iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** XI **Month of publication:** November 2018

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue XI, Nov 2018- Available at www.ijraset.com

Extraction and Conversion of Web JSON Data into Pandas Data Frame by storing it into Text File using Python

Shweta Chaturvedi[1], Prashant Kumar[2]
[1]*Department of Computer Science and Engineering, Jaipur Engineering College and Research Centre, Jaipur, India*
[2]*Department of Electronics and Communication Engineering, M.S. Ramaiah Institute of Technology, Bangalore, India*

Abstract*: In this paper, we have established a technique of extracting the web data that are available in structured JSON format from the open source by making an API call and storing it into the text file on the local disk for future references and utilization. Finally, we parse the json data of text file into python object and convert it into a Pandas DataFrame for transformation, prediction and optimization by using python script in Jupyter Notebook.
Keywords: Web Scraping, JSON data, Pandas DataFrame, Urllib2, Python

I. INTRODUCTION

As by the Facebook statistics, there are 2.23 Billion monthly active users which are also increasing by 11 percent every year. Besides this ,in each minute 136,000 snaps are uploaded , 510,000 comments and 293,000 statuses are posted by the active profiles .Similarly ,the Wikipedia pages generate by the rate of 2 edits per second statistically observed by Wikipedia analyst . At present, Wikipedia has total 5,726,316 articles of English and publish 553 new articles in twenty four hour period approximately. Do we ever imagined how these loads of data stored, collected and managed by these renowned business organization .Every Firm has its own different mechanism to handle large amount of data, for instance, Google uses its Data Centre that are located in 14 different location around the globe to handle large amount of data. Facebook has created its own method known as Presto mechanism to handle large amount of its data. There are many structured, unstructured and heterogeneous type of data file format available like CSV, Image file format ,Plain text, Binary file, XML, JSON, HTML, Excel, PDF etc. by which we can extract the data according to the demand and requirement. Data Management and Utilization are more vital in current data driven world than ever before as only those organization are dominant and ruling the industry who realized the powerful impact of data analysis at early stage. There are numerous examples which are stressed upon this fact like Netflix has replaced the blockbuster (an American-based organization for providing home theatre, video games, DVD, streaming and on demand videos), Amazon continues to shake up the offline retail markets, and Uber constantly diminishes the taxi business. Analyzing the data can be done by numerous tools as EXCEL for beginners and SAS (Base SAS and Enterprise Minor), IBM (Cognos and SPSS), ANGOSS, R, KXEN, Weka, Pentaho, StatSoft, JMP, Rapid-I, Python etc. for expertise plays the significant role for exploration, preparation, scrubbing and visualizing the data which can also be useful for performing the complex algorithms such as Machine learning, Artificial Intelligence, Deep learning, Neural network, Natural Language Processing (NLP). As data is increasing persistently, it leads to the introduction of new terminology in the field of Analytics known as “Big Data” for analyzing a large number of structured and unstructured data which is beyond the scope of commonly used software tool .To handle these kind of data sets, different modern tool such as Apache Pig, HPCC, Flink, Kaggle, Hadoop, Spark etc. are used to analyze and visualize it.

II. WEB SCRAPING

Web Scraping is a technique of extracting structured or unstructured data from the websites available in different format like JSON, CSV, EXCEL, HTML and XML etc. [1] for future studies and researches. Data exhibit by websites can only be discern by web browser as they do not offer the functionality to save the data into the local drive of the system .One way is to copy and paste the data manually which is a tedious, protracted and impractical process as it may take from few hours to few days to complete it. Therefore, web scraping is basically an automatic process by which data can queried and extracted from the website having public API within a fraction of seconds. In addition one major advantage of web scrapping is that it can penetrate the information to much extant than a traditional search engine. For instance, if we searched “cheapest five star Hotel accommodation in Bangalore”, then google will answer the query with uncontrollable advertisement and popular hotel search sites. However if we do same thing with

seasoned web scrapers then we can get the information of which five star hotel in Bangalore is best in terms of cost for accommodation across variety of websites and the best time to book the hotels. Sometimes webpage administrator does not want to share the web page information so they use API key for the security of data .This is where web Scrapping comes into picture with some exception, as it can access the information from the webpage using API call through Python script and store it in the Database and then we can perform any modification with the data.

