A.B.M. Fahim Shahriar

abmfahimshahriar@iut-dhaka.com

Web Scraping

Web scraping refers to the extraction of data from a website. This information is collected and then exported into a format that is more useful for the user. Be it a spreadsheet or an API.

Although web scraping can be done manually, in most cases, automated tools are preferred when scraping web data as they can be less costly and work at a faster rate.

But in most cases, web scraping is not a simple task. Websites come in many shapes and forms, as a result, web scrapers vary in functionality and features.

How do the Web Scrapers work?

Automated web scrapers work in a rather simple but also complex way. After all, websites are built for humans to understand, not machines.

First, the web scraper will be given one or more URLs to load before scraping. The scraper then loads the entire HTML code for the page in question. More advanced scrapers will render the entire website, including CSS and Javascript elements.

Then the scraper will either extract all the data on the page or specific data selected by the user before the project is run.

Ideally, the user will go through the process of selecting the specific data they want from the page. For example, you might want to scrape an Amazon product page for prices and models but are not necessarily interested in product reviews.

Lastly, the web scraper will output all the data that has been collected into a format that is more useful to the user.

Most web scrapers will output data to a CSV or Excel spreadsheet, while more advanced scrapers will support other formats such as JSON which can be used for an API.

The given task

Create a solution that crawls for articles about **coronavirus** from a news website, cleanses the response, stores in a cloud storage.

My solution:

For creating the solution I chose the Scrapy as my crawler framework. I used “bbc.com” website to crawl for news articles and filtered news related to coronavirus from the website. Four types of information were extracted from each news article:

1. News title
2. News link
3. News text
4. News tags

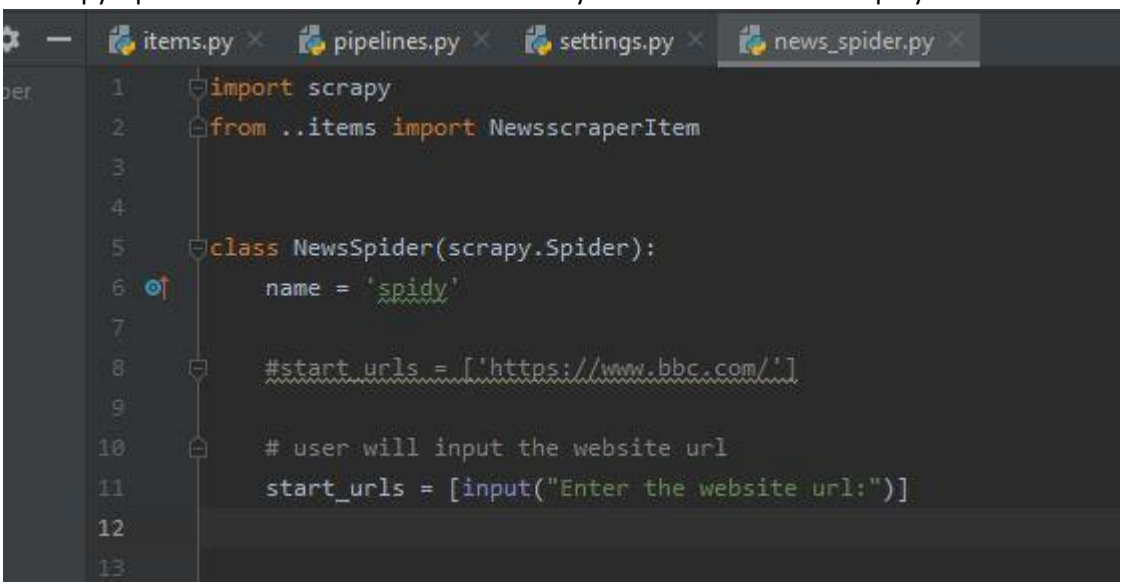
The extracted data was stored into mongoDB as well as CSV and JSON files were generated to show the results.

What is Scrapy?

Scrapy is a free and open-source web-crawling framework written in Python. Originally designed for web scraping, it can also be used to extract data using APIs or as a general-purpose web crawler. It is currently maintained by Scrapinghub Ltd., a web-scraping development and services company.

