Google Colab Based Analysis of Web Links using Python Library

Shamsa Kauser Yasini[1], Vivek Shukla[2], S.R.Tandan[3]
[1]M.Tech Scholar, Department of Computer Science and Engineering
[2, 3]Assistant Professor, Dr. C V Raman University, Bilaspur, CG, India

Abstract—**One of the prominent source of information is internet. The rapid advancement of Morden technologies makes it more convenient for the user. Use of various networking sites create huge unstructured data. In this paper, we have tried to retrieve the links of website and analyse it to verify it's security level. Web scraping is used to extract URL using python library. Security levels of links have been analysed on the basis of number of URL retrieved. URL retrieval technique has been developed and done on online data. URL extraction method has been implemented to classify the web links using python and beautifulSoup.**

Keywords— **Security Analysis, Web Information Retrieval, Google Colab, Information Processing**

I. INTRODUCTION

The continuous growth in information technology requires a machine which can handle the system and extract the information correctly is the need of the current generation. The information retrieval system is to assist the users while accessing the retrieval environment. The major user group of commercial applications is using a traditional retrieval process of information which is available in structured format. The traditional system limitation is a difficulty with dealing unstructured data. The system should be able to process the information available in web which are mostly unstructured in nature. The application of modern machine learning tools are capable to process the unstructured data and make it useful.

The most essential element of an IR system is the Textual Archive which consist of textual units known as Document, and Document Retrieval Engine. The user enters the query with the required document information. The Document Retrieval Engine searches the similar document against query term from the knowledge base and responds with all possible lists of documents which are most relevant for the user.

The continuous growth in information technology requires a machine which can handle the system and extract the information correctly is the need of the current generation. This concept is classically known as Big Data. The deep investigation of intelligence and meaningful patterns from Big Data is known as Big Data Analytics. A number of researchers and scientists are working in this domain of Big Data using assorted technologies and tools. There are number of approaches by which the live data can be obtained for research and development. One of these approaches is getting data from Open Data Portals. The open data portals provide authentic data sets for research and development in multiple domains. The data sets can be downloaded from these portals in multiple formats including XML, CSV, JSON and many others.

Many times data is not easily accessible – although it does exist. As much as we wish everything was available in CSV or the format of our choice – most data is published in different forms on the web. What if you want to use the data to combine it with other datasets and explore it independently? [1] One of the solutions is Screen Scraping. Screen Scraping is the technique to capture the data that is being displayed in human readable format on the destination terminal and to replicate it at the source terminal for further processing. Screen scraping is sometimes referred to as terminal emulation [2]. Though there are other ways to get the data out of the web i.e., from web-based APIs, such as interfaces provided by online databases and many modern web applications (including Twitter, Facebook and many others). This is a fantastic way to access government or commercial data, as well as data from social media sites [3]. Extracting information from PDFs is beyond the scope of this paper, but there are some tools and tutorials that may help you do it [3]. But the advantage of scraping is that you can do it with virtually any web site — from weather forecasts to government spending, even if that site does not have an API for raw data access. However, screen scraping is not an independent process. Before scraping the output, Crawlers are responsible to navigate to the destination terminal. The search key entered at the source machine, engages the crawlers to navigate through the links on the web. Once the crawlers successfully reaches the correct page that matches up with the search string, scraping process starts.

II. RELATED WORK

Rahul Dhawani et al - The digital world is growing with a pace that exceeds the speed of any man made fastest prime movers. Here the term growing is used in context to the size of data. At 487bn gigabytes (GB), if the world's rapidly expanding digital content were printed and bound into books it would form a stack that would stretch from Earth to Pluto 10 times. The main contributors to this digital warehouse are social media, government surveillance cameras and plenty of other independent websites which are updated on daily basis such as inventories system of companies, their daily revenues as well as E-Commerce websites that comes up with FMCG's on daily basis. In this digital age, this web data is the most essential resource for any business. The main focus of this paper is to highlight the collection of data through scraping as API's are not available for each and every data source [19].

