

A.B.M. Fahim Shahriar

[abmfahimshahriar@iut-dhaka.com](mailto: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

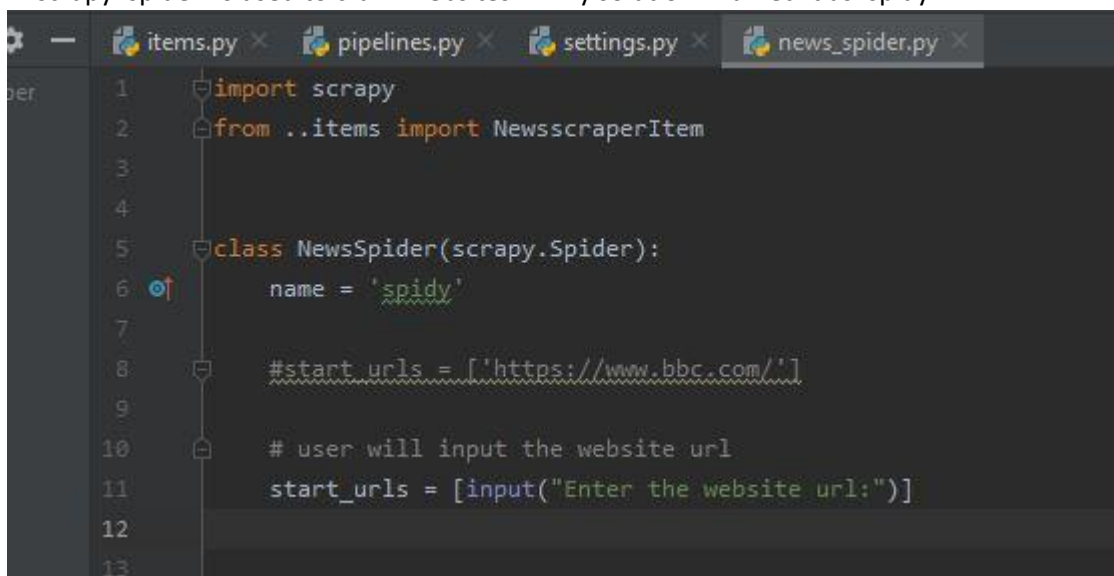
The extracted data was stored into mongoDB as well as CSV and JSON files were generated to show the results.

### **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’.



```
1  import scrapy
2  from ..items import NewsscrapecItem
3
4
5  class NewsSpider(scrapy.Spider):
6      name = 'spidy'
7
8      #start_urls = ['https://www.bbc.com/']
9
10     # user will input the website url
11     start_urls = [input("Enter the website url:")]
12
13
```

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.

```

13
14 # parser function that will crawl the website
15 def parse(self, response):
16
17     items = NewsscraferItem()
18
19     all_blocks = response.css("li.media-list__item")
20
21     for q in all_blocks:
22
23         # CSS selector to fetch specific data from website
24         news_title = str(q.css("a.block-link__overlay-link::text").extract())
25         news_link = str(q.css("a.block-link__overlay-link").xpath("@href").extract())
26         news_article = str(q.css("div.media__content").css("p.media__summary::text").extract())
27         news_tag = str(q.css("div.media__content").css("a.media__tag::text").extract())
28
29         if ('corona' or 'Corona' or 'virus' or 'Covid') in news_title:
30             # data will be stored in the items
31             items['news_title'] = news_title
32             items['news_link'] = news_link
33             items['news_article'] = news_article
34             items['news_tag'] = news_tag
35             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.