Step by step solution:

1. Firstly, I created a project in pycharm naming News Scraper.
2. Next, the Scrapy package was installed from pycharm.
3. After that, a scrapy project was created naming NewsScraper by the command: “scrapy startproject NewsScraper”. It automatically generated necessary files to crawl websites.
4. In scrapy ‘spider’ is used to crawl websites. In my solution I named it as ‘spidy’.



5. In the NewsSpider class the name of the spider, start ulrs and parse functions are defined.
6. In the parse function, I wrote codes for crawling the ‘bbc.com’ website for relevant. I used css selector and xpath to fetch specific data from the website.

```python

    # parser function that will crawl the website
    def parse(self, response):

        items = NewsscraperItem()

        all_blocks = response.css("li.media-list__item")

        for q in all_blocks:

            # CSS selector to fetch specific data from website
            news_title = str(q.css("a.block-link__overlay-link::text").extract())
            news_link = str(q.css("a.block-link__overlay-link").xpath("@href").extract())
            news_article = str(q.css("div.media__content").css("p.media__summary::text").extract())
            news_tag = str(q.css("div.media__content").css("a.media__tag::text").extract())

            if ('corona' or 'Corona' or 'virus' or 'Covid') in news_title:
                # data will be stored in the items
                items['news_title'] = news_title
                items['news_link'] = news_link
                items['news_article'] = news_article
                items['news_tag'] = news_tag
                yield items
```

NewsSpider

7. After scraping data from the website I stored them in the items dictionary.
8. To store the data into database ‘pipelines.py’ is used. But before that, in ‘settings.py’ file I had to uncomment ‘ITEM_PIPELINES’. I also had to install mongoDB in my computer.After that, I wrote code in the ‘pipelines.py’ file to store the data into database. This code may vary depending on which database you use. For my case, it was mongoDB.

```python

# Configure item pipelines
# See https://docs.scrapy.org/en/latest/topics/item-pipeline.html
ITEM_PIPELINES = {
    'NewsScraper.pipelines.NewsscraperPipeline': 300,
}
```

```python
# import pymongo package for mongoDB
import pymongo

class NewsscraperPipeline(object):
    def __init__(self):
        # creating connection with the database
        self.conn = pymongo.MongoClient(
            'localhost',
            27017
        )
        db = self.conn['BBCnews']
        self.collection = db['news_tb']

    # storing the data of items into the mongoDB
    def process_item(self, item, spider):
        self.collection.insert(dict(item))
        return item
```

NewsscraperPipeline

9. The following two commands will start the web crawling and store the extracted data in the database. “cd NewsScraper” and “scrapy crawl spidy”.