Manu Bansal - Web Monitoring, Scraping and digital forensic is one of the prominent areas in the domain of Big Data and Sentiment Analysis. A number of software products and tools are available in the technology market which are used to guards the network infrastructure and confidential data against cyber threats and attacks. From long time, the monitoring of servers and forensic analysis of network infrastructure is done using packet capturing (PCAP) tools. These activities are performed using PCAP and related tools available in the market which includes open source software as well as commercial products. As far as the fame and usage of the software suites is concerned, the open source market is getting popularity because of the scope of customization and organization specific personalization the software products. In this research paper, an approach is depicted for the fetching and analysis of live data from social media portals and using such approaches the sentiment data analysis can be implemented effectively [20].

The digital world is growing with a pace that exceeds the speed of any man made fastest prime movers. Here the term growing is used in context to the size of data. At 487bn gigabytes (GB), if the world's rapidly expanding digital content were printed and bound into books it would form a stack that would stretch from Earth to Pluto 10 times. The main contributors to this digital warehouse are social media, government surveillance cameras and plenty of other independent websites which are updated on daily basis such as inventories system of companies, their daily revenues as well as E-Commerce websites that comes up with FMCG’s on daily basis. In this digital age, this web data is the most essential resource for any business. The main focus of this paper is to highlight the collection of data through scraping as API’s are not available for each and every data source.

Web Monitoring, Scraping and digital forensic is one of the prominent areas in the domain of Big Data and Sentiment Analysis. A number of software products and tools are available in the technology market which are used to guards the network infrastructure and confidential data against cyber threats and attacks. From long time, the monitoring of servers and forensic analysis of network infrastructure is done using packet capturing (PCAP) tools. These activities are performed using PCAP and related tools available in the market which includes open source software as well as commercial products. As far as the fame and usage of the software suites is concerned, the open source market is getting popularity because of the scope of customization and organization specific personalization the software products. In this research paper, an approach is depicted for the fetching and analysis of live data from social media portals and using such approaches the sentiment data analysis can be implemented effectively.

III. Simulation of Proposed methodology

The proposed technique is based on the retrieval of web links and analysing its accessibility. The proposed work is simulated in google cloud based Colaboratory environment. Colaboratory is a Google research project created to help disseminate machine learning education and research. It's a Jupyter notebook environment that requires no setup to use and runs entirely in the cloud. In this work, we have developed working model. By using this methodology web link transform into visual blocks. A visual block is actually segment of webpage. The system is automatic top-down; tag tree independent approach to detect web content structure. Basically, the block-based page content structure is obtained by using python script in BeautifulSoup.

1. Initializing Google Colab
2. Environment Setting
3. Installation of python library
4. Execution of script
5. Content structure construction