```

64
65 # Configure item pipelines
66 # See https://docs.scrapy.org/en/latest/topics/item-pipeline.html
67 ITEM_PIPELINES = {
68     'NewsScrafer.pipelines.NewsscraferPipeline': 300,
69 }

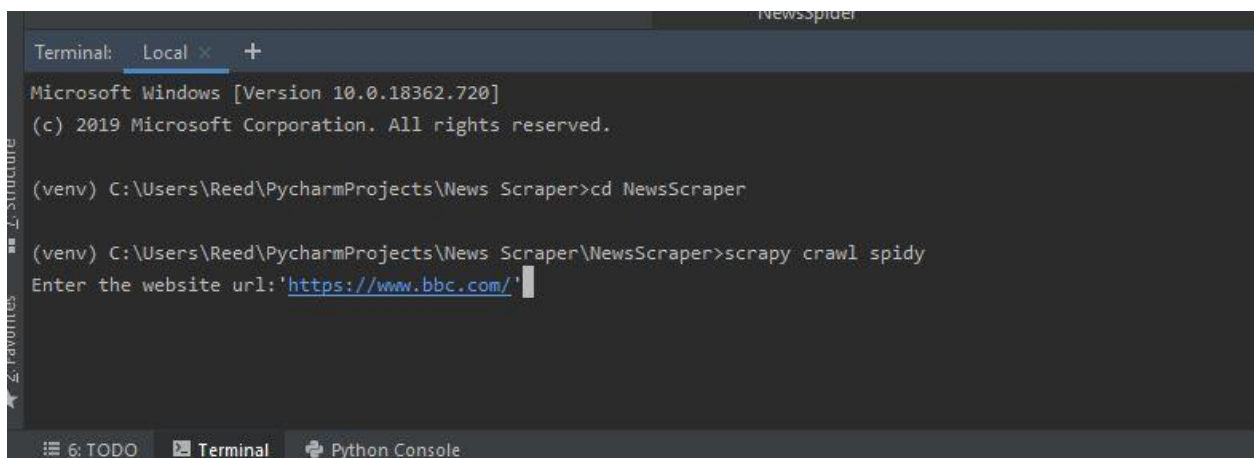
```

```

7
8  # import pymongo package for mongoDB
9  import pymongo
10
11  class NewsscraPerPipeline(object):
12      def __init__(self):
13          # creating connection with the database
14          self.conn = pymongo.MongoClient(
15              'localhost',
16              27017
17          )
18          db = self.conn['BBCnews']
19          self.collection = db['news_tb']
20
21      # storing the data of items into the mongoDB
22      def process_item(self, item, spider):
23          self.collection.insert(dict(item))
24          return item
25
26
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”.



The screenshot shows a terminal window with the following content:

```

Terminal: Local x +
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/'

