Smart Web Investigation Framework

G.Mallikarjuna Rao[1], B.Ramakrishna Reddy[2] and and P.Vishnu[3]

[1] Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

[2] Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

[3]Gokaraju Rangaraju Institute of Engineering and Technology, Bachpally, Hyderabad, India

gmr_333@yahoo.com, brkgriet@yahoo.com, vishnupeesapati@gmail.com

Abstract. In this paper we have developed a frame work for web scraping and text analysis. Most of the real world applications require text analysis and web scraping, however due their disjoint behaviour there is no unified frame work which will take care about both needs. In the proposed scalable, portable frame work, centralized toolkit is designed to scraping and text analysis. Further it allow analysis of large data in one go and provide inferences about change in web-based text over the time. The frame work will give inferences about web-based topics as well summarize the data in a crisp format which will make it easy for the user to analyse the text-based data of their choosing from the internet. Proposed frame work is very much useful when analyzing web based reviews of various products, entertainment and the news about a specific topic in www.

Keywords: : Lemmatization and Stemming, TF-IDF Score, Latent Semantic Analysis.

.

1 Introduction

Huge amounts of data are posted on the World Wide Web related to socio, technical or entertainment. However trustworthiness of this data is always questionable [5][6]. The customer is in a better position to judge the correctness of interested topic if they get more elaborative facts on a specific topic. The problem here is how to get them as the information showed by most sites must be seen utilizing an internet browser. They don't give the usefulness of sparing a duplicate of this information for individual use. At that point, the main decision is to physically reorder the information which is an exhausting activity that can take a few hours to finish or even days. Web scratching will be handy here, it is robotic based technique, hence avoid physically duplicating the information from the respective sites.

1.1 Web Scraping

Web Scraping is a method used to extract large amount of data from the websites. This extracted data from the websites is usually stored in a local file in your personal computer (or) in a table format in a database [3][4]. This process is performed with the help of a bot (or) a web crawler. Some of the techniques used are:

- **Copy and Paste Manually:** Most of this activity can be automated except the cases where webpages set-up blocks to prevent the web crawlers from scrapping the page. Here only alternative is human involvement.

- **Comparing the Text Pattern:** The UNIX grep command, matching regular expressions in programming languages like Perl or Python can be a robust idea to scrap the information from the web pages.

- **Socket Programming:** Using HTTP programming, we can retrieve data from static and dynamic webpages by posting HTTP requests to the remote web server.

- **HTML parsing:** Wrapper allow the collection of the bulk data from dynamic/static web sites and translate them into relational from. Wrapper program assumes that the input from web pages can be associated with URL schemes which can be used to parse HTML pages and then extract and transform the page content.

- **Document Object Model parsing:** Web browser, like the Safari (or) the Google Chrome, we can extract the information which is dynamically generated by the client-side scripts. These browsers also parse the web pages into a DOM tree which help us to extract information from parts of the pages too.

- **Vertical percentage aggregation:** Many companies have devised vertical harvesting platforms. This platform is used to automate the creation of bots with the available knowledge base. The quality of the platform is calculated with the help of the information it will get and the number of the sites it extracted.

- **Recognizing semantic annotation:** The scraped pages can also contain metadata or semantic mark-ups and annotations, which can be used to discover unique snippets of statistics.

- **Computer vision web-page analysis:** There are efforts using machine gaining knowledge of and laptop imaginative and prescient that try to become aware of and extract facts from net pages by deciphering pages visually as an individual might.

1.2 Text Analysis

Data recovery, lexical appraisal to check word recurrence conveyances, test acknowledgment, labelling/comment, measurement extraction, information mining techniques which incorporate connection and affiliation assessment, representation, and perceptive examination are associated with content examination[1][2]. Content examination ordinarily incorporates the way toward organizing the info's literary substance (typically sifting, including some inferred phonetic highlights, evacuating others, and resulting consolidation into a database), determining designs inside existing information, lastly separating and deciphering the subtleties. In-Text audit, 'High calibre' typically alludes to a couple of blends of significance, oddity, and intrigue.

