Web Scraping and its Applications

[1]Kaushal sahu, [2]Harsh Kumar Verma, [3]Mr. Advin Manhar

[1,2]Research Scholar, [3]Proffesor
Amity University
Raipur, Chhattisgarh, India

***Abstract*- The Internet affords a vast scope of data sources established by humans. Though it consists of an innumerable assortment of dissimilar and struggling organized data, it is difficult to collect in a physical sense and problematic for its usage in mechanical processes. In recent times, various procedures and outfits have been developed to permit data gathering and alteration into organized information by B2C and B2B systems. This paper will focus on various aspects of web scraping, beginning with the basic introduction and a brief discussion on various software's and tools for web scraping. We had also explained the process of web scraping with an elaboration on the various types of web scraping techniques and finally concluded with the pros and cons of web scraping. The opportunities taking advantage of these data are numerous which shall include expanses concerning Open Government Data, Big Data, Business Intelligence, aggregators and comparators, development of new applications and mashups amongst formers.**

***Keywords*- Web Scrapping, Internet, Big Data, Business Intelligence.**

I. INTRODUCTION

Presently the internet world is enormously enormous considering the web pages with huge quantity of explanatory substances obtainable with dissimilar designs such as text, graphical, audio- video, etc. which will focus on the contradiction in repossession of facts owing to the insignificance regarding the fact user is seeing. The data that is displayed by the websites can only be viewed using a web browser. They do not offer the functionality to save a copy of this data for personal use. The only option then is to manually copy and paste the data exhibited on the website at browser into the hard drive of our computer which is quite tiresome job. This is where web scraping comes into play. Web scraping (also known as Screen Scraping, Web Data Extraction, and Web Harvesting etc.) is a procedure of automatic web data extraction instead of manually copying it. It is a technique in which meaningful data from the HTML of websites are extracted and stored into a central local database or spreadsheet. It uses the URL of the website for this purpose. It is performed by web scrapers with the help of specially coded programs. It can either be traditionally assembled for some precise website or can be one that is organized easily for working with any website. The goal of a Web scraper is concentrated on conversion of unstructured data while.

preserving in organized databases. Few Web scraping procedures are HTTP programming, DOM parsing, and HTML parsers. The data generated is later used for retrieval or analysis. It is a huge advantage as it provides us with error-free data, saves our time to give lightening quick results and stores all the data in one place. We can also choose the format in which it should be available to us. This allows an ease of access and makes life easier in analysing the data. Web scraping is presently cast-off on various aspects including online price comparison, weather data monitoring, website change detection, Web mashup, Web research and Web data integration. Further, it may be noted that Web scraping might be alongside the tenures of usage of few websites.

II. Social media sentiment analysis and eCommerce pricing

The shelf life of social media posts is very little, however, when looked at collectively they show valuable trends. While most social media platforms have APIs that let 3rd party tools access their data, this may not always be sufficient. In such cases scraping these websites gives access to real-time information such as trending sentiments, phrases, topics, etc. Many eCommerce sellers often have their products listed on multiple marketplaces. With scraping, they can monitor the pricing on multiple platforms and make a sale on the marketplace where the profit is higher.

III. Machine learning and identify goals.

Machine learning models need raw data to evolve and improve. Web scraping tools can scrape many data points, text, and images in a relatively short time. Machine learning is fuelling today's technological marvels such as driverless cars, space flight, image, and speech recognition. However, these models need data to improve their accuracy and reliability.

A good web scraping project follows these practices. These ensure that you get the data you are looking for while being non-disruptive to the data sources. Any web scraping project begins with a need. A goal detailing the expected outcomes is necessary and is the most basic need for a scraping task. The following set of questions need to be asked while identifying the need for a web scraping project:

- What kind of information do we expect to seek?
- What will be the outcome of this scraping activity?
- Where is this information typically published?
- Who are the end-users who will consume this data?
- Where will the extracted data be stored? For e.g., on Cloud or on-premises storage, on an external database, etc.
- How should this data be presented to its end-users? For e.g., as a CSV/Excel/JSON file or as an SQL database, etc.
- How often are the source websites refreshed with new data? In other words, what is the typical shelf-life of the data that is being collected and how often does it have to be refreshed?

- Post the scraping activity, what are the types of reports you would want to generate?