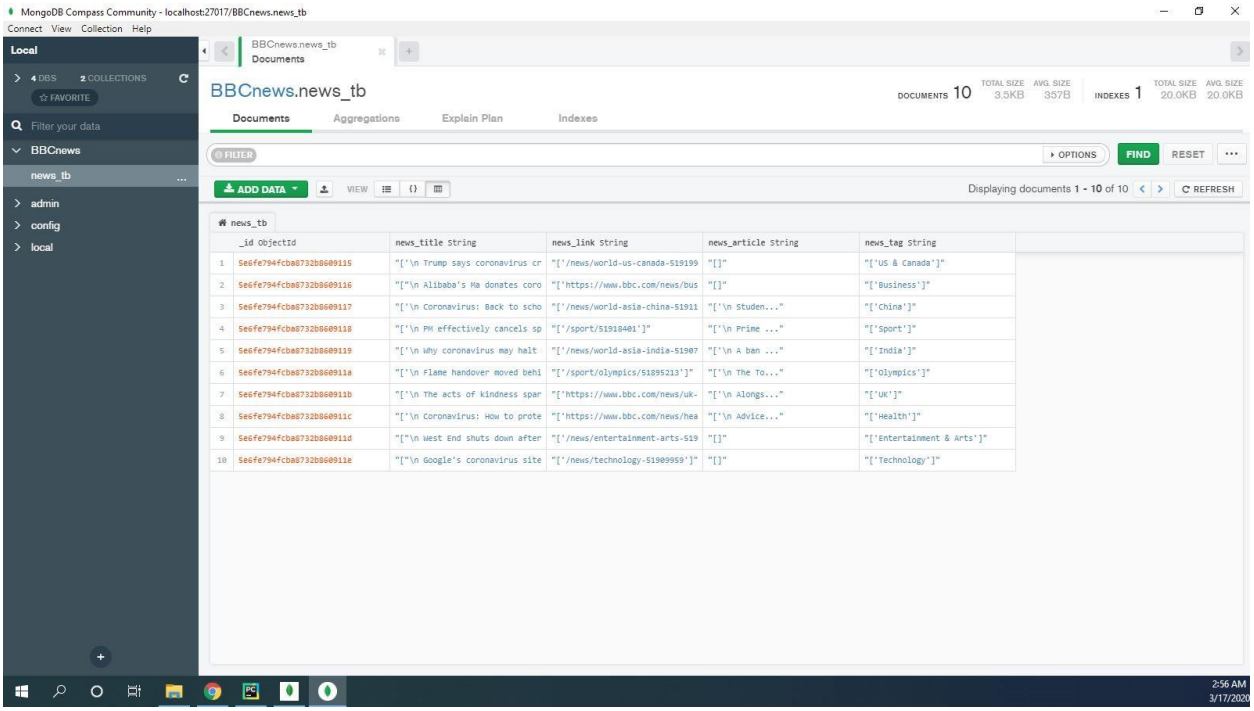
```
Terminal:  Local  +
Microsoft Windows [Version 10.0.18362.720]
(c) 2019 Microsoft Corporation. All rights reserved.

(venv) C:\Users\Reed\PycharmProjects\News Scraper>cd NewsScraper

(venv) C:\Users\Reed\PycharmProjects\News Scraper\NewsScraper>scrapy crawl spidy
Enter the website url:'https://www.bbc.com/'
```

