JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Scraping and Visualization of Product Data from E-commerce Websites

V. Srividhya[1*], P.Megala[2]

[1, 2] Dept of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, Tamilnadu, India - 641043
[*]*Corresponding Author: vidhyavasu@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v7i5.14031407 | Available online at: www.ijcseonline.org

Accepted: 12/May/2019, Published: 31/May/2019

Abstract— the paper entitled as "Scraping and Visualization of Product Data from E-commerce Websites"**.** Extracting data from websites is called as web scrapping**.** The main advantages of the scrapping are inexpensive, easy to implement, low maintenance and speed. The main objective of the work is to scrap the data from websites and store the extracted data in Comma-separated values (CSV) format for analysis. The data available in the websites are in the form of unstructured information. Web scraping helps to collect these unstructured data and store it in a structured form. The process of Web scraping is to extract the data using various methods from the internet. Millions of people consider universally accessible resource as internet. The rise in the usage of internet has commonly been increased day by day and there is high growth in competition between the organizations in their business. This work consists with three phases. The first phase of the work is web scrapping. In this phase, the extracted data will be stored as a csv file. The second phase of the work is data analysis. In this phase, the data is imported from the csv format and analyzed using statistical analysis. The third phase of the work is visualization and in this the extracted data has been visualized with the help of different charts.

Keywords: Web scrapping, Data Analysis, Visualization, data mining, websites.

I. INTRODUCTION

In today's scenario the need for on-line shopping over the Traditional Shopping is being increased day by day. To perform this people are utilizing electronic gadgets such as tablets, cell phones, PC and work area to get into an E-Commerce sites through the Internet. Web scrapping is a generally new strategy for gathering the web information. The term depicts the mechanized procedure of getting into sites and downloading explicit data, costs, item name and ratings. The main aim of this work is to gather and store information. The accessible innovations are connected might be in various languages like java, python, php and so on. Since there is ascend in new online business through web this adversary affects the purchasers also. The papers[1], has adopted the web scraping techniques in the web advertising field which also explains the collaborative filtering ways of web scraping with preferred implementation ads. As the scraping has a usage in adverting field this has become important in learning various methods. Internet showcasing analyst use web scrapping techniques to get some data from different contenders, for example, emails, directed catchphrases, joins and furthermore traffic sources.

The analysis of data is particularly required in a general public way to extract any data and to change it into particular format. In this way, web scrapping services impact the result which is required from the data accumulation. The change of the helpful substance on websites into significant business resources is the procedure of Web data extraction.

The scrapping methods are utilized for individuals just as business usage. Each and every procedure accessible has its own advantages and disadvantages to defeat this and there is a need for the unmistakable thought on the use of these systems in social networking.

The paper is organized with five sections. Section I deals with the concept of introduction. The Section II gives the related work of web scrapping. The methodology flow diagram explained in Section III. Section IV describes the results and discussion. The final part of the paper Section V deals with conclusion and future research work directions.

II. RELATED WORK

Web scrapping is the procedure of naturally gathering data from the World Wide Web. It is a field with dynamic advancement imparting a shared objective to the semantic web vision, an eager activity that still requires a leap forward in content handling, semantic understanding, artificial intelligence and human computer interactions [2].

The related works on the web scraping techniques involve [3] in this paper have different viewpoints, for example, extraordinary web scrapping dimensions and the sentimental methodology has been incorporated. This paper gives the study on human opinion mining where screen scraping plays the significant role. The most widely recognized accessible tools and strategies which are free and simple to utilize have been utilized by numerous clients.

The paper [4] uses the algorithm to explain XPath using Tree edit distance matching algorithm. The problem of computing the tree edit distance between trees is a variation of the classic string edit distance problem for extracting data. The authors of [5] have a study on the methods of web content mining and relate web scraping tools that are accessible.

As these papers comprise numerous topics under data mining which gives the reasonable thought on the distinctive accessible methods under data mining and the comparison have been made with web scraping. Web Crawler/scrapers are addressed for extracting URLs from various E-commerce sites.

Crawlers are capable of exploring to the destination terminal. The search key entered at the source machine, engages the crawlers to explore through the connections on the web. When the crawler effectively achieves the right page that match up with the search string, scrapping process begins [6].

