International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

PLeveraging Django and Redis using Web Scraping

K. M. Anandkumar, Abhinav R, Abhinav Raman, Abilash R

Abstract: ***Web scraping is also known as data scraping and it is used for extracting data from sites. The software used for this may directly access the World Wide Web by using the Hypertext Transfer Protocol or by using a web browser. Over the years, due to advancements in web development and its technology, various frameworks have come in use and almost all of websites are dynamic with their content being served from CMS. This makes it tough to extract data since there is no common template for extracting data. Hence, we use RSS. Rich Site Summary is a kind of timeline allowing users and also applications to gain access to the updates on websites in a standardized, computer-readable format. This project combines the use of RSS to extract data from websites and serve users in a robust and easy way. The differentiation is that this project uses server side caching to serve users almost instantaneously without the need to perform data extraction from the requested site all over again. This is done using Redis and Django.***

Keywords: python, Django, redis, RSS

I. INTRODUCTION

Web pages except for core contents are also composed of other elements namely banners, copy right information, external links, navigational elements etc. Extracting or taking out relevant information from the data which is of semi-structured is of use for several purposes, namely document, identification, data extraction, news article etc. There are two types of scaling in the current system for workloads, Namely, horizontal scaling and vertical scaling. The above two scaling prove to be very expensive as the number of users increases. Since web scraping is a process intensive task, the system may slow down when there are many user requests. To avoid these issues we use server side caching for scraping data by using Redis instead of the scaling methods. The priority for the users is given by queue and queuing is implemented by celery, thus making the system cost efficient and robust. This technique proves to be very valuable for startup companies.

Manuscript received on April 02, 2020.
Revised Manuscript received on April 15, 2020.
Manuscript published on May 30, 2020.
* Correspondence Author
Dr. K. M*. Anandkumar, Professor, Department of Computer Science and Engineering, Easwari Engineering College, Chennai, India.
Abhinav R , UG Students, Department of Computer Science and Engineering, Easwari Engineering College, Chennai, India.
Abhinav Raman, UG Students, Department of Computer Science and Engineering, Easwari Engineering College, Chennai, India.
Abilash R, UG Students, Department of Computer Science and Engineering, Easwari Engineering College, Chennai, India.

[1]Belen Vela, Jose, Ploma and Carlos developed the framework for extraction of information to give accessibility for urban people. This paper specifically focuses on the extraction associated process of the present data regarding conveyance and its accessibility for the generation of an open data repository within which to store this information. This technique permits the extraction of conveyance information and also the existing accessibility data from a specific website. The advantage delineated through this paper is that it provides users a convenient means of accessing the data for individuals with special quality needs. The problem arises once winding up the programming task manually as it could be a terribly time intense task. So the process of internet scrappers in the general transport is to be machine controlled so as to cut back the programming effect needed.

[2]K.Sundarmoorthy and R.Durga developed a strategy to dedicate recent information in simpler words .The contents of the rich site summary is Uniform Resource Locator is fetched by the crawler. The application developed by them derives information from the news websites and seperates them in a platform. The projected system implies gathering all the recent news using an RSS aggregator and displaying them beneath one complete roof .The advantage of this paper is it gets a unique reading experience using its straight forward UI. It gathers clean news story from three websites while not having the requirement to manually copy and paste the article .The main issue of this method is that it provides solely source uniform resource locator of the news. In this method, it simply displays the name of the digital newspapers on the market within the net. The user has got to head to every and each link to browse the news. Thus it is a time intense method.

[3]Deborah and Deny Triawan conducted a survey on data extraction and abilities on e-commerce websites. The system uses web scraping that makes use of simple Direct Object Message as tag Hyper Text Markup Language required for the website. The program of web extraction starts by compiling the Hyper Text Transfer Protocol request to retrieve resources by the online sites. This request can be given in a Uniform resource locator that has a GET request consisting of a POST query. Going through one by one of the information resources and comparing data from every information sources visited will increase much more time to the rediscovering the information. It takes an approach that can put together information from several sources into a separate entity to continue the process of information retrieval. The proposed system is composed of three e-commerce websites as information sources.

It does web scraping techniques on web browsers or search engines by using hypertext preprocessor and my SQL database is used to accumulate the search results. Since the precision rate is 93.9% the system provides results which are not accurate. Thus the success of the information retrieval is relatively very low.