```

10. To create CSV and JSON file, you just need to run these two 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

- DBS
- COLLECTIONS
- FAVORITE

Filter your data

BBCnews

- news\_tb
- admin
- config
- local

BBCnews.news\_tb

Documents Aggregations Explain Plan Indexes

DOCUMENTS 10 TOTAL SIZE 3.5KB AVG. SIZE 357B INDEXES 1 TOTAL SIZE 20.0KB AVG. SIZE 20.0KB

ADD DATA VIEW

Displaying documents 1 - 10 of 10

#	news_tb	_id ObjectId	news_title String	news_link String	news_article String	news_tag String
1	Se6fe794fcb07320609115	Se6fe794fcb07320609115	"['\n Trump says coronavirus cr	"['/news/world-us-canada-519195	"[']	"['US & Canada']"
2	Se6fe794fcb07320609116	Se6fe794fcb07320609116	"['\n Alibaba's Ma donates coro	"['https://www.bbc.com/news/bus	"[']	"['Business']"
3	Se6fe794fcb07320609117	Se6fe794fcb07320609117	"['\n Coronavirus: Back to scho	"['/news/world-asia-china-51911	"['\n Studen..."	"['China']"
4	Se6fe794fcb07320609118	Se6fe794fcb07320609118	"['\n PM effectively cancels sp	"['/sport/51918401']"	"['\n Prime ..."	"['Sport']"
5	Se6fe794fcb07320609119	Se6fe794fcb07320609119	"['\n why coronavirus may halt	"['/news/world-asia-india-51907	"['\n A ban ..."	"['India']"
6	Se6fe794fcb0732060911a	Se6fe794fcb0732060911a	"['\n Flame handover moved behi	"['/sport/olympics/5189521a']"	"['\n The To..."	"['Olympics']"
7	Se6fe794fcb0732060911b	Se6fe794fcb0732060911b	"['\n the acts of kindness spar	"['https://www.bbc.com/news/uk-	"['\n Along..."	"['UK']"
8	Se6fe794fcb0732060911c	Se6fe794fcb0732060911c	"['\n coronavirus: How to prote	"['https://www.bbc.com/news/hea	"['\n Advice..."	"['Health']"
9	Se6fe794fcb0732060911d	Se6fe794fcb0732060911d	"['\n West end shuts down after	"['/news/entertainment-arts-519	"[']	"['Entertainment & Arts']"
10	Se6fe794fcb0732060911e	Se6fe794fcb0732060911e	"['\n Google's coronavirus site	"['/news/technology-51909959']"	"[']	"['Technology']"

2:56 AM 3/17/2020

MongoDB Compass Community - localhost:27017/BBCnews.news\_tb

Connect View Collection Help

Local

- DBS
- COLLECTIONS
- FAVORITE

Filter your data

BBCnews

- news\_tb
- admin
- config
- local

BBCnews.news\_tb

Documents Aggregations Explain Plan Indexes

DOCUMENTS 5 TOTAL SIZE 2.1KB AVG. SIZE 430B INDEXES 1 TOTAL SIZE 20.0KB AVG. SIZE 20.0KB

ADD DATA VIEW

Displaying documents 1 - 5 of 5

#	news_tb	_id ObjectId	news_title String	news_link String	news_article String	news_tag String
1	Se70ccec6411ee85ad75cfd	Se70ccec6411ee85ad75cfd	"['\n Social curbs ramped up ar	"['/news/live/world-51921683']"	"['\n Europe..."	"['World']"
2	Se70ccec6411ee85ad75cfc	Se70ccec6411ee85ad75cfc	"['\n Spanish coronavirus death	"['/news/world-europe-51927798'	"['\n The 1a..."	"['Europe']"
3	Se70ccec6411ee85ad75cfd	Se70ccec6411ee85ad75cfd	"['\n Euro 2020 postponed until	"['/sport/football/51909518']"	"['\n Euro 2..."	"['Football']"
4	Se70ccec6411ee85ad75cfe	Se70ccec6411ee85ad75cfe	"['\n Norwegian FA says Euro 20	"['/sport/live/51924564']"	"['\n All th..."	"['Sport']"
5	Se70ccec6411ee85ad75cff	Se70ccec6411ee85ad75cff	"['\n India shuts down Taj Maha	"['/news/world-asia-india-51909	"['\n India ..."	"['India']"

7:14 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 7 TOTAL SIZE 2.9KB AVG. SIZE 430B INDEXES 1 TOTAL SIZE 20.0KB AVG. SIZE 20.0KB

Documents Aggregations Explain Plan Indexes

FILTER

ADD DATA VIEW

Displaying documents 1 - 7 of 7

#	news_tb	_id Objectid	news_title String	news_link String	news_article String	news_tag String
1	5e70e15a60b0783205ca80de		"['\n Social curbs ramped up ar	"['/news/live/world-51921683']"	"['\n Europe..."	"['world']"
2	5e70e15a60b0783205ca80df		"['\n Spanish coronavirus death	"['/news/world-europe-51927798']"	"['\n The 1a..."	"['Europe']"
3	5e70e15a60b0783205ca80e0		"['\n Euro 2020 postponed until	"['/sport/football/51909518']"	"['\n Euro 2..."	"['Football']"
4	5e70e15a60b0783205ca80e1		"['\n Euro 2020 postponed, plus	"['/sport/live/51924564']"	"['\n All th..."	"['sport']"
5	5e70e15a60b0783205ca80e2		"['\n Horse racing in Britain s	"['/sport/horse-racing/51930209']"	"['\n All ho..."	"['Horse Racing']"
6	5e70e15a60b0783205ca80e3		"['\n India shuts down Taj Maha	"['/news/world-asia-india-51909']"	"['\n India ..."	"['India']"
7	5e70e15a60b0783205ca80e4		"['\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 7 TOTAL SIZE 2.9KB AVG. SIZE 430B INDEXES 1 TOTAL SIZE 20.0KB AVG. SIZE 20.0KB