Web scraping is a form of data mining. The essential and significant aim of the web scrapping process is to mine data from unstructured sites and change it into an intelligible structure like spreadsheets, database or a comma-separated values (CSV) file[7].

The paper [8] uses Web Scrapping to extract HTML information from URL's and use it for individual reason. As this is price comparison website, data is scrapped from multiple e-commerce sites. The clients before obtaining the items on-line, they have to visit diverse E-business sites to locate the specific items at the least expensive cost.

Real-time product Analysis utilizing Data Mining takes care of this issue for the client by giving client the items from various E-trade sites at one spot with various costs and plans and offers by various E-business firms [9].

To view and think about costs of a specific item from various sites and buy the item which he/she finds suitable for him/her is conceivable in Real Time Product Analysis utilizing Data Mining. In general, this will decrease Time and effort put by client giving client ease and agreeable outcomes. It will in general save client from savage pricing systems forced by various E-trade sites. In 2009 the usefulness and execution of e-commerce shopping bots for E-business were researched [10].

The paper [11] uses web scrapping for business and personal requirements and they are endless. Every business or individual has his or her very own particular requirement for gathering data.

The paper [12] provides a short-survey on state of the art of the discipline. It discuss with Web Data Scrapper System, Taxonomy for characterizing Web Data Scrapper Tools and an overview about Web Data Scrapper Tools.

Many earlier works have attempted to collect the historical data using web scraping. The paper [13] gives the detailed explanation on how web scrapping was done using libraries such as Selenium and Beautiful Soup. It also discusses about data cleaning and feature engineering to generate several features.

III. METHODOLOGY

This work comprises of three phases. The first phase is to scrap the product details like product name, product price and their ratings from e-commerce websites – Flipkart, Snapdeal and store it in a csv format. The second phase of the work is data analysis. The third phase is visualization. The following figure 1 shows the overall methodology diagram.

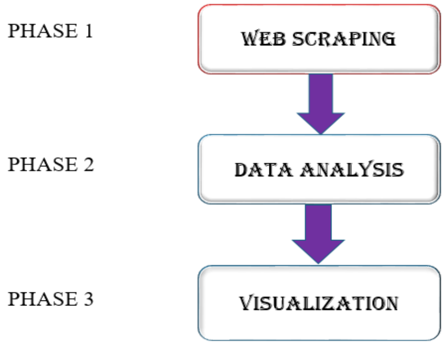


Figure 1. Methodology diagram

A. Web Scraping

The usage of Web Scrapping is to extract HTML data from URL's and use it for personal purpose. In this work python library requests and beautifulsoup4 are used for performing web scraping. To parse html pages a python library Beautifulsoap4 is used. The product details from various website are scrapped and stored in csv file with the help of the library.

There are various approaches to scrap sites, such as online services, APIs or composing own code. A few sites permit web scratching and some don't. To know whether a website permits web scrapping or not, one can come across at the site's "robots.txt" file.

Find this file by appending "/robots.txt" to the URL that want to scrape. In this paper Flipkart website is checked with the "robots.txt" file. The workflow of web scrapping is represented in the following figure 2.

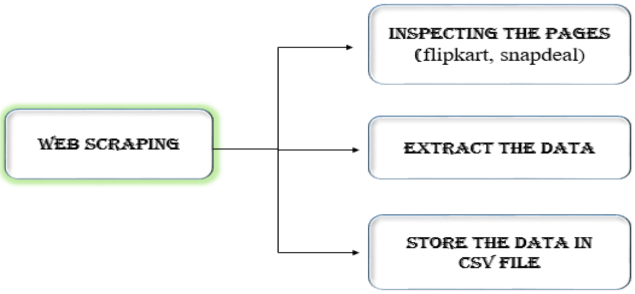

WEB SCRAPING
INSPECTING THE PAGES (flipkart, snapdeal)
EXTRACT THE DATA
STORE THE DATA IN CSV FILE

Figure 2. Workflow of web scraping

To extract data using web scraping, one needs to follow the basic steps:

Step 1: In today's online business world, there are many E-commerce websites available on internet. In this step find the URL of the required e-commerce websites that wants to be scraped.
Step 2: The data in website is usually nested in tags. The selected e-commerce website pages are inspected to find under which tag the required product falls.
Step 3: In this step using web scraping, one can extract their required data from websites by finding the data one want to extract from that website.
Step 4: Many programming languages are used to extract the data from the e-commerce websites. Here the coding is built for scraping.
Step 5: Run the code to scrap and extract only the required data from webpage.
Step 6: After extracting the data, one want to store it in a format. This format varies depending on one's requirement. In this paper, the extracted data is stored in a CSV (Comma Separated Value) format.

B. Statistical Data Analysis

Statistical data analysis is the part of data analytics. NumPy is the fundamental package for scientific computing with Python. A good understanding of Numpy will help to use tools like Pandas effectively. In this work, Numpy is used to perform operations like mean, standard deviation, count, maximum, minimum, head and tail. In this phase the statistical data analysis has been done on the csv file, where the product details of two e-commerce website is stored using web scraping.

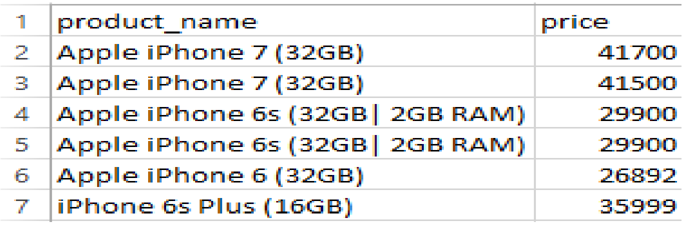
C. Visualization

Graphical portrayal of data and information is called as data Visualization. Tools like Charts, graphs, and maps helps to understand trends, outliers, and patterns in data. Nowadays we are able to gain business insights with the help of Interactive graphs and charts which are commonly used. These charts are created using data visualization tools. Exploratory data analysis (EDA) or graphical data analysis allows the analyst to:
• Examine the inter-relationships among the attributes,
• Identify the interesting subsets of the observations, and
• Build up an underlying thought of potential relationship between the attributes and the target variable.

IV. RESULTS AND DISCUSSION

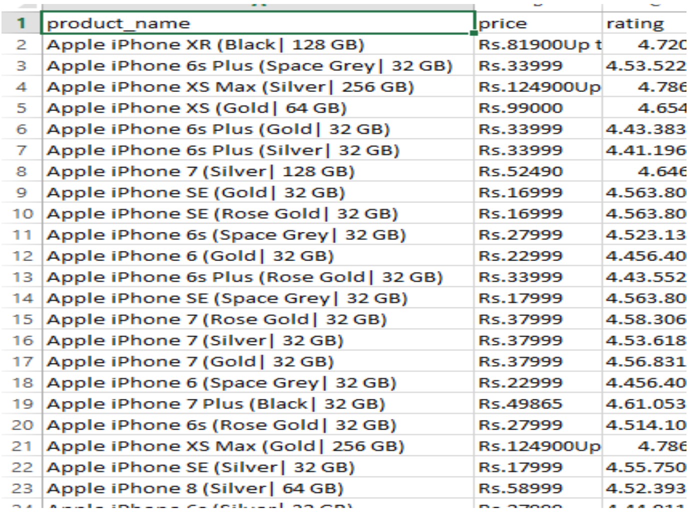
In this work, the Flipkart and Snapdeal websites are taken as the two different e-commerce websites for web scraping. The result of web scraping is the extracted details of Apple iPhone. Through web scraping only the required fields are scraped; they are product name, price and rating. Table 1 and table 2 show the scraped data from two different websites and they are stored as CSV files.