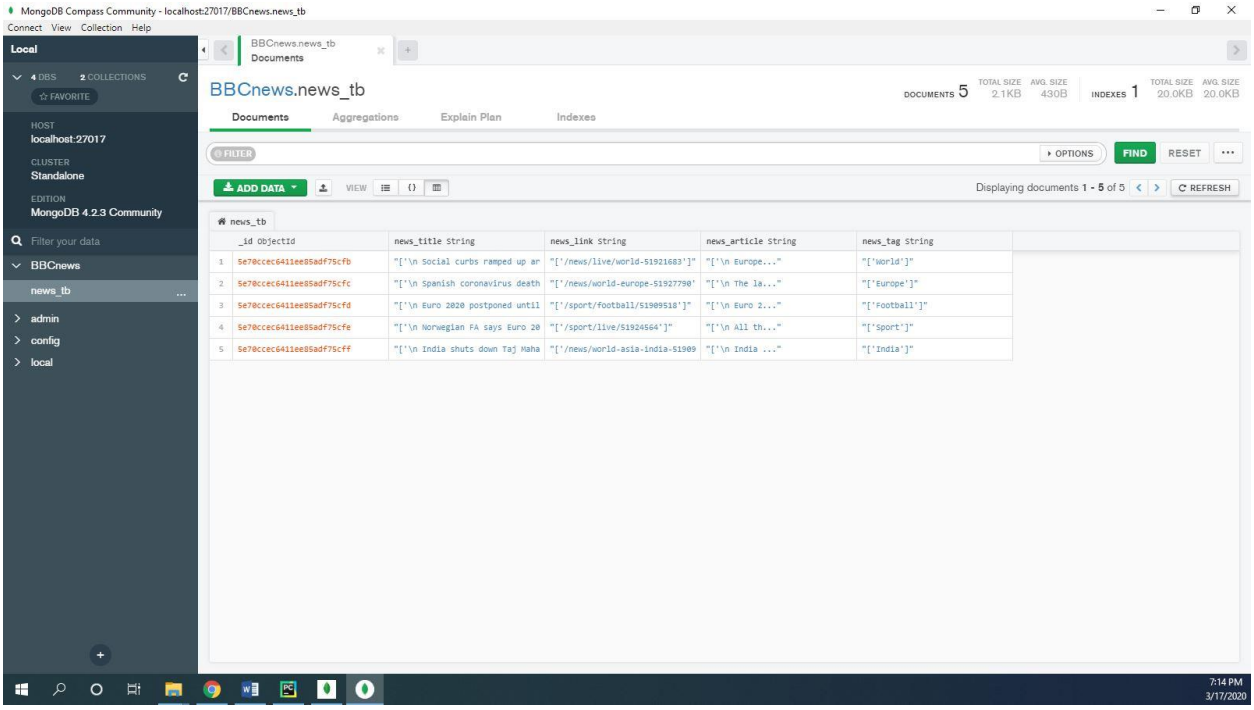
6: TODO Terminal Python Console

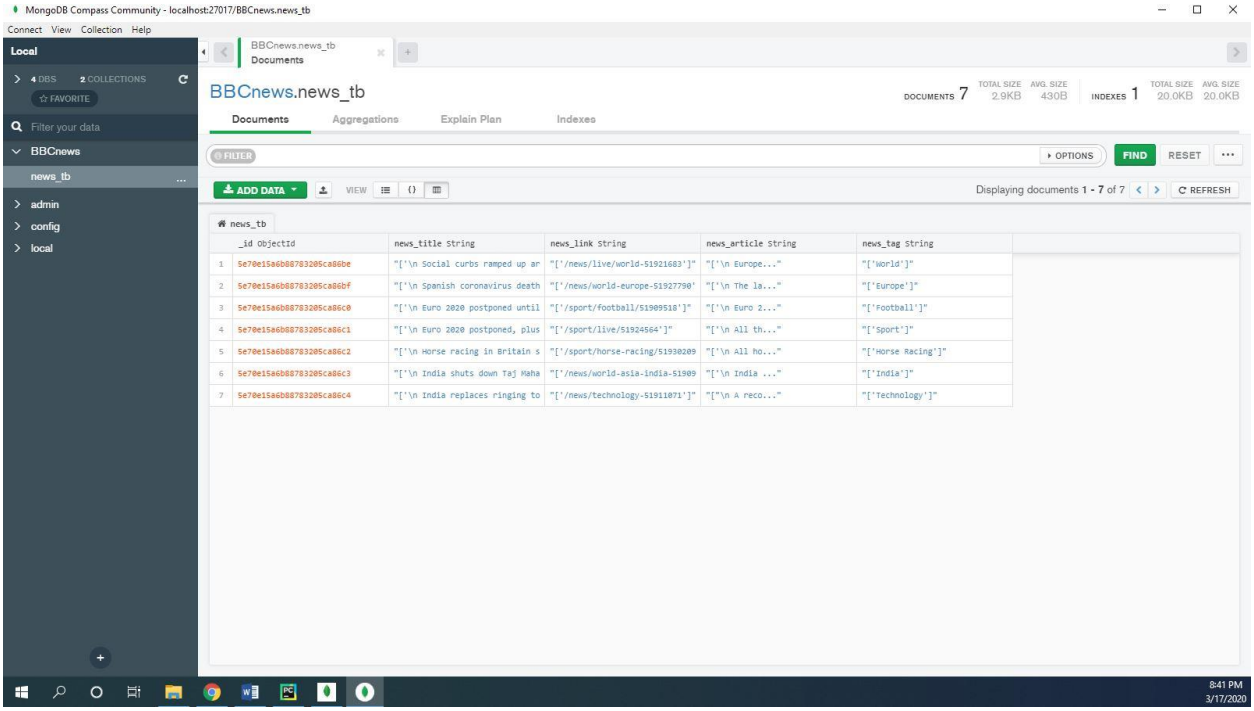
10. To create CSV and JSON file, you just need to run these to commands in the terminal. For JSON: “scrapy crawl spidy –o BBC_corona.json” and for CSV: “scrapy crawl spidy –o BBC_corona.csv”.