A. JSON Format Data

JSON (JavaScript Object Notation) is a most commonly used data format for data interchange between two machine applications across the globe or within same hardware system as it is a language-independent and most accessible data format in the world due to its simplicity in understanding and can be extracted without any tool like Oracle, MySQL and Microsoft-EXCEL [2]. JSON has bi-functional characteristics as it is in human as well as machine readable format. Indeed, while applications/libraries can parse the JSON data, humans can also look and interpret the data along with its meaning. JSON data mainly contains curly braces, text, square brackets, commas, colons, double quotes, inverted comma and some other characters in its format. There are many online as well as offline tool, formatter and software available to convert JSON data format to many other format like CSV, XML, EXCEL, SQL, C#, JAVASCRIPT etc.

B. Pandas DataFrame

Pandas is one of the most popular packages of data science which provide powerful, flexible and expressive data structure on which data manipulation and analysis can be applied. Pandas works on tabular and structured frame data which is either one dimensional (Series) or two dimensional (DataFrame). Pandas DataFrame is a two dimensional labeled data structure which generally consist of three main components: Data, Index and Columns [3]. DataFrame may contain the data from few hundred to thousands in terms of rows. We can view them at any point of time selectively by using appropriate functions. To view the rows we can use head (…) and tail (…) functions, which by default shows the first five and last five rows of data , if no input is provided in their arguments, otherwise shows specific number of rows from top to bottom.

III.PROPOSED METHODOLOGY AND IMPLEMENTATION

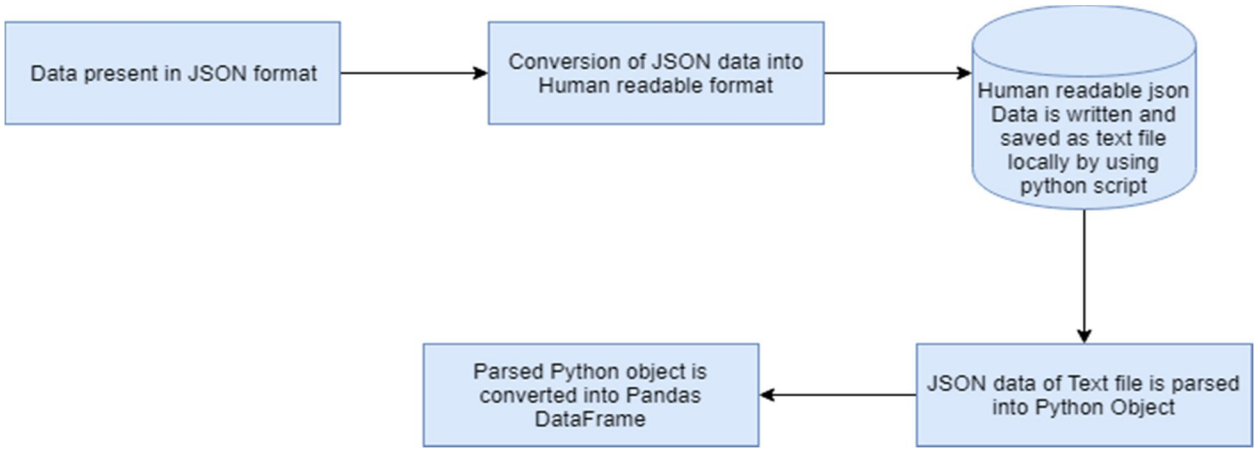


Fig. 1. Basic Block Diagram of Proposed Technique**.**

In our implementation, we are using the web based IPython2 (command shell for providing a platform to multiple programming language to execute, especially for data science) notebook known as Jupyter notebook for implementing the codes of python programming language. Python has the extensive framework and libraries like NumPy, Pandas, SciPy, Matplotlib, Sklearn etc. for performing various task in data science. In our methodology, we are using Pandas library, Urllib2 module and Json Serializer (json.loads and json.dumps) for the conversion of Json Data into Pandas DataFrame by following process:

A. Use of Urllib2 module for open the URL provided by user [4].
B. Use of json.loads () function for parsing the json data to python objects and json.dumps () function to decode the json data.
C. Use of write () function to write the json data into Text file and open () function to open the same Text File stored in local drive in the system.
D. Parsing of JSON data (available in text file) to python object for converting into Pandas DataFrame for future analysis.