Table 1. Extraction from snapdeal website



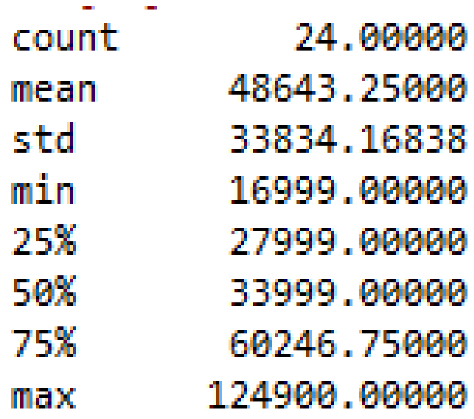
1	product_name	price
2	Apple iPhone 7 (32GB)	41700
3	Apple iPhone 7 (32GB)	41500
4	Apple iPhone 6s (32GB| 2GB RAM)	29900
5	Apple iPhone 6s (32GB| 2GB RAM)	29900
6	Apple iPhone 6 (32GB)	26892
7	iPhone 6s Plus (16GB)	35999

Table 2. Extraction from flipkart website



1	product_name	price	rating
2	Apple iPhone XR (Black| 128 GB)	Rs.81900Up t	4.720
3	Apple iPhone 6s Plus (Space Grey| 32 GB)	Rs.33999	4.53.522
4	Apple iPhone XS Max (Silver| 256 GB)	Rs.124900Up	4.786
5	Apple iPhone XS (Gold| 64 GB)	Rs.99000	4.654
6	Apple iPhone 6s Plus (Gold| 32 GB)	Rs.33999	4.43.383
7	Apple iPhone 6s Plus (Silver| 32 GB)	Rs.33999	4.41.196
8	Apple iPhone 7 (Silver| 128 GB)	Rs.52490	4.646
9	Apple iPhone SE (Gold| 32 GB)	Rs.16999	4.563.80
10	Apple iPhone SE (Rose Gold| 32 GB)	Rs.16999	4.563.80
11	Apple iPhone 6s (Space Grey| 32 GB)	Rs.27999	4.523.13
12	Apple iPhone 6 (Gold| 32 GB)	Rs.22999	4.456.40
13	Apple iPhone 6s Plus (Rose Gold| 32 GB)	Rs.33999	4.43.552
14	Apple iPhone SE (Space Grey| 32 GB)	Rs.17999	4.563.80
15	Apple iPhone 7 (Rose Gold| 32 GB)	Rs.37999	4.58.306
16	Apple iPhone 7 (Silver| 32 GB)	Rs.37999	4.53.618
17	Apple iPhone 7 (Gold| 32 GB)	Rs.37999	4.56.831
18	Apple iPhone 6 (Space Grey| 32 GB)	Rs.22999	4.456.40
19	Apple iPhone 7 Plus (Black| 32 GB)	Rs.49865	4.61.053
20	Apple iPhone 6s (Rose Gold| 32 GB)	Rs.27999	4.514.10
21	Apple iPhone XS Max (Gold| 256 GB)	Rs.124900Up	4.786
22	Apple iPhone SE (Silver| 32 GB)	Rs.17999	4.55.750
23	Apple iPhone 8 (Silver| 64 GB)	Rs.58999	4.52.393

The following figure 3 shows the result of statistical data analysis. The mean, standard deviation, count, maximum, minimum and percentile are calculated.



```
count         24.00000
mean       48643.25000
std        33834.16838
min        16999.00000
25%        27999.00000
50%        33999.00000
75%        60246.75000
max       124900.00000
```

Figure 3. Result of statistical analysis.

In this work Bar Chart is used for visualization. The following figure 4 shows the number of occurrences of the Apple iPhone price in the csv file, where the scraped data of e-commerce websites are stored. In this figure, bar chart is used to visualize the result based on price. The y axis represents the price of the apple iPhone and x axis represents the number of records.