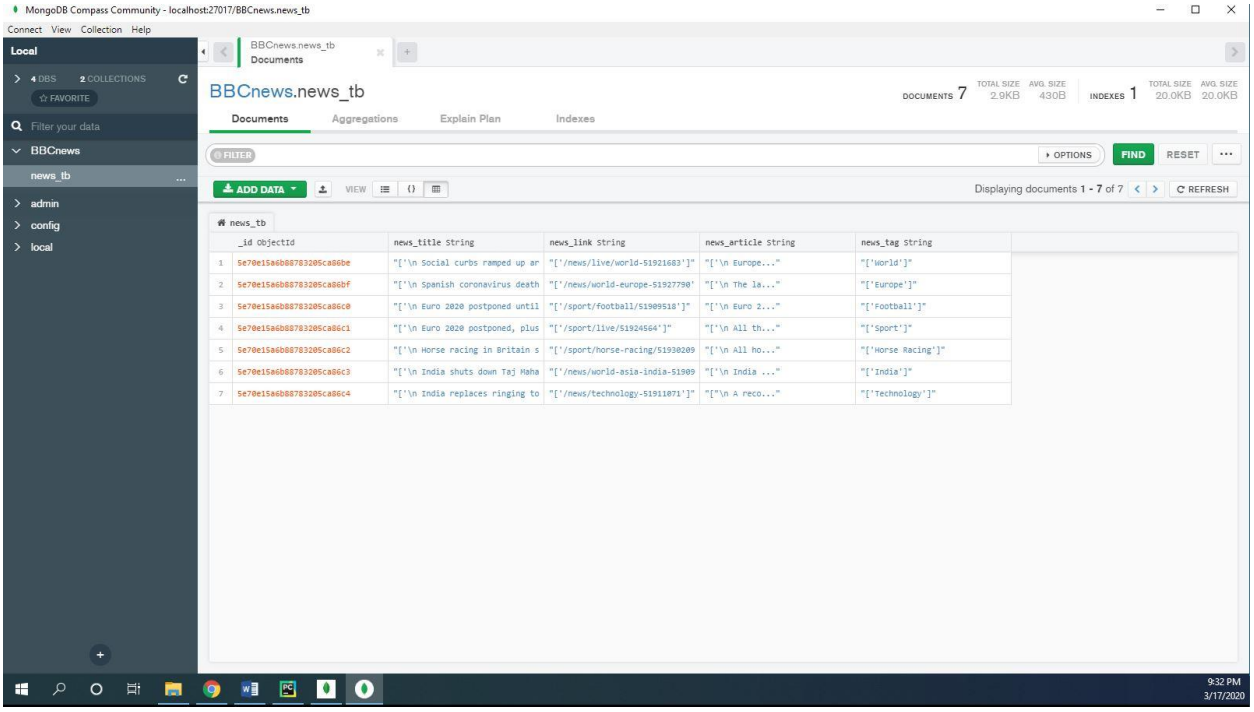
Results

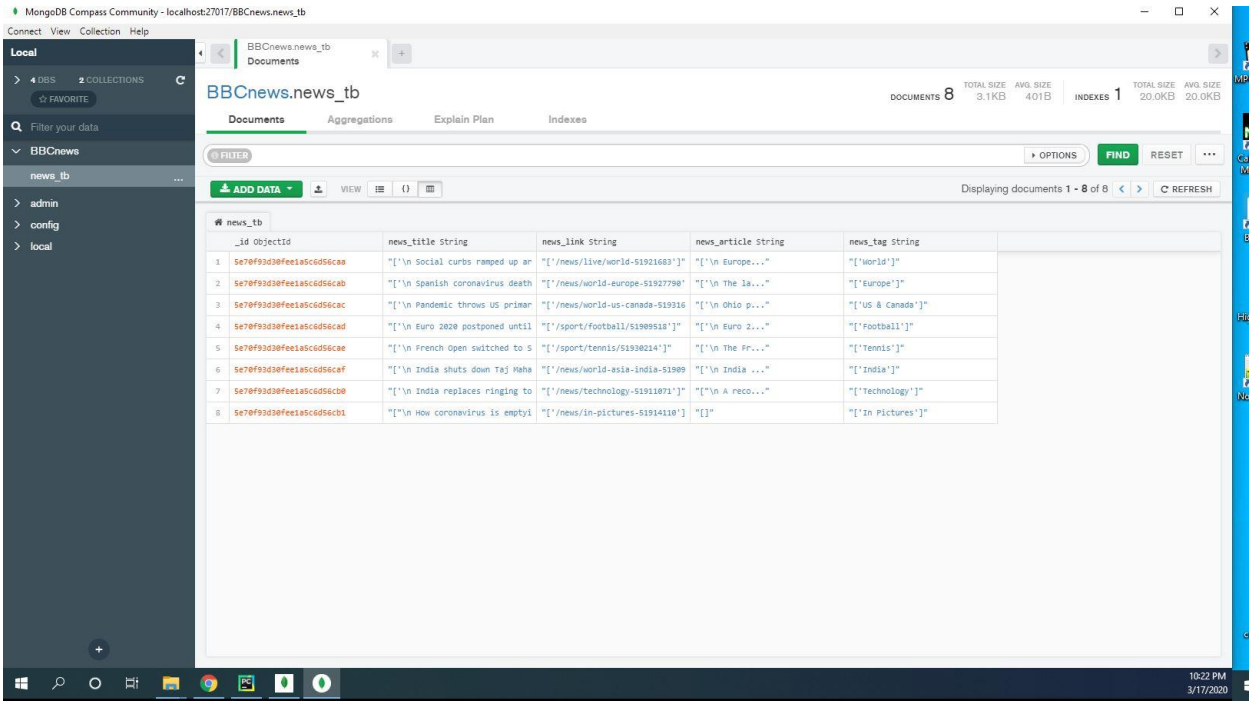
These are some screenshots of mongoDB containing the extracted data from the BBC website which were taken on different times while the news articles were updated on the website.