1.3 Natural Language Processing

Content pre-handling is customarily an indispensable advance for common language preparing (NLP) undertakings [7]. It changes content into a more noteworthy edible structure with the goal that device considering estimations can complete better. In NLP, textual content pre-processing is the primary step within the process of building a model.

The various text pre-processing steps are:

- Regular expressions
- Stop words removal
- Stemming
- Lemmatization
- Tokenization of text

1.4 Text Classification

This can be performed either manually (accomplished with a human agent's effort reading and categorizing texts) or automatically (including machine learning techniques and algorithms to identify the texts more easily and cost-effectively). Companies receive text facts all the time. Be it emails, chats, social media comments, support tickets, or NPS responses, all these texts are very wealthy sources of information. However, they're not structured, so you need to tag and examine those texts before you can make sense of them and acquire insights.

2. Proposed Frame Work

2.1 Pre-processing

The retrieved HTML document contains many HTML tags. Using regular language expressions, the HTML document is cleaned of all tags. The content that is meant to be analyzed from the web-site is generally present in tags or a variation of or tags. The necessary text is then extracted from those tags. This text is then pre-processed through removal of stop words, converting all words to lowercase, lemmatization and stemming. The stops words are generally considered to be conjunctions, prepositions and determiners. These words are removed. Then lemmatization is done in order to trace the word to its root by comparing the word with a dictionary and then removing its inflectional endings. The stemming is done in order to remove all the word's suffixes and prefixes. This results in only leaving the important parts of the text in the document.

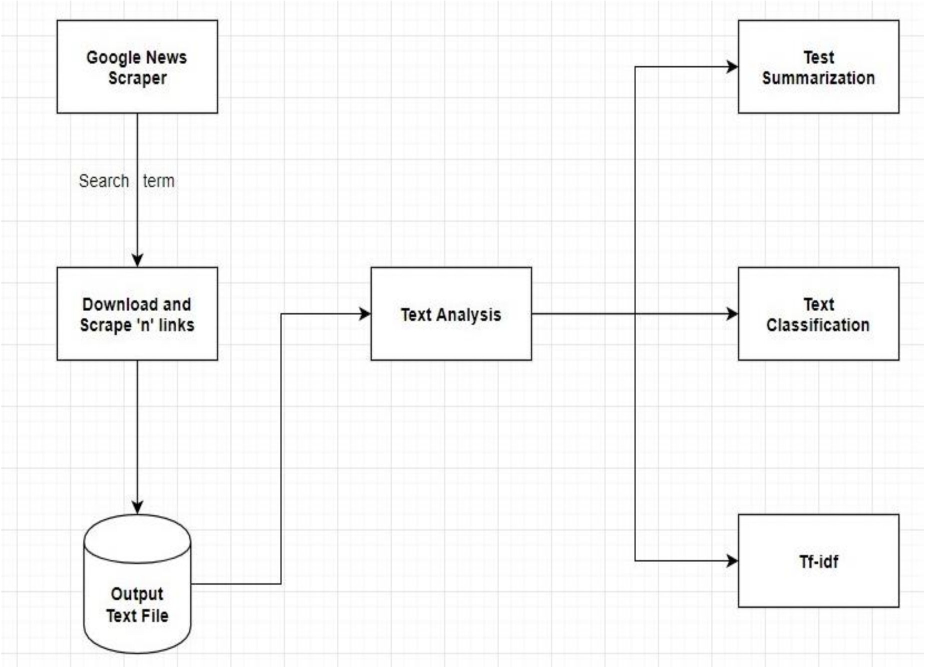


Fig. 1. Proposed Frame Work

2.2 Working of the Web-Scraper

The web-scraper is written completely in C. We have written a web-scraping library in C completely from scratch. The library uses Linux socket libraries and the open source Open SSL library. The web-scraper takes the web-site name as an input It is then sends a DNS request to retrieve the IP address of the server of the website. The scraper initiates a TCP connection with the server of the website. The SSL library is used create a secure SSL connection on top of the TCP connection. The request is sent to the server therefore downloading the entire HTML file of the entire website. The SSL connection is then closed and the TCP connection is closed next. Figure 2 denote web-scraping process.