[4]Achmad, Windi Eka and Muhamat Abdul developed a technique of web scraping based on sequence of characters that define a search pattern. This technique has three steps i.e analyzing web structure news, pattern of regex construction and implementing it as a set of rule in web scraping. Two different types of patterns are used i.e content pattern(original text article of news is extracted) and filter pattern(non -news elements are eliminated).This system is based on general web scraping approaches by retrieving general text from HTML page .However, less different from those works, the proposed system focuses on getting five different elements by giving title, publication date, author, clean text article and URL address of news article from HTML page of websites .From the final result, we may know that each news website has a special layout to present their news .Every news website provided a unique HTML element for the article link, title, author, and publication date .These news elements could be retrieved by simple means by providing a corresponding Regex for each. However, this approach may not be implemented on news article content as it has a few non-news elements in every news website .Thus we have to give with two kinds of Regex, Regex for filtering and Regex for extraction.

[5]Sandeep Sirsat and Vinay Chavan conducted a survey on pattern matching from news sites. Though there are several available existing strategies that formulate the particular content identification downsides as a DOM tree node choice downsides, everyone has some kind of lacunae. The planned approach is predicated on pattern matching technique. This method uses straight forward heuristic for extraction of core contents from sites that are principally semi-structured in nature. This approach additionally uses devised algorithms that applies regular expressions (Regexes) to spot the right pattern for extracting the particular text contents from these news documents. The advantage is that it deals with news sites of any size and extracts core contents with high potency and high accuracy. This approach does not exploit the options of DOM structure. It's template freelance and isn't hooked on to any tag sort. It utilizes restricted top down mapping (RTDM) that supports post order traversal of trees. But the disadvantage is that it supports the ambiguous assumption that the new website content may be split into teams that share common format and layout characteristics. Therefore it is not appropriate to use RTDM approach for the news websites having heterogenous structure and page layout.

[6]Li Zhao and Si-Feng Du designed a paper supported content management system. Web site content management system is a foundational web site application platform for web site style and data distribution and is an auxiliary tool system for web site development. This paper proposes the key methodology within which the system info distribution module is made within the webpage component and guide manner to totally scale back the quality of the system style. This software system is outlined supported the open design and object minded methodology and is developed by victimization java primarily based on pure object minded framework. The guide will mix the contents with the pages within the style to manage web site content and mechanically generate the web site. Correcting a page event will simply update the total web site.The guide will embrace freelance page components of any webpage. Benefits within the system guarantees the superb cross platform capability victimization pure Java. The system will run on Windows and Linux and might migrate to the operating system massive scale machine. The disadvantage of the positioning is that it contains sizable amount of files and will leave the files at risk of errors. Restricted flexibility in style is additionally a serious issue.

[7]Shreya Upadhyay and Vishal Pant constructed a web scraper for massive data extraction. There exist several automated techniques to information retrieval from the web. A lot of these methods are ad-hoc and domain related. The system makes an effort to show the advantages of automatic information extraction tools and the roles of them as a significant component in the improvement of knowledge based systems. The requirement for a robust, automated, simplified framework for retrieval of content from the web with vey less human effort possible across domains are enticing. The architecture developed by the authors for a web scraper gives the gap to harvest information from the web. Advantages: The framework gives a feasible and simple approach for parsing and retrieving data on a large scale from several online sites with very less human interaction. Few of the special aspects of the design structure are the simplicity of operations, adapting to various range of domains, applicability to a wide range of popular file formats and providing data instantaneously. The issues faced are: The statically loaded content is easy to mine but dynamically generated scripting tools like java script have its own problem. Such a system has no value for organizations operating at the enterprise level or for research institutions by giving unprecedented control to volumes of different data.

[8]David Mathew Thomas and Sandeep Mathur analyzed web scraping using python. The web scrapers conniving ethics and procedures are juxtaposed, it tells how the operation of the scraper is premeditated. The method of it is seperated into three fragments: the web scraper takes the required links from the web and then the information is extracted to get the data from the native links and in the end stowing that information into a csv file. The operation is carried out by using the python language. The purpose of the paper is to take away the information from several sources with the help of programming called the web scrawler scrape making use of the programming language python 3.6. All the unstructured data from various sources are collected into a database and then analyzed by the analytic technique of its specifications, assembling, organizing, cleaning, re-analyzing, applying models, algorithms and giving the required results.