MongoDB Compass Community - localhost:27017/BBCnews.news_tb
Connect View Collection Help
Local
4 DBS 2 COLLECTIONS
FAVORITE
Filter your data
BBCnews
news_tb
admin
config
local
BBCnews.news_tb
Documents
BBCnews.news_tb
DOCUMENTS 7 TOTAL SIZE 2.9KB AVG. SIZE 430B
INDEXES 1 TOTAL SIZE 20.0KB AVG. SIZE 20.0KB
Documents Aggregations Explain Plan Indexes
FILTER
OPTIONS
FIND
RESET
ADD DATA
VIEW
Displaying documents 1 - 7 of 7
REFRESH
news_tb
_id ObjectId
news_title String
news_link String
news_article String
news_tag String
1 5e70e15a6b88783205ca86be "['\n Social curbs ramped up ar" "['/news/live/world-51921683']" "['\n Europe..." "['World']"
2 5e70e15a6b88783205ca86bf "['\n Spanish coronavirus death" "['/news/world-europe-51927790'" "['\n The la..." "['Europe']"
3 5e70e15a6b88783205ca86c0 "['\n Euro 2020 postponed until" "['/sport/football/51909518']" "['\n Euro 2..." "['Football']"
4 5e70e15a6b88783205ca86c1 "['\n Euro 2020 postponed, plus" "['/sport/live/51924564']" "['\n All th..." "['Sport']"
5 5e70e15a6b88783205ca86c2 "['\n Horse racing in Britain s" "['/sport/horse-racing/51930209" "['\n All ho..." "['Horse Racing']"
6 5e70e15a6b88783205ca86c3 "['\n India shuts down Taj Maha" "['/news/world-asia-india-51909" "['\n India ..." "['India']"
7 5e70e15a6b88783205ca86c4 "['\n India replaces ringing to" "['/news/technology-51911071']" "["\n A reco..." "['Technology']"
8:41 PM
3/17/2020