Step 1: Data Present in Json format.


```
{"a2gov_org":{"name":"a2gov_org","title":"Ann Arbor,
Michigan","url":"http://www.a2gov.org/services/data/Pages/default.aspx","author":"City of Ann Arbor","publisher":"City of
Ann Arbor","issued":"","publisher_classification":"","description":"City of Ann Arbor's Open Data Catalog (USA)","tags":
["ctic","unitedstates"],"license_id":"","license_url":"","place":"Ann Arbor,
Michigan","location":"42.2681569,-83.7312291","country":"US","language":"en","status":"active","metadatacreated":"2011-06-
27T18:12:57.439Z","generator":"","api_endpoint":"Not
apparent","api_type":"","full_metadata_download":"","id":"a2gov_org","description_html":"<p>City of Ann Arbor&#39;s Open
Data Catalog (USA)</p>\n","groups":[]},"acikveri-sahinbey-bel-tr":{"name":"acikveri-sahinbey-bel-tr","title":"Açık Veri
Portali - Test Yayını","url":"http://acikveri.sahinbey.bel.tr/dataset","author":"pinardag","publisher":"SahinBey Belediyesi
","issued":"31/01/2015","publisher_classification":"Government","description":"The first official open data portal of
Turkey","tags":
["turkey","national"],"license_id":"Unknown","license_url":"","place":"Gaziantep,Turkey","location":"37.0587715,37.380137",
"country":"TR","language":"tr","status":"active","metadatacreated":"","generator":"","api_endpoint":"","api_type":"","full_
metadata_download":"","id":"acikveri-sahinbey-bel-tr","description_html":"<p>The first official open data portal of
Turkey</p>\n","groups":[]},"africa_open_data":{"name":"africa_open_data","title":"Africa Open
Data","url":"http://africaopendata.org/","author":"Africa Open Data","publisher":"Africa Open
Data","issued":"","publisher_classification":"","description":"Africa's largest central repository for Government, Civil
Society, Corporate and Donor Agency Data.","tags":
```


Fig. 2. Json Format Data contains text, commas, colons, square brackets, and some other characters.

Step 2: Website URL is entered to convert JSON data into human readable format.

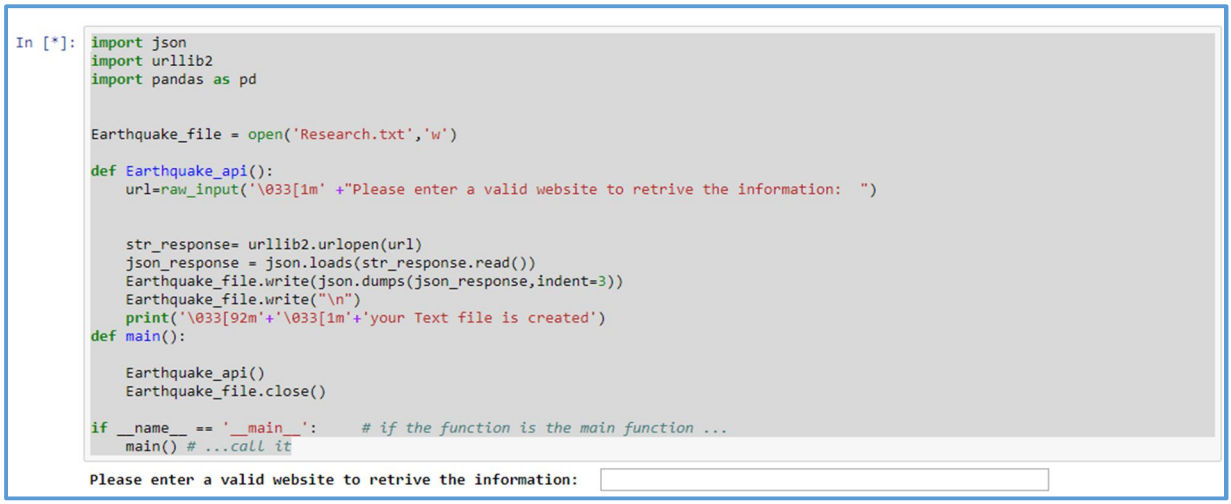


Fig. 3. Python script after running ask for URL from user.