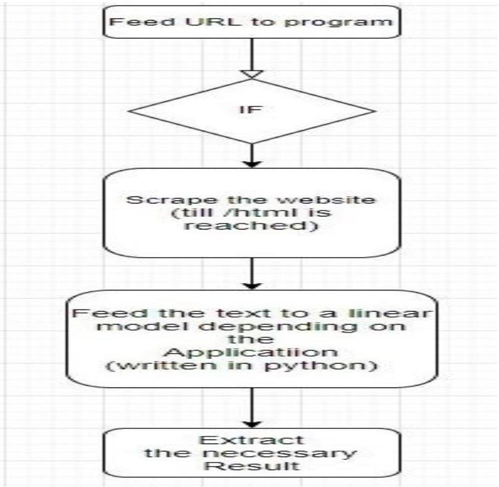

Feed URL to program
IF
Scrape the website (till /html is reached)
Feed the text to a linear model depending on the Applicatiion (written in python)
Extract the necessary Result

Fig. 2. Web Scraping

2.3 Text Summarization

The project uses extractive text summary techniques to summarize the document. The flow of the summarization is shown in Figure 3.

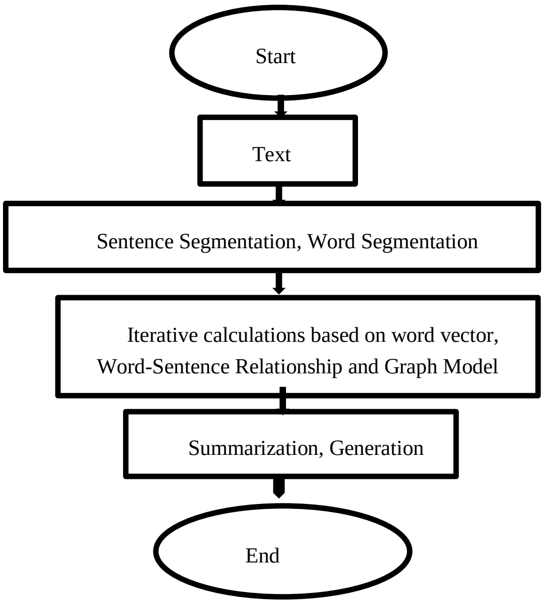

Start
Text
Sentence Segmentation, Word Segmentation
Iterative calculations based on word vector, Word-Sentence Relationship and Graph Model
Summarization, Generation
End

Fig. 3. Process Flow

The document is cleaned and pre-processed using regular language expressions which denote the HTML tags as well as punctuation marks. The document is divided into multiple sentences and is then stored in a list of lists format in the program. The sentences are divided into word-based tokens. The term frequencies of each word are stored in a dictionary with the word being the key and its term frequency being the value. This dictionary of words and its term-frequency is compared to the sentences within the scraped text. Each sentence is then scored by calculating the sum of term-frequency of the words it holds. Then the respective sentences and its words are stored in a dictionary with the sentence being the keys and their scores being values of the dictionary. An issue with this approach is that longer sentences tend to have bigger scores. To deal with this problem, the lengths of the sentences are fitted to a normal distribution. Then over the normal distribution only a range of few numbers is considered for summarizing. Depending on the number of words and sentences you want the summarized text to hold, the sentences in the dictionary are selected and the outputted together. This makes up the text summary module.

2.4 Sentiment Analysis

The document is cleaned and pre-processed using regular language expressions which denote the HTML tags as well as punctuation marks. The stop words are then removed from the document. The words in the document is the put through lemmatization. Using the one-hot encoding the document is converted into a vector. We have previously created a dataset using news articles and trained on supporting and criticizing news articles and reviews. We have trained a linear regression model on the dataset after one-hot encoding the dataset. The Figure 3, Flowchart of the process of summarization cleaned and pre-processed document which was retrieved through the scraper is then fed to the model which will then classify the document as being supportive or criticizing.