MongoDB Compass Community - localhost:27017/BBCnews.news_tb
Connect View Collection Help
Local
4 DBS 2 COLLECTIONS
FAVORITE
Filter your data
BBCnews
news_tb
admin
config
local
BBCnews.news_tb
Documents
BBCnews.news_tb
DOCUMENTS 7 TOTAL SIZE 2.9KB AVG. SIZE 430B
INDEXES 1 TOTAL SIZE 20.0KB AVG. SIZE 20.0KB
Documents Aggregations Explain Plan Indexes
FILTER
OPTIONS
FIND
RESET
ADD DATA
VIEW
Displaying documents 1 - 7 of 7
REFRESH
news_tb
_id ObjectId
news_title String
news_link String
news_article String
news_tag String
1 5e70e15a6b88783205ca86be "['\n Social curbs ramped up ar" "['/news/live/world-51921683']" "['\n Europe..." "['World']"
2 5e70e15a6b88783205ca86bf "['\n Spanish coronavirus death" "['/news/world-europe-51927790'" "['\n The la..." "['Europe']"
3 5e70e15a6b88783205ca86c0 "['\n Euro 2020 postponed until" "['/sport/football/51909518']" "['\n Euro 2..." "['Football']"
4 5e70e15a6b88783205ca86c1 "['\n Euro 2020 postponed, plus" "['/sport/live/51924564']" "['\n All th..." "['Sport']"
5 5e70e15a6b88783205ca86c2 "['\n Horse racing in Britain s" "['/sport/horse-racing/51930209" "['\n All ho..." "['Horse Racing']"
6 5e70e15a6b88783205ca86c3 "['\n India shuts down Taj Maha" "['/news/world-asia-india-51909" "['\n India ..." "['India']"
7 5e70e15a6b88783205ca86c4 "['\n India replaces ringing to" "['/news/technology-51911071']" "["\n A reco..." "['Technology']"
9:32 PM
3/17/2020



Here is the output of the CSV file:

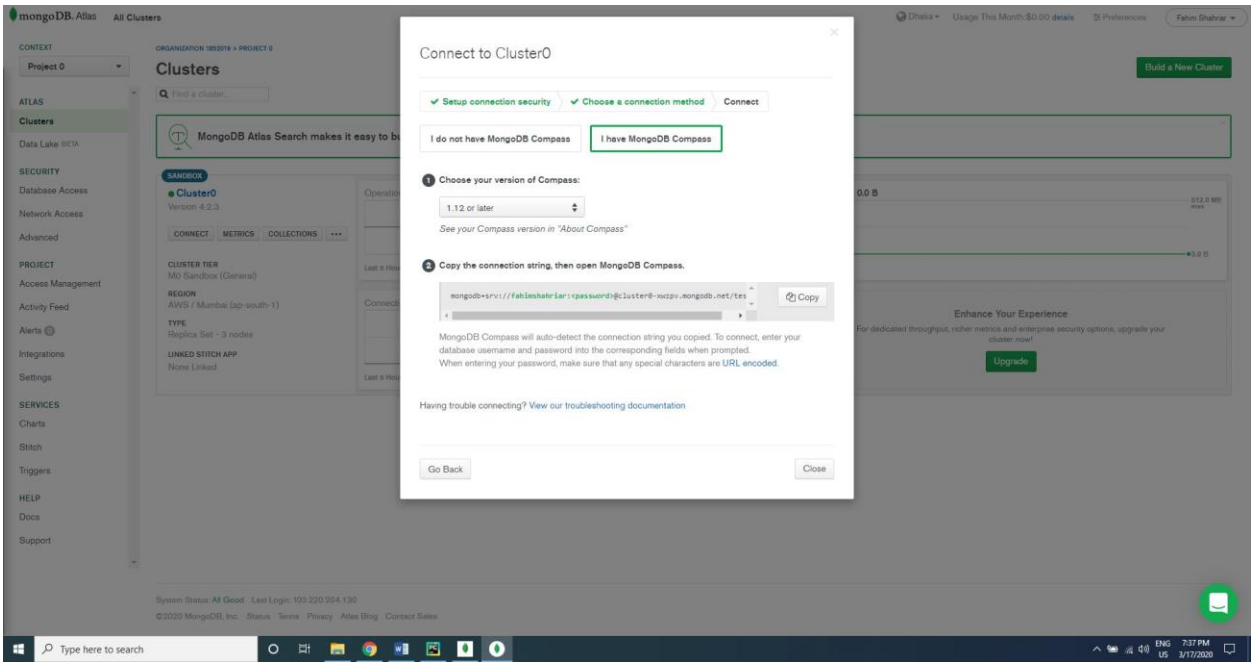
In [1]:
```python
import pandas as pd


# Importing the data
X = pd.read_csv("BBC_corona.csv")
df = pd.DataFrame(X)

#showing the csv file
df
```

