']

0

rows2 = table[0].find\_all('tr')

for i,item in enumerate(rows2):

import dask.dataframe as dd

df2 = dd.from\_pandas(dt2, npartitions=2)

dt2 = pd.DataFrame(data2)

#rows2 = table2.find\_all('tr', attrs={'style' : ''})

data2.append(item.text.strip().split("\n")[:5])

warnings.simplefilter(action='ignore', category=FutureWarning)

Type

**Public** 

IPGMER and SSKM Hospital None

School of Tropical Medicine None

implementing Web Scraping on it:

#csv file name: Details\_of\_Testing\_Lab\_t1

19 in West Bengal:

for i,item in enumerate(rows3):

Out[558... [['Samples Tested', '', '1,314,772'], ['Per million population', '', '14,609'],

['Active cases', '', '27,299'],

['Home quarantine', '', '39,731'],
['Recovered', '', '86,771'],
['Recovery rate', '', '74.48%'],
['Deceased', '', '2428'],
['Mortality Rate', '', '2.08%'],

['Under quarantine'],

import warnings

#dt3

dt3.head()

-t2-0.csv']

**for** tab **in** table:

data4 = []

import warnings

#data4

#dt4

dt4.head()

0

2

District

Alipurduar

Bankura

Birbhum

4 Cooch Behar

-t2-0.csv',

-t2-1.csv',

Total

rows4 = table[2].find\_all('tr')

for i,item in enumerate(rows4):

import dask.dataframe as dd

19 in West Bengal:

**Total cases** 

8,80,894

8,599

16,574

24,447

14,271

df4 = dd.from\_pandas(dt4,npartitions=4)

dt4 = pd.DataFrame(data4)

2

Samples Tested

Samples Tested

Positive cases

Active cases

Positive percentage

**1** Per million population

import dask.dataframe as dd

df3 = dd.from\_pandas(dt3,npartitions=1)

1,314,772

1,314,772

14,609

116,498

8.86%

27,299

implementing Web Scraping on it:

WebPage for COVID-19 in West Bengal:

data4.append(item.text.strip().split("\n")[:9])

warnings.simplefilter(action='ignore', category=FutureWarning)

Recoveries

7,49,296

8,003

14,122

19,348

12,996

dt4 = pd.DataFrame(data4[2:], columns=data4[1][:9]) #Formatting the header

Deaths

11,637

115

74

To Store in the form of 'csv' file the extracted important

df4.to\_csv('.../DWDM\_LAB\_CSE5021\_ASGN-6--Web\_Mining/Statistics-Samples\_Tested-t2-\*.csv')

Bengal after implementing Web Scraping on it:

dt3 = pd.DataFrame(data3)

['Positive cases', '', '116,498'], ['Positive percentage', '', '8.86%'],

['Institutional quarantine', '', '2,661'],

rows3 = table[1].find\_all('tr')

data3.append(item.text.strip().split("\n")[:3])

warnings.simplefilter(action='ignore', category=FutureWarning)

dt3 = pd.DataFrame(data3[0:], columns=data3[0][:3]) #Formatting the header

tor tab in table:

data3 = []

data3

Midnapore Medical College and Hospital None

3 North Bengal Medical College and Hospital None

dt2 = pd.DataFrame(data2[1:], columns=data2[0][:5]) #Formatting the header

DataFrame Representation of the "Details of Testing Lab" table

National Institute of Cholera and Enteric Dise...

Name of Testing Lab

None None

None None

None None

None

None

Method

RT-PCR

None

None

None

None

of the WikipediA WebPage for COVID-19 in West Bengal:

To Store in the form of 'csv' file the extracted important

WikipediA WebPage for COVID-19 in West Bengal after

df2.to\_csv('../DWDM\_LAB\_CSE5021\_ASGN-6--Web\_Mining/Details\_of\_Testing\_Lab\_t1-\*.csv')

information from the A) "Details of Testing Lab" table of the

Out[556... ['C:/Users/NRika/Documents/DWDM-CSE-5021-LAB-/DWDM\_LAB\_CSE5021\_ASGN-6--Web\_Mining/Details\_of\_Testing\_Lab\_t1

Extracting /Pulling important information from the B) "Statistics

(Samples Tested)" table of the WikipediA WebPage for COVID-

['Data provided by the Health and Family Welfare department as of 16 August 2020.[21]']]

DataFrame Representation of the "Statistics (Samples Tested)"

table of the WikipediA WebPage for COVID-19 in West Bengal:

To Store in the form of 'csv' file the extracted important

WikipediA WebPage for COVID-19 in West Bengal after

df3.to\_csv('../DWDM\_LAB\_CSE5021\_ASGN-6--Web\_Mining/Statistics-Samples\_Tested-t2-\*.csv')

Out[561... ['C:/Users/NRika/Documents/DWDM-CSE-5021-LAB-/DWDM\_LAB\_CSE5021\_ASGN-6--Web\_Mining/Statistics-Samples\_Tested

pandemic in West Bengal by district" table of the WikipediA

Extracting /Pulling important information from the C) "COVID-19

DataFrame Representation of the "COVID-19 pandemic in West

Bengal by district" table of the WikipediA WebPage for COVID-

Active cases

1,19,961

507

2,337

4,962

1,201

information from the A) "COVID-19 pandemic in West Bengal by

district" table of the WikipediA WebPage for COVID-19 in West

Out[566... ['C:/Users/NRika/Documents/DWDM-CSE-5021-LAB-/DWDM\_LAB\_CSE5021\_ASGN-6--Web\_Mining/Statistics-Samples\_Tested