2.5 TF-IDF Scoring:

The document is cleaned and pre-processed using regular language expressions which denote the HTML tags as well as punctuation marks. The stop words are generally considered to be determiners, conjunctions, prepositions. These occur highly in any kind of document and hold a high term frequency score but they are not necessary to the document. These are then removed from the document. The words in the document is the put through lemmatization. The lemmatization function then scans each term and traces it to the root term of the word effectively removing prefixes and suffixes. The term-frequency of each of term is calculated. Depending on the term-frequency of each word in the document and the number of documents retrieved by the scraper, the TF-IDF score is calculated and displayed using

$$W_{i,j} = tf_{i,j} * \log\left({}^{N}/_{df_i}\right) \quad ------1$$

Where $tf_{i,j}$ denote the number of occurrences of i in j, df_i is the number of documents containing i, and N is the total number of documents.

2.6 **Topic Modelling**:

Topic Modelling tries to push the prominent words present in a the corpus into corresponding topics. We are achieving this using Latent Semantic Analysis or LSA for short. It fits all the words in the document after pre-processing into a Bag of Words Model. After converting the text into a bag of words model it then applies matrix decomposition on the bag of words model. We used Singular Value Decomposition to breakdown the bag of words matrix and extracted the corresponding topic - words matrix. We then chose the word with the highest term frequency in the topic as the name of the topic.

3. Experimentation

Python technology: A translator is a type of programming program that executes elective projects. At the point when we compose the Python programs, it changes over source code composed with the guide of the engineer into the middle language that is fairly converted into the local language/contraption language with the expectation to be executed [8]. Figure 5 denotes compilation and interpreting process

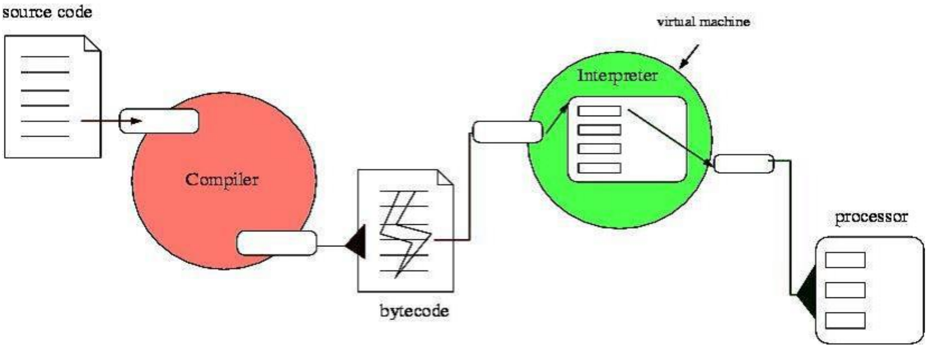
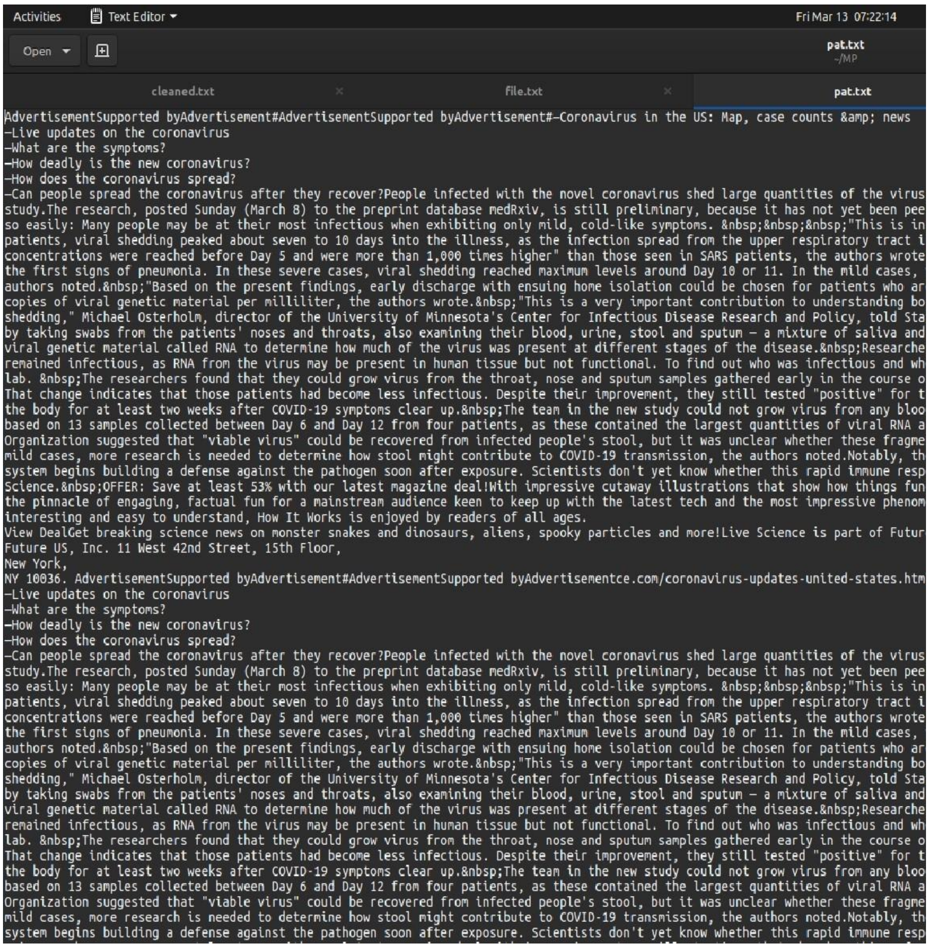