Documents Aggregations Explain Plan Indexes

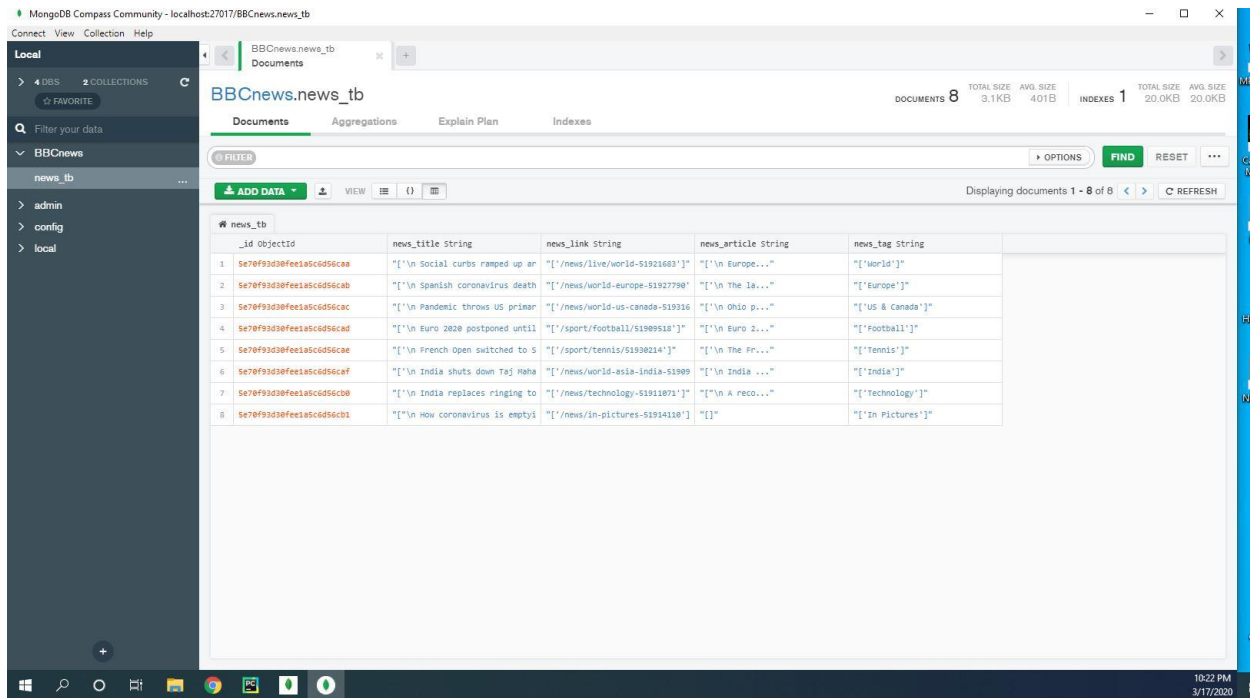
FILTER

ADD DATA VIEW

Displaying documents 1 - 7 of 7

#	news_tb	_id Objectid	news_title String	news_link String	news_article String	news_tag String
1	5e70e15a60b0783205ca80de		"['\n Social curbs ramped up ar	"['/news/live/world-51921683']"	"['\n Europe..."	"['world']"
2	5e70e15a60b0783205ca80df		"['\n Spanish coronavirus death	"['/news/world-europe-51927798']"	"['\n The 1a..."	"['Europe']"
3	5e70e15a60b0783205ca80e0		"['\n Euro 2020 postponed until	"['/sport/football/51909518']"	"['\n Euro 2..."	"['Football']"
4	5e70e15a60b0783205ca80e1		"['\n Euro 2020 postponed, plus	"['/sport/live/51924564']"	"['\n All th..."	"['sport']"
5	5e70e15a60b0783205ca80e2		"['\n Horse racing in Britain s	"['/sport/horse-racing/51930209']"	"['\n All ho..."	"['Horse Racing']"
6	5e70e15a60b0783205ca80e3		"['\n India shuts down Taj Maha	"['/news/world-asia-india-51909']"	"['\n India ..."	"['India']"
7	5e70e15a60b0783205ca80e4		"['\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]: import pandas as pd