C:/Users/NRika/Documents/DWDM-CSE-5021-LAB-/DWDM\_LAB\_CSE5021\_ASGN-6--Web\_Mining/Statistics-Samples\_Tested

C:/Users/NRika/Documents/DWDM-CSE-5021-LAB-/DWDM\_LAB\_CSE5021\_ASGN-6--Web\_Mining/Statistics-Samples\_Tested

'C:/Users/NRika/Documents/DWDM-CSE-5021-LAB-/DWDM\_LAB\_CSE5021\_ASGN-6--Web\_Mining/Statistics-Samples\_Tested

information from the A) "Statistics (Samples Tested)" table of the

C:/Users/NRika/Documents/DWDM-CSE-5021-LAB-/DWDM\_LAB\_CSE5021\_ASGN-6--Web\_Mining/Details\_of\_Testing\_Lab\_t1'

Name of Testing Lab Method Type Public National Institute of Cholera and Enteric Diseases RT-PCR **IPGMER** and SSKM Hospital Midnapore Medical College and Hospital North Bengal Medical College and Hospital School of Tropical Medicine Malda Medical College and Hospital Command Hospital Chittaranjan National Cancer Institute R.G. Kar Medical College and Hospital Murshidabad Medical College Burdwan Medical College and Hospital Medical College and Hospital, Kolkata Nil Ratan Sircar Medical College and Hospital Bankura Medical College and Hospital Suri District Hospital College of Medicine & JNM Hospital Raigunj Government Medical College and Hospital ESIC Medical College Asansol District Hospital TRUENAT Jhargram District Hospital Rampurhat Medical College and Hospital Calcutta National Medical College Falakata Super Speciality Hospital Diamond Harbour Government Medical College **Balurghat District Hospital** Serampore District Hospital **Uluberia Super Speciality Hospital** Jangipur Sub-District Hospital Egra Sub-District Hospital M R Bangur Hospital Chanchal Sub-District Hospital Biswa Bangla Krirangan COVID Hospital Purulia Government Medical College & Hospital Alipurduar District Hospital **Basirhat District Hospital** Contai Sub-District Hospital College of Medicine & Sagore Dutta Hospital **CBNAAT Gangarampur District Hospital** Private Apollo Gleneagles Hospitals RT-PCR Tata Medical Center Dr. Lal PathLabs Suraksha Diagnostic Peerless Hospital Sanaka Hospitals Medica Hospital **Desun Hospital RN Tagore Hospital** Jagannath Gupta Institute of Medical Sciences Remedy Lifecare AMRI Hospitals, Salt Lake AMRI Hospitals, Dhakuria **TRUENAT** Calcutta Medical Research Institute Anandaloke Sonoscan, Siliguri Woodlands Hospital **CBNAAT** SRL Lab Samples tested in labs outside the state RT-PCR Total Type Name of Testing Lab Method National Institute of Cholera and Enteric Diseases RT-PCR Public **IPGMER** and SSKM Hospital Midnapore Medical College and Hospital North Bengal Medical College and Hospital

School of Tropical Medicine Malda Medical College and Hospital

Command Hospital

Chittaranjan National Cancer Institute R.G. Kar Medical College and Hospital Murshidabad Medical College Burdwan Medical College and Hospital Medical College and Hospital, Kolkata Nil Ratan Sircar Medical College and Hospital Bankura Medical College and Hospital Suri District Hospital College of Medicine & JNM Hospital Raigunj Government Medical College and Hospital **ESIC Medical College** Asansol District Hospital TRUENAT Jhargram District Hospital Rampurhat Medical College and Hospital Calcutta National Medical College Falakata Super Speciality Hospital Diamond Harbour Government Medical College **Balurghat District Hospital** Serampore District Hospital Uluberia Super Speciality Hospital Jangipur Sub-District Hospital Egra Sub-District Hospital M R Bangur Hospital Chanchal Sub-District Hospital Biswa Bangla Krirangan COVID Hospital Purulia Government Medical College & Hospital Alipurduar District Hospital **Basirhat District Hospital** Contai Sub-District Hospital College of Medicine & Sagore Dutta Hospital CBNAAT Gangarampur District Hospital Private **Apollo Gleneagles Hospitals** RT-PCR Tata Medical Center Dr. Lal PathLabs Suraksha Diagnostic Peerless Hospital Sanaka Hospitals Medica Hospital **Desun Hospital** 

TRUENAT

**RN Tagore Hospital** Jagannath Gupta Institute of Medical Sciences Remedy Lifecare AMRI Hospitals, Salt Lake

AMRI Hospitals, Dhakuria Calcutta Medical Research Institute Anandaloke Sonoscan, Siliguri

**Woodlands Hospital** CBNAAT

SRL Lab

Samples tested in labs outside the state RT-PCR

Total

Samples Tested	13,14,772
Samples Tested	13,14,772
Per million population	14,609
Positive cases	1,16,498
Positive percentage	8.86%
Active cases	27,299
Under quarantine	
Institutional quarantine	2,661
Home quarantine	39,731
Recovered	86,771
Recovery rate	74.48%
Deceased	2428
Mortality Rate	2.08%
Data provided by the Health and Family Welfare department as of 16 August 2020.[21]	

## **COVID-19 pandemic in West Bengal by district**

District	Total cases	Recoveries	Deaths	Active cases
Total	8,80,894	7,49,296	11,637	1,19,961
Alipurduar	8,599	8,003	89	507
Bankura	16,574	14,122	115	2,337
Birbhum	24,447	19,348	137	4,962
Cooch Behar	14,271	12,996	74	1,201
Dakshin Dinajpur	10,157	9,149	83	925
Darjeeling	25,236	21,887	237	3,112