Fig. 4. Compiling and interpreting source code

Experimentation is performed using corei3 processor of 3Ghz using Anaconda (Python 3.8), Linux Shell/Windows Command-Prompt and python-requests, pip version 20.0.3

4. **Results and Conclusions:**

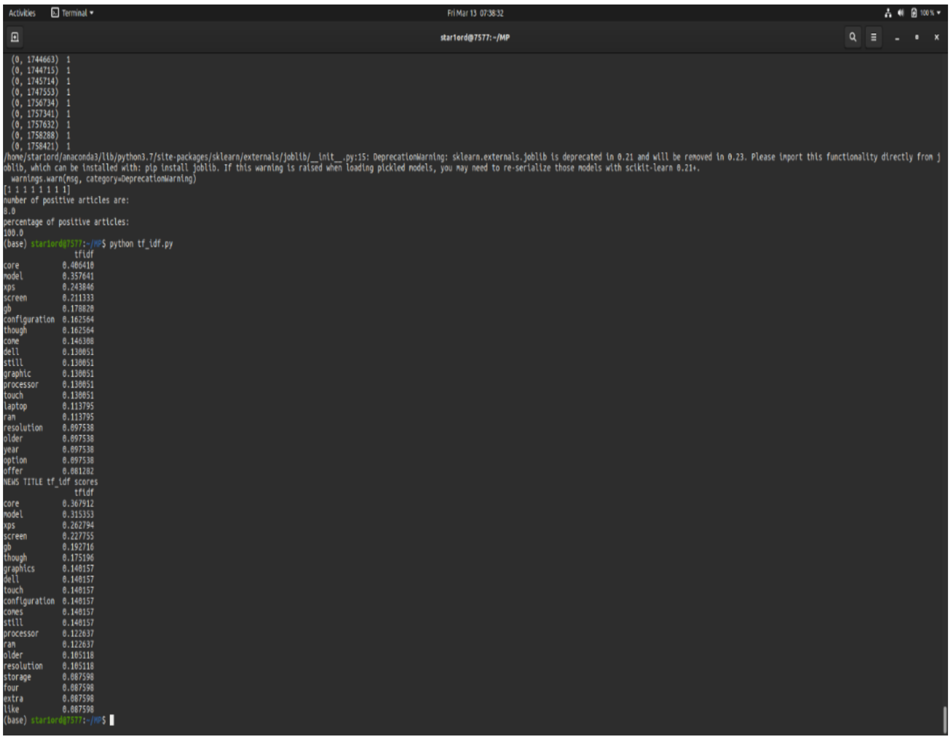
Web Scraping is achieved by connecting to the server of the website, establishing SSL connection and downloading the web page. Sub-string search algorithms are used to isolate intended elements. Preliminary analysis is required before scraping the website. Depending on the application, the model must be trained on the required datasets to perform specific analysis such as sentiment analysis, cosine similarity, etc. For sentiment analysis we are using linear classifiers such as linear regression. Using the tools, we have written, we have constructed a Google news scraper to analyze news reports [8][9][10]. The architecture designed for the Google news scraper using tools from our framework is represented in Figure 5.



```
AdvertisementSupported byAdvertisement#AdvertisementSupported byAdvertisement#—Coronavirus in the US: Map, case counts &amp; news
—Live updates on the coronavirus
—What are the symptoms?
—How deadly is the new coronavirus?
—How does the coronavirus spread?
—Can people spread the coronavirus after they recover?People infected with the novel coronavirus shed large quantities of the virus
study.The research, posted Sunday (March 8) to the preprint database medRxiv, is still preliminary, because it has not yet been pee
so easily: Many people may be at their most infectious when exhibiting only mild, cold-like symptoms. &nbsp;&nbsp;&nbsp;"This is in
patients, viral shedding peaked about seven to 10 days into the illness, as the infection spread from the upper respiratory tract i
concentrations were reached before Day 5 and were more than 1,000 times higher" than those seen in SARS patients, the authors wrote
the first signs of pneumonia. In these severe cases, viral shedding reached maximum levels around Day 10 or 11. In the mild cases,
authors noted.&nbsp;"Based on the present findings, early discharge with ensuing home isolation could be chosen for patients who ar
copies of viral genetic material per milliliter, the authors wrote.&nbsp;"This is a very important contribution to understanding bo
shedding," Michael Osterholm, director of the University of Minnesota's Center for Infectious Disease Research and Policy, told Sta
by taking swabs from the patients' noses and throats, also examining their blood, urine, stool and sputum — a mixture of saliva and
viral genetic material called RNA to determine how much of the virus was present at different stages of the disease.&nbsp;Researche
remained infectious, as RNA from the virus may be present in human tissue but not functional. To find out who was infectious and wh
lab. &nbsp;The researchers found that they could grow virus from the throat, nose and sputum samples gathered early in the course o
That change indicates that those patients had become less infectious. Despite their improvement, they still tested "positive" for t
the body for at least two weeks after COVID-19 symptoms clear up.&nbsp;The team in the new study could not grow virus from any bloo
based on 13 samples collected between Day 6 and Day 12 from four patients, as these contained the largest quantities of viral RNA a
Organization suggested that "viable virus" could be recovered from infected people's stool, but it was unclear whether these fragme