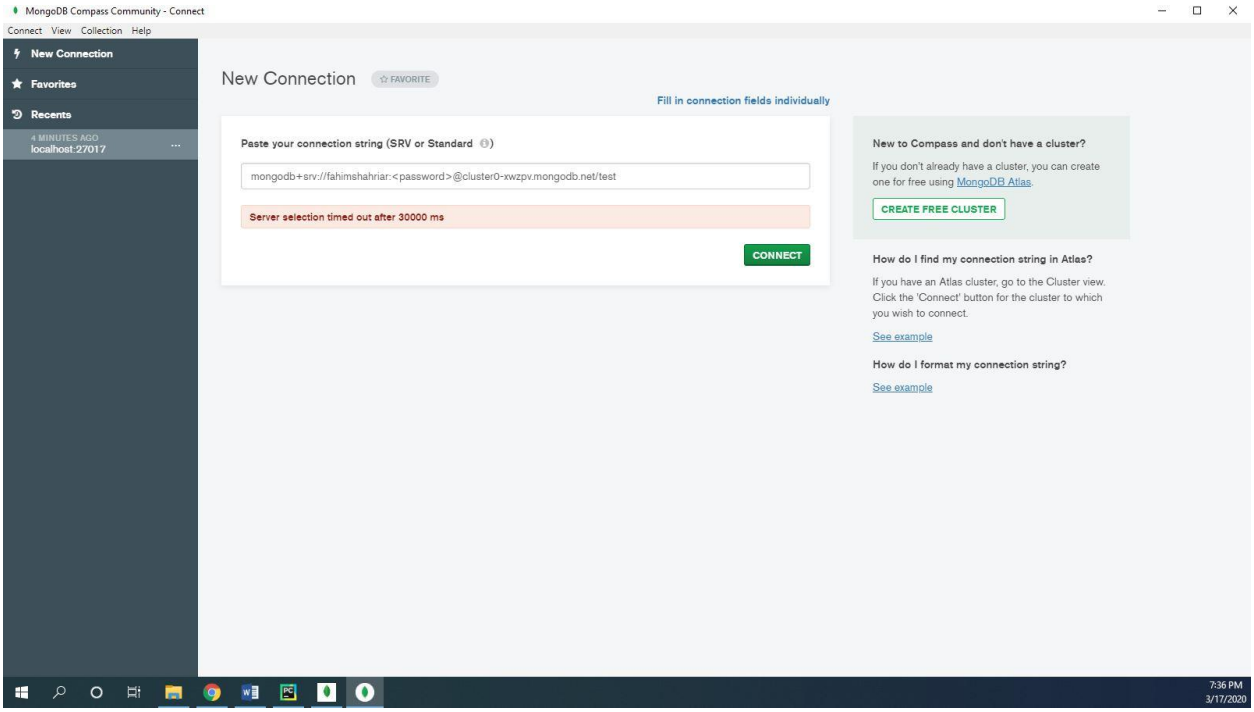
Out[1]:

	news_article	news_link	news_tag	news_title
0	[]	['/news/world-us-canada-51919945']	['US & Canada']	['\n Trump says coronavirus...
1	[]	['https://www.bbc.com/news/business-51904379']	['Business']	['\n Alibaba's Ma donates c...
2	['\n ...	['/news/world-asia-china-51911870']	['China']	['\n Coronavirus: Back to s...
3	['\n ...	['/sport/51918401']	['Sport']	['\n PM effectively cancels...
4	['\n ...	['/news/world-asia-india-51907173']	['India']	['\n Why coronavirus may ha...
5	['\n ...	['/sport/olympics/51895213']	['Olympics']	['\n Flame handover moved b...
6	['\n ...	['https://www.bbc.com/news/uk-51908023']	['UK']	['\n The acts of kindness s...
7	['\n ...	['https://www.bbc.com/news/health-51873799']	['Health']	['\n Coronavirus: How to pr...
8	[]	['/news/entertainment-arts-51906370']	['Entertainment & Arts']	['\n West End shuts down af...
9	[]	['/news/technology-51909959']	['Technology']	['\n Google's coronavirus s...

In [ ]:

I stored scraped data into my local database. I tried to store them on cloud so I used mongoDB Atlas. But I faced issues while connecting my computer with the server. Here are the screenshots:





So, that's all from my side for the given task. For any queries feel free to contact me.