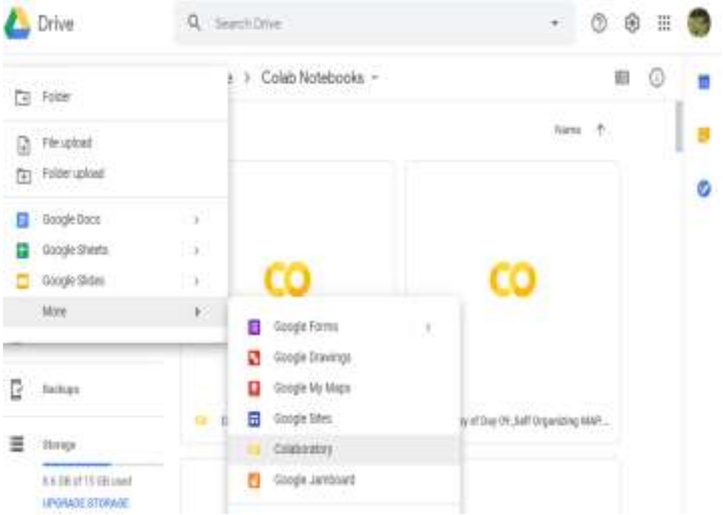
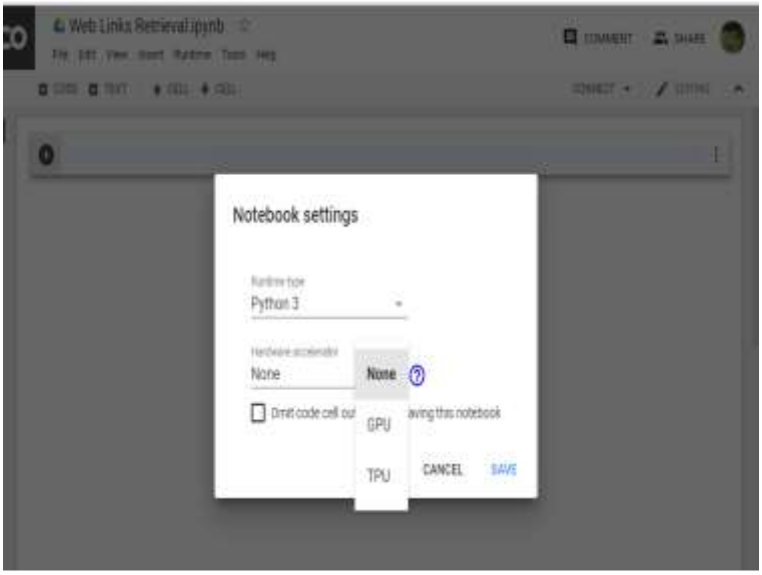


Figure 1.1 Initialization of Google Colab


Web Links Retrieval.ipynb
File Edit View Insert Runtime Tools Help
COMMENT
SHARE
CODE TEXT CELL CELL
CONNECT
EDITING
CANCEL
SAVE