mild cases, more research is needed to determine how stool might contribute to COVID-19 transmission, the authors noted.Notably, th
system begins building a defense against the pathogen soon after exposure. Scientists don't yet know whether this rapid immune resp
Science.&nbsp;OFFER: Save at least 53% with our latest magazine deal!With impressive cutaway illustrations that show how things fun
the pinnacle of engaging, factual fun for a mainstream audience keen to keep up with the latest tech and the most impressive phenom
interesting and easy to understand, How It Works is enjoyed by readers of all ages.
View DealGet breaking science news on monster snakes and dinosaurs, aliens, spooky particles and more!Live Science is part of Futur
Future US, Inc. 11 West 42nd Street, 15th Floor,
New York,
NY 10036. AdvertisementSupported byAdvertisement#AdvertisementSupported byAdvertisementce.com/coronavirus-updates-united-states.htm
—Live updates on the coronavirus
—What are the symptoms?
—How deadly is the new coronavirus?
—How does the coronavirus spread?
—Can people spread the coronavirus after they recover?People infected with the novel coronavirus shed large quantities of the virus
study.The research, posted Sunday (March 8) to the preprint database medRxiv, is still preliminary, because it has not yet been pee
so easily: Many people may be at their most infectious when exhibiting only mild, cold-like symptoms. &nbsp;&nbsp;&nbsp;"This is in
patients, viral shedding peaked about seven to 10 days into the illness, as the infection spread from the upper respiratory tract i
concentrations were reached before Day 5 and were more than 1,000 times higher" than those seen in SARS patients, the authors wrote
the first signs of pneumonia. In these severe cases, viral shedding reached maximum levels around Day 10 or 11. In the mild cases,
authors noted.&nbsp;"Based on the present findings, early discharge with ensuing home isolation could be chosen for patients who ar
copies of viral genetic material per milliliter, the authors wrote.&nbsp;"This is a very important contribution to understanding bo
shedding," Michael Osterholm, director of the University of Minnesota's Center for Infectious Disease Research and Policy, told Sta
by taking swabs from the patients' noses and throats, also examining their blood, urine, stool and sputum — a mixture of saliva and
viral genetic material called RNA to determine how much of the virus was present at different stages of the disease.&nbsp;Researche
remained infectious, as RNA from the virus may be present in human tissue but not functional. To find out who was infectious and wh
lab. &nbsp;The researchers found that they could grow virus from the throat, nose and sputum samples gathered early in the course o
That change indicates that those patients had become less infectious. Despite their improvement, they still tested "positive" for t
the body for at least two weeks after COVID-19 symptoms clear up.&nbsp;The team in the new study could not grow virus from any bloo
based on 13 samples collected between Day 6 and Day 12 from four patients, as these contained the largest quantities of viral RNA a
Organization suggested that "viable virus" could be recovered from infected people's stool, but it was unclear whether these fragme
mild cases, more research is needed to determine how stool might contribute to COVID-19 transmission, the authors noted.Notably, th
system begins building a defense against the pathogen soon after exposure. Scientists don't yet know whether this rapid immune resp
```