II. PROPOSED MODEL

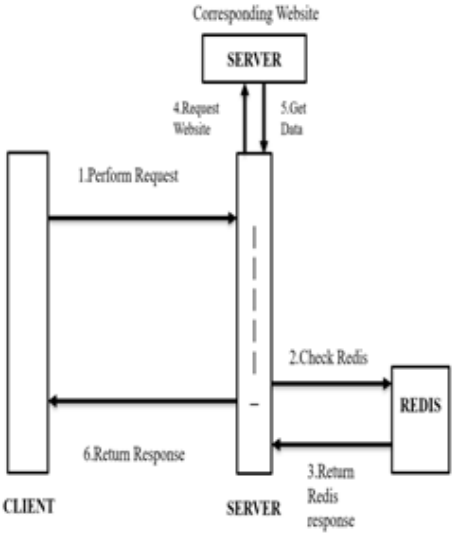

Fig 1: System Architecture

The above diagram is the flow of the proposed system where the client requests the server to access a website. The data is structured through Redis which deals with very difficult programming situation and problems with easy commands executed within the data store. The requested website is accessed through the server and the response data is sent back to the client.

III. EXPERIMENTAL SETUP:

The proposed system has the following modules to achieve the result

A. List of modules:

- Web scraping
- Django with Redis
- User Interface Implementation

B. B. Web scraping:

In this module website URL's to be extracted is given as input by the client. All the websites in the modern world uses content delivery network (CDN) which makes getting contents of a webpage cumbersome since they are either encrypted or point to other URL's. Due to this reason every website provide Rich Site Summary (RSS) or XML template with all their content .Hence we use BeautifulSoup, a package for parsing HTML, XML responses.

C. Django with Redis:

In this module the processed and cached data from corresponding website URL's is acquired and returns a response to the user. If another user wants a summary of the same site, instead of pinging the corresponding site, we retrieve it from our cache. Django is a python framework while redis is an in-memory data structure that supports various kinds of data.

D. User interface:

The UI provides a means of communication between the two interfaces which provides simple statistics on how long it takes to process the request and how many words were used totally on the webpage.

IV. RESULT

The system uses a technique of web extraction in order to take data from online sites as requested by the user. Most existing systems are inefficient since many users access a website at a time. This makes it difficult to retrieve information thereby decreasing the rate of success.

V. CONCLUSION

The proposed system uses server side caching for scraping data using Redis which is an inbuilt memory data structure that supports various kinds of data. The priority for the users is given by queue and the queuing is done by huey or celery. This approach makes the system robust, easy to access and cost efficient.

REFERENCES

1. Belen Vela, Jose, Ploma, "A Semi-Automatic data scraping method for the public transport domain", IEEE access, vol.7, no.10, pp. 335-339, 2019.
2. K. Sundarmoorthy, R. Durga, "An aggregation system for news using web scraping method", IEEE International conference on Technical Advancements in Computers and Communications (ICTACC), pp.1340-1343, 2017.
3. Deborah, Deny, "Increased information retrieval capabilities on e-commerce website using scraping techniques", IEEE International conference on sustainable information engineering and technology (SIET), pp. 829-834, 2017.
4. Abdul, Windi Eka, Muhamat Abdul, "An approach on web scraping on news website based on regular expressions", IEEE 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT), pp.906-923, 2018.
5. Sandeep Sirsat, Vinay Chavan, "Pattern matching for extraction of core contents from news web pages", IEEE Second International Conference on Web Research, pp.51-54, 2016.
6. LI Zhao, SI-Feng Du, "Design and implementation of website content management system", IEEE International Conference on Information Management and Engineering, pp.5-8, 2018.
7. Shreya Upadhyay, Vishal Pant, "Articulating the construction of a web scraper for massive data extraction", IEEE Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp.1-3, 2017
8. David Mathew Thomas, Sandeep Mathur, "Data analysis by web scraping using python", IEEE 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp.6179-6186, 2019.

AUTHORS PROFILE

Abhinav R Pursuing BE at Easwari Engineering College. Doing my 4th year of computer science and engineering .I am interested in latest technologies like machine learning, Blockchain

Abhinav Raman Pursuing BE at Easwari Engineering College. Doing my 4th year of computer science and engineering .I am interested in latest technologies like machine learning, Human Computer Interaction

Abilash R Pursuing BE at Easwari Engineering College. Doing my 4th year of Computer Science and Engineering. I am interested in latest technologies like machine learning, Blockchain

Dr. Anand Kumar M.Tech, Ph.D, Professor, Department of Computer Science and Engineering department, Easwari Engineering College