IV. Marketing & sales

Price intelligence data collection for every price elastic product in the market, setting optimal prices is one of the most effective ways to improve revenues. However, competitor pricing needs to be known to determine the most optimal prices. Companies can also use these insights in setting prices. Sponsored: Bright is a web scraper that can be used to extract competitors' pricing data and this is the most common web scraping use CA mentioned by most companies in the space. A web crawler can programmed to make requests on various competitor websites' product pages and then gather the price, shipping information, and availability data from the competitor website. Another price intelligence use case is ensuring Minimum Advertised Price (MAP) compliance. Manufacturers can scrape retailers' digital properties to ensure that retailers follow their pricing guidelines.

Fetching product data Specifically, in e-commerce, businesses need to prepare thousands of product images, features, and descriptions that have already been written by different suppliers for the same product. Web scraping can automate the entire process and provide images and product descriptions faster than humans. Below is an example of extracted product data from an e-commerce company website. For example, Amazon is one of the largest e-commerce companies that enables companies to analyse their competitors, generate leads, and monitor their customers. Web scraping tools help companies to extract products' reviews, images features, and stock availability from Amazon product pages automatically. To learn more about how you can leverage Amazon data for a competitive edge, check out our in-depth guide on scraping Amazon data.

V. TOOLS FOR WEBSCRAPING

A. revest

The revest platform is the workhorse toolkit. The workflow characteristically is as follows:

Reading a webpage with the usage of function read HTML () which downloads the HTML and stores so that revest can traverse it.

1. Selection of essentials we require with usage of function HTML nodes (). This function yields an HTML object (from read HTML) accompanied by CSS or Path selector (e.g., p or span) and preserve every component which matches the selector. Selector Gadget can be supportive in this aspect.
2. Extraction of constituents of nodes being selected with usage of functions like HTML tag () (the name of the tag), HTML text () (every text within the tag), html_attr() (substances of a solitary element) and html_attrs() (every elements).

The revest package comprises of certain additional characteristics like its ability in filling forms on websites and navigating websites like using a browser.

B. Regular Expressions

Frequently we will view a pattern in text which is needed to be exploited. For illustration, a novel variable might continually monitor a colon which comes after a single word in a new line. Regular expressions (or regex) specifically describe these patterns. They're very fundamental for web scraping and text analysis. In R, few regex commands can be used are:

- grep (pattern, string) which revenues a string vector and returns a vector of the indices of the string which matches the pattern string = c ("this is", "a string", "vector", "this") grep ("this", string)
- grip (pattern, string) which revenues a string vector with length n as an input returning a logical vector of length in which says whether the string resembles the pattern. Example:

```
grepl("this", string)
## [1] TRUE FALSE TRUE
```

- gsub(pattern, replacement, string) which bargains every occurrences of pattern in string and substitutes it with auxiliary Example:

```
gsub(pattern="is", replacement="WTF", string) ##  "that WTF" "a string" "vector" "thWTF"
```

VI. OPERATING STANDARD OF WEB SCRAPPER

For understanding the thought of web scraping, together with the visual lined web services, it is significant in understanding the technical working values of the technology.

Web scraping is prepared with the usage of definite techniques on the type of data to be gathered and combined. To facilitate its achievement, a sound perception of programming, web technologies like HTML, and the arrangement of web data is necessary. This requisite information and indulgence is condensed with a web scraping API.

Automated web scraping can be classified into 3 major methods which are extensively worn by web scraping software.

- Syntactic Web Scraping
- Semantic Web Scraping
- Computer vision webpage analysing

A. **Syntactic Web Scraping**

Syntactic web scraping mines information from the arrangement of website by parsing HTML, CSS, and further distinctive web languages.

Visual selectors. They can be exploited with choice of nodes. HTML nodes are provided with a group of illustrative properties specified by used browser. It is a familiar fact that humans desire in consistent web designs. Web designers thus formulate essentials of similar type to be provided with comparable visual belongings for identifying assistance. It is thus a circumstance to facilitate.

VII. STRENGTHS AND WEAKNESSES OF WEB SCRAPPING

Here we shall be centring supplementary on topic of mechanized APIs scraping through visual interface for web scraping. The two major web services to facilitate this technology are Kimono Labs and Import.io. They mutually necessitate user for creation of an account for usage of service.