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

#showing the csv file
df

```

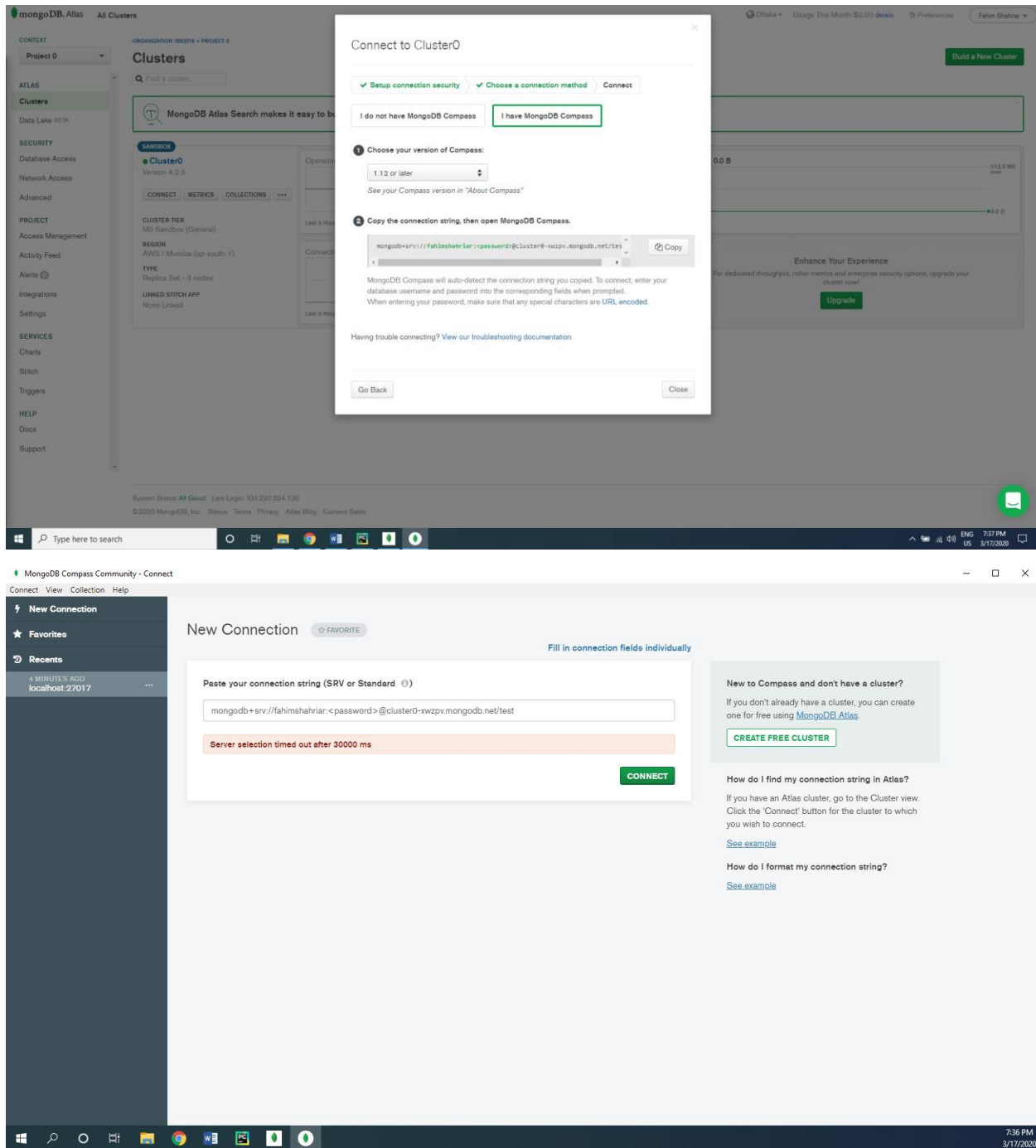
Out[1]:

	news_article	news_link	news_tag	news_title
0	['\n ...	['/news/world-us-canada-51919945']	['US & Canada']	"['\n Trump says coronavirus...
1	['\n ...	['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	['\n ...	['/news/entertainment-arts-51906370']	['Entertainment & Arts']	"['\n West End shuts down af...
9	['\n ...	['/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:





The issue that I mentioned earlier about connecting to the cloud is resolved. I have connected mongoDB compass with the mongoDB Atlas. I tried connecting with the compass manually and successfully did it. Here are some screenshots.

The screenshot shows the MongoDB Atlas web interface. The top navigation bar includes the MongoDB logo, the cluster name 'NewsScraper', and the version '4.2.3'. The left sidebar contains a navigation menu with options like 'Context', 'Atlas', 'Security', 'Project', 'Services', and 'Help'. The main content area displays the 'Collections' tab for the 'BBCnews' database. It shows a table with columns: Collection Name, Documents, Documents Size, Documents Avg, Indexes, Index Size, and Index Avg. The table lists one collection, 'news\_tb', with 6 documents, a size of 2.12KB, and an average document size of 362B. There is also a 'CREATE COLLECTION' button.

Collection Name	Documents	Documents Size	Documents Avg	Indexes	Index Size	Index Avg