Notebook settings

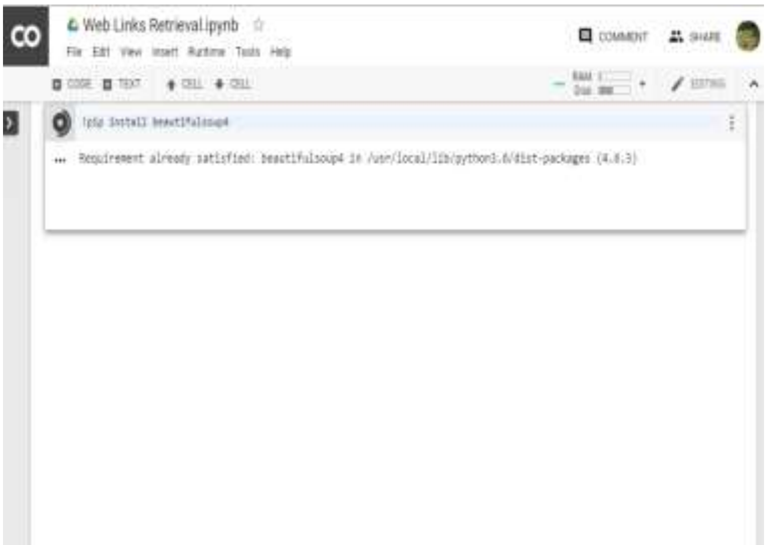
Runtime type
Python 3

Hardware accelerator
None

None
GPU
TPU

□ Omit code cell ou aving this notebook

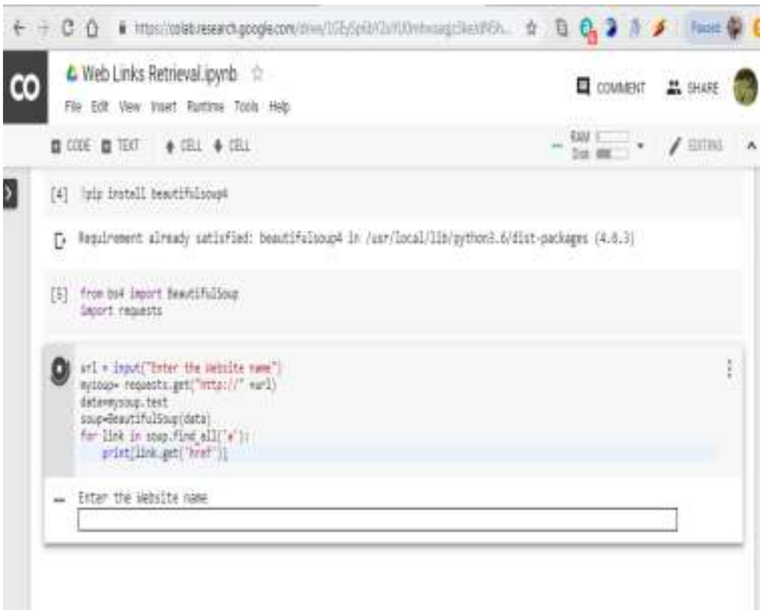
Figure 1.2 Environment Setting


Web Links Retrieval.ipynb
File Edit View Insert Runtime Tools Help
COMMENT
SHARE
CODE TEXT CELL CELL
RAM
Disk
EDITING

```
!pip install beautifulsoup4
```

```
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.6/dist-packages (4.6.3)
```

Figure 1.3 Installation of Python Library


https://colab.research.google.com/drive/1GEySpkbY2uYU0mtwsagcSkeIN6h...
Web Links Retrieval.ipynb
File Edit View Insert Runtime Tools Help
COMMENT
SHARE
CODE TEXT CELL CELL
RAM
Disk
EDITING

```
[4]  !pip install beautifulsoup4
```

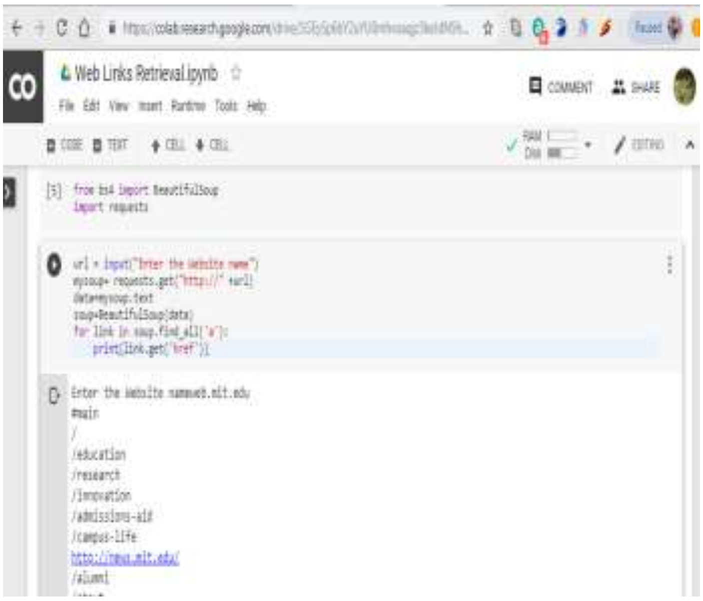
```
Requirement already satisfied: beautifulsoup4 in /usr/local/lib/python3.6/dist-packages (4.6.3)
```

```
[5]  from bs4 import BeautifulSoup
     import requests
```

```
url = input("Enter the Website name")
mysoup= requests.get("http://" +url)
data=mysoup.text
soup=BeautifulSoup(data)
for link in soup.find_all('a'):
    print(link.get('href'))
```

```
Enter the Website name
```

Figure 1.4 Execution of Script


Web Links Retrieval.ipynb
File Edit View Insert Runtime Tools Help
COMMENT
SHARE
CODE TEXT CELL CELL
RAM
Disk
EDITING

```
[5] from bs4 import BeautifulSoup
    import requests
```

```
url = input("Enter the website name")
mysoup= requests.get("http://" +url)
data=mysoup.text
soup=BeautifulSoup(data)
for link in soup.find_all('a'):
    print(link.get('href'))
```

```
Enter the website nameweb.mit.edu
#main
/
/education
/research
/innovation
/admissions-aid
/campus-life
http://news.mit.edu/
/alumni
```