A. Scraping by Coding

To start with, the web scraping offers a possibility in getting whichever data we desire in a structured method. This arrangement may be based on syntactic, computer vision or semantic abstraction technologies. An immense power of web scraping is the reality that it facilitates user in structuring data in the means to which it outfits the finest for primary project. We can fully govern the data which we do not govern. This data is the strong point of scraping which is founded on the data being figured for website audiences. These revenues the utmost precedence of web developer for keeping this data modern which delivers scraper through topmost eminence modern data. This opinion moreover enlightens the betterment of scraping data as an alternative of consuming a public API. The essential purpose of maximum web services lies in upholding the html front end for enabling their users to view. Connected to the point's overhead lies a detail which has no consistent rate restrictions for data queries. Roughly the flawless of scraping web data are the following. Foremost, an appraisal on the design of the website, or simply the retitling of definite essentials in the CSS could proceed to a failing scraper. Furthermore, we shall require programming skills for writing a scraper, and require a server for running the scraper. Finally, there is certainly no documentation about the procedure of scrapping the website.
Individually case is dissimilar and necessitates a practice for manufacturing a scraper.

B. Scraping with Visual Interfaced Services

While comparing the visual strengths and weaknesses with mechanized APIs for instance Kimono or Import.io, it is informal for starting by means of the weaknesses and explaining the assets of automatic web scrapers. For using Kimono or Import.io the user may not require any expertise in programming. The graphic CSS element collection efforts healthy in mutually services with a proposal to user for creation of whole working API lacking a little programming knowledge. It similarly empowers users in generating a web scraper shorn of without requiring any server as everything goes online. A huge benefit lies on the fact that the resultant API shall use typical API assemblies in several layouts, thus the data will be distributed smoothly through extra inventors. A supplementary service which can be obtained from Kimono and Import.io lies on the statement on disastrous apprises of the API signifying the needlessness in checking whether our application is functional. The detail which the user is delivered through these facilities implies the resources which can be relied on third party services for providing the fed data. It is very significant for keeping in mind about the fact that presently both are in beta rank. It also revenues of the fact that its usage and value can alter in the forthcoming days. Likewise, there can be few rate restrictions, somewhat like a physical scraper which has not been come across. Finally, it is significant in realizing that if we have the necessary skills, we can build our individual scraper shall continuously be an additional customization than some of these services. If we require in scraping data which is not effortlessly recognized via CSS, such as script produced textboxes change based on context of the position of our mouse location which has the potency to be healthier in writing our own scraper.

VIII. CONCLUSION

%Web scraping %has a long history with important modern applications. Numerous professionals and researchers need "free" data, while people dealing with business-to-business scenarios require the admittance of data from several sources. We have reviewed the various aspects of %Web scraper %tools and software starting with the operating principle, strength, and drawbacks and finally seen how they are used in applications.

REFERENCES:

1. Osmar Castrillo-Fernández, "Web Scraping: Applications and Tools", European Public Sector Information Platform Topic Report No. 2015 / 10, December 2015.
2. Kanehisa M, Goto S, Sato Y, et al. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 2012;40: D109–14.
3. Glez-Pen‹a et al., "Web scraping technologies in an API world", Briefings in Bioinformatics Advance Access, doi:10.1093/bib/bbt026, published April 30, 2013
4. http://jsoup.org/
5. http://adamsoft.sourceforge.net/
6. https://nutch.apache.org/
7. https://lucene.apache.org/solr/
8. James, G., Witten, D., Hastie, T., Tibshirani R. (2013), An Introduction to Statistical Learning with Applications in R, Springer Texts in Statistics
9. Giulio Barcaroli et Al, "Use of web scraping and text mining techniques in the Istat survey on "Information and Communication Technology in enterprises"", European Conference on Quality in Official Statistics (Wien 2014), June 2014
10. Grimmer, Justin. 2013. Representational Style in Congress: What Legislators Say and Why It Matters. Cambridge University Press.
11. William Marble, "Web Scraping With R", stanford.edu, August 11, 2016
12. Carlos A. Iglesias Mercedes Garijo Jose Ignacio Fernandez-Villamor, Jacobo Blasco-Garcia. A Semantic Scraping Model for Web Resources, Applying Linked Data to Web Page Screen Scraping.
13. Muntasir Mashuq MichelZiyan Zhou. Web Content Extraction Through Machine Learning.