news_tb	6	2.12KB	362B	1	20KB	20KB

The screenshot shows the MongoDB Compass desktop application. The left sidebar displays the 'Local' connection and the 'BBCnews' database. The main content area shows the 'Documents' tab for the 'BBCnews.news\_tb' collection. It displays a table with 6 documents, each with a unique '\_id' and a 'news\_title' field. The table also includes columns for 'news\_link', 'news\_article', and 'news\_tag'. The status bar at the bottom indicates 'Displaying documents 1 - 6 of 6'.

#	news_tb	_id	news_title	news_link	news_article	news_tag
1	5e72434f5940dfc376b4a8fa	5e72434f5940dfc376b4a8fa	"['\n Global coronavirus cases	"['\n School..."	"['\n World..."	"['\n World..."
2	5e7243528940dfc376b4a8fb	5e7243528940dfc376b4a8fb	"['\n Stocks slide despite coro	"['\n The FT..."	"['\n Business..."	"['\n Business..."
3	5e7243528940dfc376b4a8fc	5e7243528940dfc376b4a8fc	"['\n EU entry ban hits travel	"['\n A 30-d..."	"['\n Europe..."	"['\n Europe..."
4	5e7243538940dfc376b4a8fd	5e7243538940dfc376b4a8fd	"['\n Virus could shake up worl	"['\n The re..."	"['\n Business..."	"['\n Business..."
5	5e7243538940dfc376b4a8fe	5e7243538940dfc376b4a8fe	"['\n Eurovision Song Contest c	"['\n Entertainment & Arts..."	"['\n Entertainment & Arts..."	"['\n Entertainment & Arts..."
6	5e7243538940dfc376b4a8ff	5e7243538940dfc376b4a8ff	"['\n How coronavirus is empty	"['\n In Pictures..."	"['\n In Pictures..."	"['\n In Pictures..."

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