Figure 1.5 Retrieval Process of Website

TABLE 1.1 URL EXTRACTIONS OF VARIOUS WEBSITES

S.No	Name of Website	No. of Link Retrieved	Organization
01	bitmesra.ac.in	246	Birla Institute of Technology, Mesra Ranchi, INDIA
02	nitttrbpl.ac.in	82	National Institute of Technical Teachers Training, Bhopal, INDIA
03	cvru.ac.in	230	Dr C V Raman University, Bilaspur, INDIA
04	web.mit.edu	69	Massachusetts Institute Technology, Cambridge, USA
05	nasa.gov	0	National Aeronautics and Space Administration, USA
06	isro.gov.in	148	Indian Space Research Organization, INDIA
07	barc.gov.in	123	Bhabha Atomic Research Centre, INDIA
08	fbi.gov	152	Federal Bureau of Investigation, USA
09	bell-labs.com	108	Bell Laboratory, USA
10	tcs.com	0	Tata Consultancy Services Ltd, INDIA
11	infosys.com	148	Infosys Consultants Pvt. Ltd, INDIA
12	iitb.ac.in	248	Indian Institute of Technology, Bombay, INDIA
13	iiitnr.ac.in	138	International Institute of Information Technology, Naya Raipur, INDIA
14	mciindia.org	139	Medical Council of India, INDIA
15	drdo.gov.in	0	Defence Research Development Organization, New Delhi, INDIA
16	bseindia.com	72	Bombay Stock Exchange, Bombay, INDIA
17	aajtak.intoday.in	776	Aaj Tak News Channel, Noida, INDIA

The above table shown in Table 1.1 have been retrieved using python based beautifulsoup web scrapping library. The data given are the numeric values which is number of links retrieved. Based on the retrieved links we have made following observations.

i. It is a very powerful method to retrieve the all links.

ii. Its shows that webpages consist of dummy links.

iii. Secure website links cannot be retrieved so we can easily observe its security level by running proposed technique.

iv. In the Table 1.1 we have shown that the technique could not retrieve any links while analyzing the website like TCS, DRDO (Defence Research Development Organization) and NASA links extraction method failed to fetch any link as of highly secured sites.