Step 3: Human Readable text file is created.

Fig. 4. Text file will be create after getting the notification “Your file is created”.

Step 4: Text File is Created on Jupyter Notebook and saved locally.

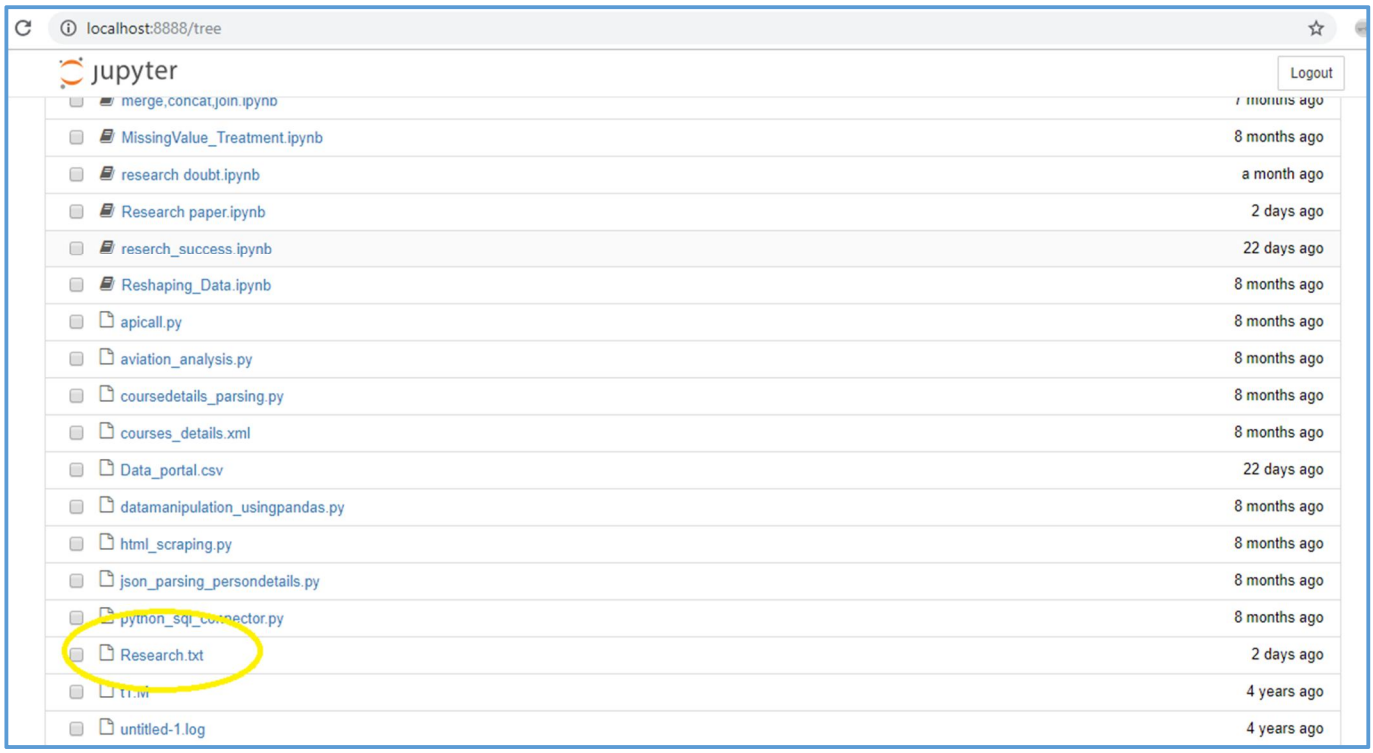


Fig. 5. Text file creation highlighted in the figure

Step 5: Created Human Readable JSON Text file.