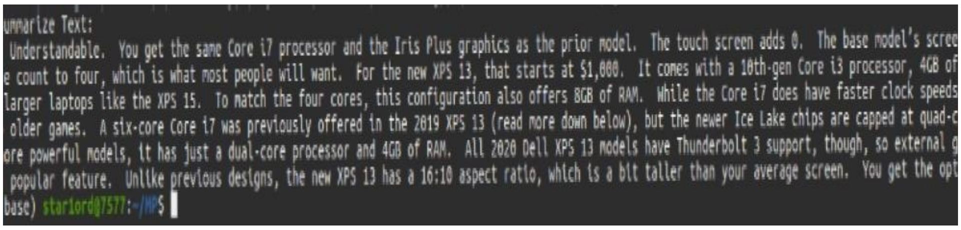
Fig 5. Web-scraping on keyword corona or Covid-19(Zoom appropriately)

We have run the tool on the search ‘Dell XPS i3 Laptop Reviews’. Inferring from the sentiments from the results, we got a positive articles percentage of 80%. The TF-IDF scores for the results from Google News on Corona is shown in Figure 6(a) and Summarization of the same is shown in Figure. 6(b).



```
  (0, 1744663)  1
  (0, 1744715)  1
  (0, 1745714)  1
  (0, 1747553)  1
  (0, 1756734)  1
  (0, 1757341)  1
  (0, 1757632)  1
  (0, 1758288)  1
  (0, 1758421)  1
/home/starlord/anaconda3/lib/python3.7/site-packages/sklearn/externals/joblib/__init__.py:15: DeprecationWarning: sklearn.externals.joblib is deprecated in 0.21 and will be removed in 0.23. Please import this functionality directly from j
oblib, which can be installed with: pip install joblib. If this warning is raised when loading pickled models, you may need to re-serialize those models with scikit-learn 0.21+.
  warnings.warn(msg, category=DeprecationWarning)
[1 1 1 1 1 1 1 1]
number of positive articles are:
8.0
percentage of positive articles:
100.0
(base) starlord@7577:~/MP$ python tf_idf.py
                  tfidf
core           0.406410
model          0.357641
xps            0.243846
screen         0.211333
gb             0.178820
configuration  0.162564
though         0.162564
come           0.146308
dell           0.130051
still          0.130051
graphic        0.130051
processor      0.130051
touch          0.130051
laptop         0.113795
ram            0.113795
resolution     0.097538
older          0.097538
year           0.097538
option         0.097538
offer          0.081282
NEWS TITLE tf_idf scores
                  tfidf
core           0.367912
model          0.315353
xps            0.262794
screen         0.227755
gb             0.192716
though         0.175196
graphics       0.140157
dell           0.140157
touch          0.140157
configuration  0.140157
comes          0.140157
still          0.140157
processor      0.122637
ram            0.122637
older          0.105118
resolution     0.105118
storage        0.087598
four           0.087598
extra          0.087598
like           0.087598
(base) starlord@7577:~/MP$ ▌
```

Fig 6(a). TF-Idf Scores for Google news corona



```
ummarize Text:
 Understandable.  You get the same Core i7 processor and the Iris Plus graphics as the prior model.  The touch screen adds 0.  The base model's scree