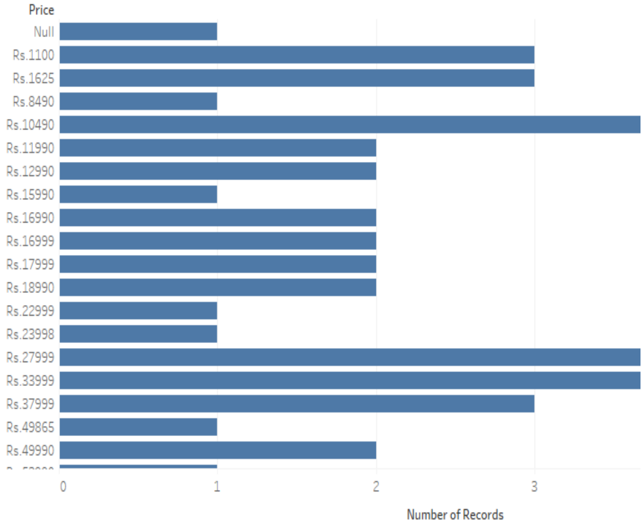
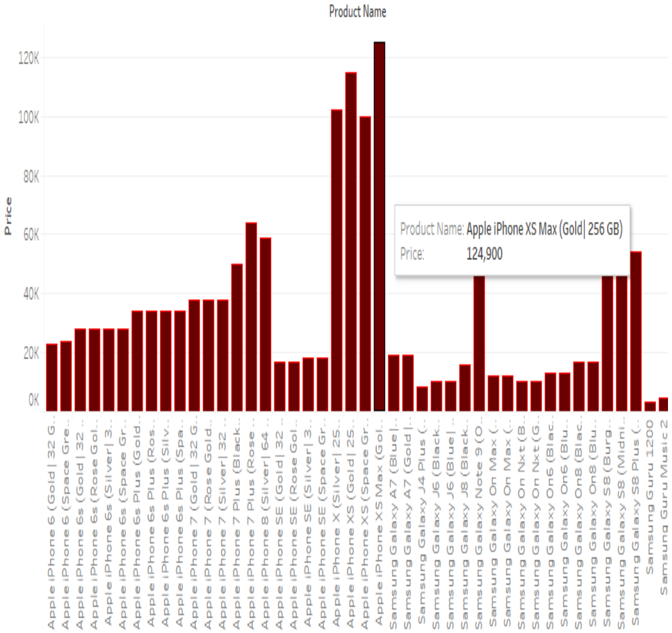


Figure 4. Bar chart result based on price

The following figure 5 shows the price level of each apple iPhone product in the csv file, where the scraped data of e-commerce websites are stored. In this figure bar chart is used to visualize the result based on product name and their price. The y axis represents the price of the apple iPhone and x axis represents the different apple iPhone product name.



.

Figure 5. Bar chart result based on product name and price.

Bar graphs are exploited to think about information crosswise over classifications. Make a bar graph by putting a measurement on the Rows rack and a measure on the Columns rack, or the other way around. In this work, the bar chart is used to visualize the price comparison between apple iPhone from two different e-commerce websites. The following figure 6 shows bar chart result based on price comparison between iPhone from two different websites and it shows that the price of apple iPhone in snapdeal is mostly high when compare to the flipkart website. In this bar chart y axis represents the price of the apple iPhone and x axis represents the different apple iPhone product name. There are two different color bars are used for comparison, blue bar represent the Flipkart price and orange bar represent the Snapdeal price for the apple iPhone product.

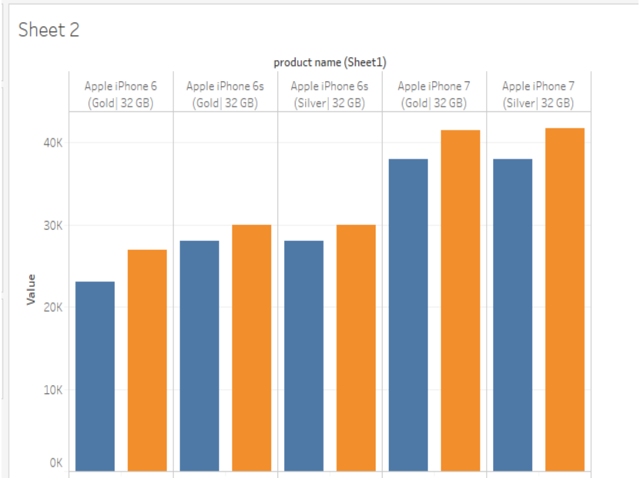


Figure 6**.** Price comparison chart of apple iphone between two websites

V. CONCLUSION AND FUTURE SCOPE

Web scraping and techniques confronting numerous difficulties as the extraction of the information are not excessively simple. These systems guarantee that the accumulated data is exact, dependable and having higher classification as the information present is in gigantic sum which is hard to oversee and keep up. In this work the e-commerce websites are scraped to extract the apple iPhone product details and the extracted data is stored in csv format. The extracted data has been visualized. The Bar Chart is used to visualize the result of the apple iPhone price comparison between Flipkart and Snapdeal e-commerce websites. The price of apple iPhone is nearly less in Flipkart when compared with Snapdeal. With the help of Customized scrapping tools the data collection and aggregation methods are easy. The developments of the tool are possible with the open source languages.