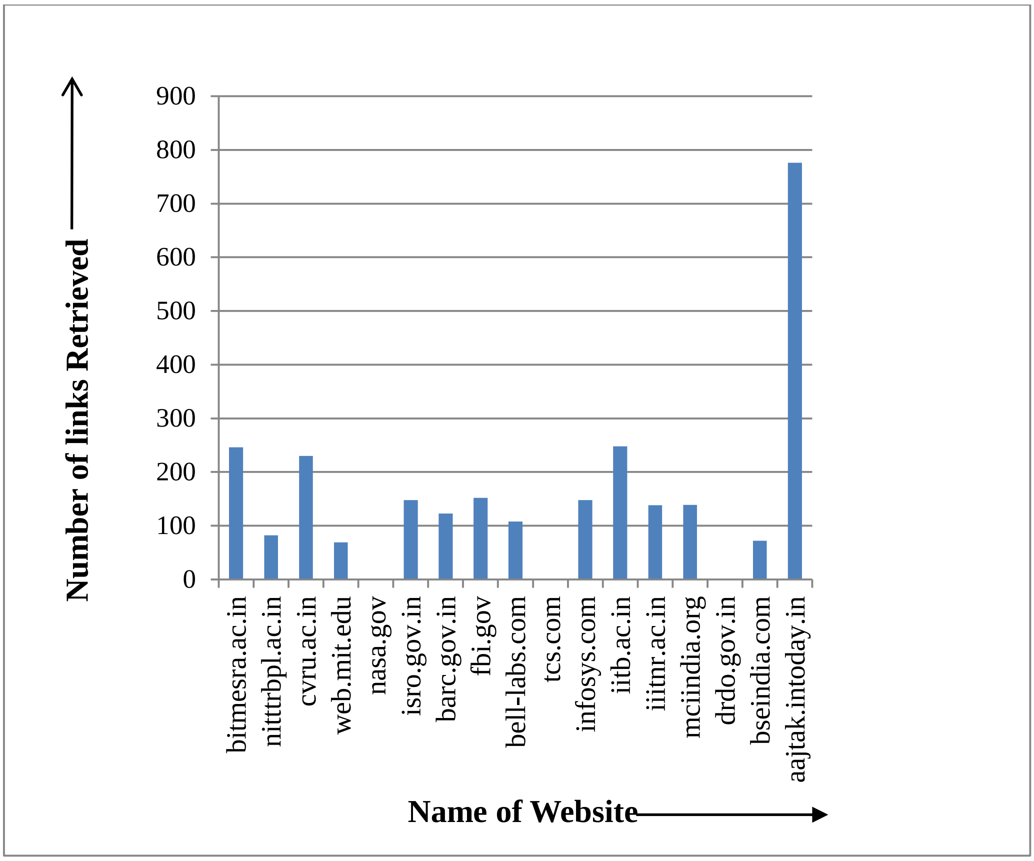


Figure 1.4 URL Extraction Chart

IV. CONCLUSIONS

The web link retrieval is one of the most recent area of research. In developed technique we have tried to retrieve the links of various website. Web information is mostly in an unstructured format. The developed method is useful to retrieve the unstructured data and make it useful. The method is useful to observe the visibility of links. The method works efficiently while retrieving web contents compared to existing techniques. Advantage of python based scripting method is fast and easy to deal with complex URLs. The developed method is useful while analyzing security level of web links. Finally the conclusion of the work is to retrieve the unstructured data from the website and also to remove the dummy links and manage the server load and make it efficient has been suggested.

ACKNOWLEDGMENT

I would like to thank my research guide Mr. Vivek Shukla and Dr. S. R. Tandan for his valuable support during research work. I would like to thank all respected faculty members of Department of Computer Science and Engineering, Dr. C V Raman Institute of Science and Technology.

REFERENCES

[1] Making data on the web useful: scraping
[2] Screen Scraping: Techopedia
[3] Getting Data from the Web: Data Journalism Handbook
[4] urllib2 — extensible library for opening URLs: https://docs.python.org
[5] Beautiful Soup Documentation – www. crummy.com
[6] Crawling the Web, Gautam Pant, Padmini Srinivasan and Filippo Menczer.
[7] Web Crawler: A Review , Md. Abu Kausar , V. S. Dhaka Dept, Sanjeev Kumar Singh
[8] Web Scraping, Wikipedia.
[9] Text Categorization by Fabrizio Sebastiani Dipartimento di Matematica Pura e Applicata Universita di Padova ` 35131 Padova, Italy
[10] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis.Foundations and trends in information retrieval, 2(1-2), 1-135.
[11] Pang, B., & Lee, L. (2004, July). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (p. 271). Association for Computational Linguistics.

[12] Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 347-354). Association for Computational Linguistics.

[13] Pak, A., & Paroubek, P. (2010, May). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 1320-1326).

[14] Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the 2nd international conference on Knowledge capture (pp. 70-77). ACM.

[15] Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing, 2, 627-666.

[16] Mullen, T., & Collier, N. (2004, July). Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In EMNLP (Vol. 4, pp. 412-418).

[17] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the Workshop on Languages in Social Media (pp. 30-38). Association for Computational Linguistics.

[18] Manu Bansal, “Sentiment Analysis from Social Media Live Feeds Using Unstructured Data Mining” International Journal of Computing and Corporate Research ISSN (Online): 2249-054x Volume 5 Issue 5 September 2015

[19] Rahul Dhawani, Mrudav Shukla, Priyanka Puvar, Bhagirath Prajapati A Novel Approach to Web Scraping Technology International Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 5, MAY 2015 ISSN: 2277 128X

[20] George Foreman (2003), An Extensive Empirical Study of Feature Selection Metrics for Text Classification. Journal of Machine Learning Research, pp.1289–1305.