e count to four, which is what most people will want.  For the new XPS 13, that starts at $1,000.  It comes with a 10th-gen Core i3 processor, 4GB of
larger laptops like the XPS 15.  To match the four cores, this configuration also offers 8GB of RAM.  While the Core i7 does have faster clock speeds
 older games.  A six-core Core i7 was previously offered in the 2019 XPS 13 (read more down below), but the newer Ice Lake chips are capped at quad-c
ore powerful models, it has just a dual-core processor and 4GB of RAM.  All 2020 Dell XPS 13 models have Thunderbolt 3 support, though, so external g
 popular feature.  Unlike previous designs, the new XPS 13 has a 16:10 aspect ratio, which is a bit taller than your average screen.  You get the opt
base) starlord@7577:~/MP$ ▌
```

Fig 6(b). Summarization

References

1. A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," Expert Systems with Applications, vol. 41, no. 16, pp. 7653–7670, 2014.

2. L. S. Adriaanse and C. Rensleigh, “Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison,” The Electronic Library, vol. 31, no. 6, pp. 727–744, 2013
3. Herrmann, Markusa and Hoyden, Laura,” Applied Webscraping in Market Research”, First International Conference on Advanced Research Methods and Analytics, CARMA2016.
4. Hossam El-Din Hassanien,” Web Scraping Scientific Repositories for Augmented Relevant Literature Search Using CRISP-DM”, Appl. Syst. Innov. 2019, 2(4), 37; https://doi.org/10.3390/asi2040037.
5. Mika V. Mäntylä, Daniel Graziotin, Miikka Kuutila,” The evolution of sentiment analysis—A review of research topics, venues, and top cited papers”, Computer Science Review, Volume 27, February 2018, Pages 16-32, ISSN 1574-0137, https://doi.org/10.1016/j.cosrev.2017.10.002.
6. P. Divya Sree, G. Mallikarjuna Rao, “A Research on Passive Forgery Detection”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019
7. Steven Bird, Ewan Klein, and Edward Loper. “ Natural Language Processing with Python” (1st. ed.). O’Reilly Media, Inc,2009.
8. W. Richard Stevens. 1990. UNIX network programming. PrenticeHall, Inc., USA.
9. The OpenSSL Project (2003). OpenSSL: The Open Source toolkit for SSL/TLS www.openssl.org.
10. OpenSSL Foundation, I., 2020. /Docs/Index.Html. [online] Openssl.org. Available at: https://www.openssl.org/docs/> [Accessed 9 May 2020].