Research - Notepad

File Edit Format View Help


```
{
   "tunisia_opendataforafrica": {
      "issued": "",
      "id": "tunisia_opendataforafrica",
      "generator": "",
      "author": "African Development Bank Group",
      "api_endpoint": "",
      "location": "33.8439408,9.400138",
      "metadatacreated": "2013-08-29T08:11:07.071Z",
      "place": "Tunesia",
      "description_html": "<p>Open data portal for Tunesia</p>\n",
      "status": "active",
      "description": "Open data portal for Tunesia",
      "tags": [
         "africa"
      ],
      "publisher_classification": "",
      "groups": [],
      "license_id": "other-at",
      "publisher": "African Development Bank Group",
      "name": "tunisia_opendataforafrica",
      "language": "en",
      "url": "http://tunisia.opendataforafrica.org/",
      "country": "tun",
      "title": "Open Data for Tunesia",
      "license_url": "",
```


Fig. 6. JSON data stored in Human Readable format

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue XI, Nov 2018- Available at www.ijraset.com

Step 6: JSON data of text file parsed into python object.

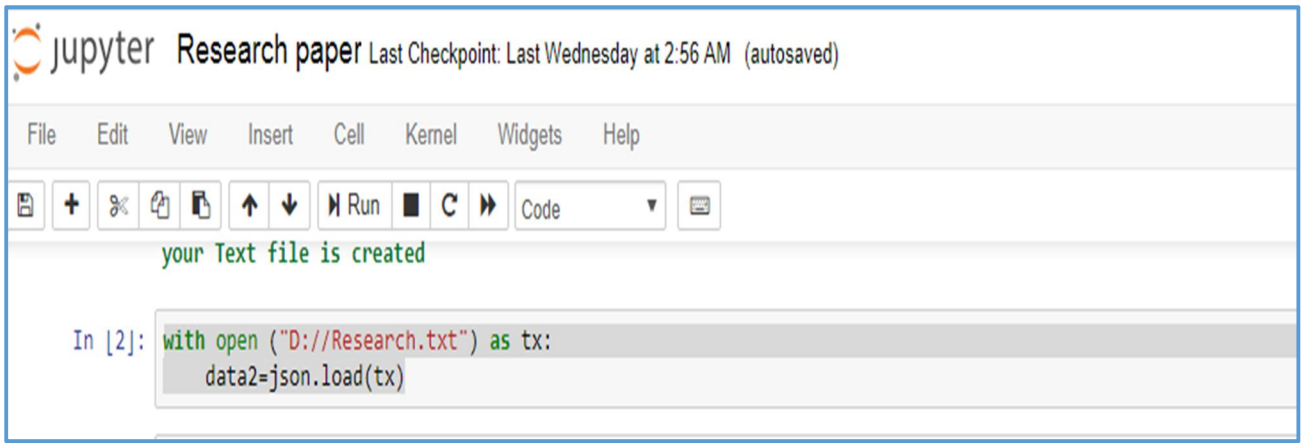


Fig. 7. Parsing of data into Python Object.

IV. RESULTS

In our implementation, we finally converts the JSON web data into pandas dataframe and call for the top five rows of the python data by the head() function. Fig.8 provide the clear evidence that how a non-human readable web json data get scraped ,stored and parsed into Pandas DataFrame for the detail examination.

```
In [7]: data3=pd.DataFrame(data2)
In [8]: data3.head()
Out[8]:
```

	a2gov_org	acikveri-sahinbey-bel-tr	africa_open_data	ajuntament-de-tarragona	ajuntament-de-terassa	alabama	albania	alberta_ca	algeria_opendataforafrica
api_endpoint	Not apparent		http://africaopendata.org/api/			Not apparent	Not apparent	Not apparent	http://opendata.aragon.es/catalogo/api
api_type									
author	City of Ann Arbor	pinardag	Africa Open Data	Ajuntament de Tarragona	Ajuntament de Terassa		ShqiperiaCom Shpk / OSI / OSFA	Province of Alberta (Her Majesty the Queen in ...	African Development Bank Group
country	US	TR	AF	ES	ES	US	AL	CA	DZ
description	City of Ann Arbor's Open Data Catalog (USA)	The first official open data portal of Turkey	Africa's largest central repository for Govern...	Open Data Tarragona	Open Data Terassa	Open Alabama Initiative	Albania Open Data Catalog	You may request datasets via [webform] (http://...	Open data portal for Algeria

5 rows × 551 columns

Fig. 8. Call for top 5 rows from converted Pandas DataFrame.