REFERENCES

[1] Eloisa Vargiu, Mirko Urru, "*Exploiting web scraping in a collaborative filteringbased approach to web advertising*", Artificial Intelligence Research, Vol. **2**, Issue **1**, pp. **44-54**, **2013**.

[2] Vasani Krunal A., "*Content Evocation Using Web Scraping and Semantic Illustration*", IOSR Journal of Computer Engineering (IOSR-JCE) Vol. **16**, Issue **3**, pp. **54-60**, **2014**

[3] Jose Ignacio Fern ´ andez-Villamor, Jacobo Blasco-Garc ´ ´ıa, Carlos A. Iglesias, Mercedes Garijo, "*A Semantic Scraping Model for Web Resources-Applying Linked Data to Web Page Screen"* in the Proceedings of the 3rd International Conference on Agents and Artificial Intelligence, Volume 2 - Agents, Rome, Italy, **2011**.

[4] Emilio Ferraraa, Pasquale De Meob, Giacomo Fiumarac, Robert Baumgartner, "*Web Data Extraction, Applications and Techniques: A Survey*", Elsevier, knowledge-Based Systems, pp. **301-323, 2014.**

[5] Faustina Johnson and Santosh Kumar Gupta. "*Web Content Mining Techniques: A Survey*", International Journal of Computer Applications Vol. **47**, Issue **11,** pp. **44-50**, **2012** .

[6] 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, Vol. **5,** Issue **5,** pp.**747-750, 2015**

[7] Anand V.Saukar, Kedar G.Pathare, Shweta A. Gode, "*An Overview on Web Scrapping Techniques and Tools"*, International Journal on Future Revolution in Compuer Science & Communication Engineering, Vol. **4**, Issue **4**, **2018.**

[8] Riya Shah, Karishma Pathan, Anand Masurkar, Shweta Rewatkar, Prof. (Ms.) P.N.Vengurlekar, "*Comparison of E-commerce Products using web mining*", International Journal of Scientific and Research Publications, Vol. **6**, Issue **5**,pp.**640-644**, **2016.**

[9] Nakash, J., Anas, S., Ahmad, S. M., Azam, A. M., Khan, T. "*Real Time Product Analysis using Data Mining"* International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. **4**, Issue **3**, pp**. 815–820**, **2015.**

[10] Rehman S.U. '*Smart agent for automated E-commerce'*, 2011 World Congress on Sustainable Technologies (WCST), **UK** IEEE pp.**124-128**, **2011.**

[11] Shikha Mahajan, Nikhit Kumar , "*A Web Scraping Approach in Node.js*", International Journal of Science, Engineering and Technology Research (IJSETR) Vol. **4**, Issue **4**, pp. **909-912, 2015.**

[12] Sneh Nain, Bhumika Lall, "*Web Data Scraper Tools: Survey*", International Journal of Computer Science and Engineering, Vol. **2,** Issue **5**, pp. **39-44, 2014.**

[13] Sarah Swain, Shriya Mishra," Prolego: A Data Science Approach to Predict the Outcome of a Football Match ", Vol. **6**, Issue **4**, pp. **132-136**, **2018**

Authors Profile

Dr.V.Srividhya has completed M.Sc., M.Phil., and Ph.D in Computer Science. She is working as Assistant Professor(SG) in the Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore. Her fields of research interest are data mining, text mining and Big Data. She has published papers in the international journals and presented research papers in international and national conferences.

P.Megala has completed BCA from Avinashilingam Institute for Home Science and Higher Education for Women. She is presently doing her final year Master of Computer Application in the same institution. Her field of interests are Data mining and Big Data.