V. CONCLUSION AND SCOPE FOR FUTURE WORK

Web scraping is the convenient way of storing the transient and frequently changing web data for the utilization in future as by this developed technique, the data available in JSON format on the web page can be stored in a text file in human readable format which can be transferable to the numerous users as well as keep in reserve for subsequent exploring, analyzing, predicting, extracting, transforming, optimizing and updating the data for various experimentation and investigation by parsing them into dataframe using python script.

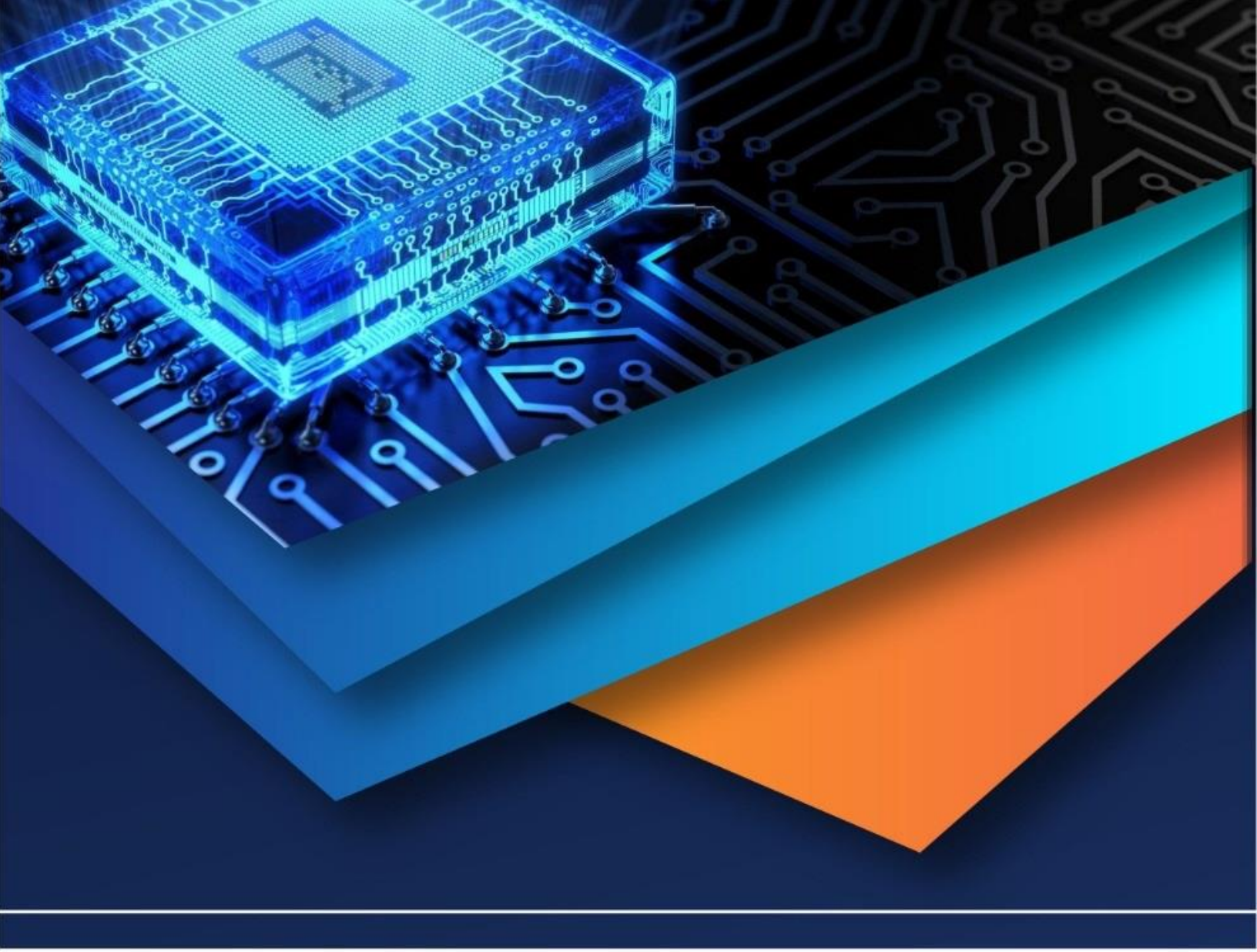
In future, we are interested in extracting, storing and exploring the variety of data like Pixel format, Image file format, Video format, Binary file format etc. and the conversion of JSON data into Html, Ajax, Xml, SQL and many other data format to make this technique more generic, proficient and containing the utility by using python script in the field of Data Science and Analytics. In future, more sophisticated algorithms can be adopted in order to save and deposit the homogenous as well heterogeneous format of data for implementing the steps of analytics altogether.

VI. ACKNOWLEDGEMENT

We would like to thank our research guide Mrs. Neelam Chaplot and Mrs. Lakshmi Shrinivasan for their valuable support during research work. We would like to thank all respected faculty members of Department of Computer Science and Engineering, Jaipur Engineering College and Research Centre and Department of Electronics and Communication Engineering, M.S.Ramaiah Institute of Technology.

REFERENCES

[1] Web Scraping, Wikipedia.
[2] "JSON Tutorial", Available from: http://www.w3schools.com/json
[3] Pandas, provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive — http://pandas.pydata.org/.
[4] Urllib2 — extensible library for opening URLs: https://docs.python.org
[5] Json.dumps & json.loads —https://www.journaldev.com
[6] Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing, 2, 627-666
[7] A. Abd El-Aziz and A. Kannan, "JSON Encryption", 2014 International Conference on Computer Communication and Informatics (ICCCI -2014), (2014) January, 03–05, pp. 1-6, Coimbatore, INDIA.
[8] Peng, L. Cao and W. Xu, "Using JSON for data exchanging in web service applications", Journal of Computational Information Systems, vol. 7, no. 16, (2011), pp. 5883-5890.
[9] Getting Data from the Web: Data Journalism Handbook
[10] 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.
[11] Rahul Dhawani, Mrudav Shukla, Priyanka Puvar, Bhagirath Prajapati A Novel Approach to Web Scraping Technology International Journal of Advanced
[12] Research in Computer Science and Software Engineering Volume 5, Issue 5, MAY 2015 ISSN: 2277 128X
[13] Human language technology and empirical methods in natural language processing (pp. 347-354). Association for Computational Linguistics.
[14] Text Categorization by Fabrizio Sebastiani Dipartimento di Matematica Pura e Applicata Universita di Padova ` 35131 Padova, Italy.
[15] Armano, G. & Vargiu, E. A unifying view of contextual advertising and recommender systems. Proceedings of International Conference on Knowledge Discovery and Information Retrieval (KDIR 2010). 2010: 463-466.
[16] Mohd Kamir Yusof, Mustafa Man. Efficiency of JSON Approach for Data Extraction and Query Retrieval. Indonesian Journal of Electrical Engineering and Computer Science. Oct 2016; 4(1): 203- 214.
[17] Miki Enoki, Jerome Simeon, Hiroshi Horri, Martin Hirzel. Event Processing over a Distributed JSON Store: Design and Performance. WISE 2014, Part II, LNCS 8787. 2014: 395–404
[18] Dunlu Peng, Lidong CAO, Wenjie XU. Using JSON for Data bn exchanging in Web Service Applications. Journal of Computational Information System. 2011; 7(16): 5883–5890.
[19] Broder, A., Fontoura, M., Josifovski, V. & Riedel, L. A semantic approach to contextual advertising. SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, New York, NY, USA. 2007; 559-566
[20] M. Kameswara Rao, Rohit Lagisetty, M.S.V.K. Maniraj, K.N.S. Dattu, B. Sneha Ganga. Commodity Price Data Analysis Using Web Scraping.International Journal of Advances in Applied Sciences (IJAAS), Vol. 4, No. 4, December 2015, pp. 146~150.
[21] Lewerenz, E. An example of website screen scraping. Proceedings of MWSUG 2009, 2009.
[22] Manning, C.D., Raghavan, P. & Schtze, H. Introduction to Information Retrieval. Cambridge University Press, NewYork, NY, USA, 2008.
[23] Mehlfuehrer, A. Web scraping - a tool evaluation. Master's thesis, Wien University, 2009.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY