International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-07, Special Issue, MAY 2021

Design and Implementation Of Domestic News Collection System Based On Python

[1]Prof. Gayatri Naik, [2]Mr. Hamsaraj Pitani, [3]Mr. Md.Ali Kuwari, [4]Mr. Vaibhav Nandurkar
[1]Asst.Professor,[2,3,4]UG Student,[1,2,3,4]Computer Engg. Dept. Shivajirao S. Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India.
[1]krishngita123@gmail.com,[2]kuwarimohali2000 @gmail.com, [3]hamsarajpitani@gmail.com, [4]vaibhavnandurkar123@gmail.com

Abstract- In this period of quick advancement of the Internet, network media has become another window for individuals to comprehend the rest of the world because of its speed and wide spread. News is a vehicle for individuals to think about Information, however large number of information are delivered consistently on the Internet , these news are required or not. How to precisely acquire the news content from the website is a great requirement in people's life. This system aims to collect news on specific websites and give it to users with concise and clear pages. This system crawls and processes the domestic financial news content, which is convenient for people to consume the information. To stay away from unnecessary news and the advertisements, In the particular execution, the framework is composed utilizing Python related to the scrapper structure and Django system, which can work on the framework partly. The practical value of the framework lies in the opportune and efficient with advantageous admittance to home grown news that individuals care about with need and interested in it.

Keywords –Python, Domestic, News Collection.

I. INTRODUCTION

The deep web is also called unvisible web. The deep web may have valuable contents. at University of California, Berkeley, it is estimated that it contains approximately 91,850 terabytes and the surface web is only about 167 terabytes in 2003. Deep web makes up about 96% of all the substance on the Internet, which is 500- 550 times bigger than the surface web. The contrary term to the profound web is surface web that can be easily seen by a search engine like google bing duckduckgo. The profound web is comprised of all scholarly data, clinical records, logical reports, government assets and some more. The deep web databases not register with any search engines since they change ceaselessly thus can't be effectively ordered by a web tool.

Subsequently, to find the profound web or covered up web contents and it needs web crawler. The size of deep web is increasing very rapidly now a days on a daily bases. The use and the structure of the web is changing on daily. Old data is getting outdated and new information is being added. The existing approach lack to efficiently locates the profound web which is covered up behinds the surface web. In this way, the need of the unique crawler emerge this paper proposed an engaged semantic crawler. The proposed crawler works in two phases, first it gathers the provided site and second stage is in-site investigating.

II. AIMS AND OBJECTIVE

a) Aim

This system aims to collect news from the specific websites and return it to the users with concise and clear pages. This system crawls & processes the domestic financial news content which is convenient for people to process the information's. To keep away from the duplication in the data, the framework has likewise executed a self-characterized de-duplications rule in it In the particular execution, the framework is composed utilizing Python language with the assistance of Scrapy structure and python Django system, which can work on the framework code somewhat. The viable estimation of the framework lies in the ideal and productive and helpful admittance to home grown monetary news that individuals care about and are keen on.

b) Objective

The primary goal of this paper is to develop a web app for Online News Paper website that can aware the peoples and to provide the daily based news and the top breaking news. Utilizes the different and unique advancements to get the required oriented information more quickly and easily and attractively. To do this more widely coverage of distribution & faster dissemination of information in a timelier way. At Whenever any place, anybody can know about the top news or information by internet at very low cost. Dynamically

provides facility. Fetches new information without complexity with latest news every day with real-time database refreshing.

III. LITERATURE SURVEY

Paper 1:

Design and Implementation of Domestic News Collection System Based on Python.

This paper studies designs and develops a convenient automatic news-gathering system. The system uses crawler analysis to collect domestic news, saves it after de-duplication, and finally provides news services for retrieving and viewing. It can help users find similar news and extract hot news that users are interested in, and improve the efficiency of reading news. This system aims to collect news on specific websites and return it to users with concise and cleared pages. . This system crawls & processes the domestic financial news content, which is convenient for people to process the information. To keep away from advertisment and pop-ups and all the aditional redirection of website with decline the user expereince .the used framework has its own self-defined rule which does not fetch Ads. In the particular execution, the framework is composed utilizing Python related to the python scrapy structure and Django system, which can work on the framework code to a limited degree. The viable estimation of this framework lies in the ideal, proficient and the helpful admittance to homegrown monetary news that individuals care about . In the particular execution, the framework is composed utilizing Python related to the Scrapy structure and Django system, which can improve on the framework code to a limited degree.

Paper 2:

Configurable News Collection System Based On Web Crawler

This paper uses web crawler technology such as regular expression and Xpath, web page analysis, and Web Magic crawler framework to realize a set of configurable news data collection system based on java. The system can realize the function of data capture, information extraction and the storage of news. The system owns high configurability. It can crawl multi-source news data. Web crawler provide important theoretical support for collecting news. It is not difficult to use web crawler to get news data from the Internet. The problem with crawlers is that users need to implement a new crawler and cannot extend the existing crawler module to reuse it in a specific scene when they want to crawl a particular page in a particular site. In other words, the crawler is not highly customizable . The Web Crawler acquires the comparing content by examining the URL address and afterward measures information. Normal articulation is the intelligent articulation for string coordinating with that utilizes some particular characters characterized ahead of time and the mix of these particular characters to shape a 'rule string' to discover or coordinate with the fixed-design text .

Paper 3:

Design and Implementation of News Collecting and Filtering System Based on RSS

The paper uses the Web information collection technology to obtain the news theme information from the RSS website. Data storage structure of the news information acquisition system was designed and the news extracting and processing model was built in it. Finally, the news information acquisition and processing system is achieved. The system able to realize real-time information acquisition and information personalized display. It helps receiving and personalized displaying information in a convenient, efficient and real-time way. By combining the characteristics of RSS and web news processing technology, an automatic news information acquisition and filtration system based on RSS was designed. On the base of technology integration of RSS and Web news collection, it realized the network information received and used in the stability, high efficiency, real-time, safely method.

IV. EXISTINGSYSTEM

Freezing the requirement before designing and the coding can proceed, a throwaway prototype is built to understand all the requirements. This model is created dependent on the current known prerequisites. by utilizing this model, customer can get a Real encounter of the framework, since the collaborations with the model can empower the customer to improve comprehension of the prerequisites of the ideal framework.

The Waterfall model in programming assumes a significant part in testing the product obviously, throughout the long term, there are various other programming measure models which have been planned and carried out, yet what is genuine is that a ton of them depend on guideline of the cascade model.

V. COMPARTIVE STUDY

SR NO.	PAPER TITLE	AUTHOR NAME	METHOD	ADVANTAGE	DISADVANTAGE
1.	Design and Implementat ion of Domestic News Collection System Based on Python	Haixia Ly	Scrapy Framework, Django Framework, Data Analysis and Processing Technology	It optimized to make the interface more concise and intuitive	Time Consuming
2.	The Design and Implementat ion of Configurabl e News Collection System Based On Web Crawler	Mengmeng Lu, Shuhong Wen, Yan Xiao, Pei Tian ,Fang Wang	Web Crawler 'rule string' to find or match the fixed-pattern text	Good Approach Explained	Difficult to understand
3.	Design and Implementation of News Collecting and Filtering System Based on RSS	ZHENG Rui-juan , ZHANG Yang-sen	RSS feed	Good Approach Explained	Time Consuming

VI. PROBLEM STATEMENT

System can handle only structured data and cannot handle unstructured data. Human intervention is required. Large data sets are required, which may not be available. Comparative less effective in feature detection.

VII. PROPOSED SYSTEM

The proposed web crawler utilizes cosine comparability calculation. .At that point the crawler will bring some pertinent and the superfluous the given URLs from web search tool like google. It at that point applies stop word expulsion and stemming measure on those URLs. After that the title and the description coordinating of the appropriate URLs is done. On the off chance that the page contains more pertinent URLs, a similar cycle is rehashed for the profound pursuit. After that the crawler will apply Cosine Similarity calculation and gives the desired result.

VIII. ALGORITHM

step1 : Start

step2 : Get all the required url and store i

step3 : retrive the data from the uRL by

making request

r = request.get(url);

step4 : Parse the require content with

<tagname> to crawl the respective

data

data = soup.find('<tagname'>

step5 : Aplly limitation the the data

fetching

data[0:LIMIT]

step6 : Repeat step 2 to step 4 to crawl more urls

step 7: END.

IX. MATHEMATICAL MODEL

To Calculate cosine score There are two vectors.

1 Crawler search query

2 Using this score URL documents are fetched. Use the count to calculate

$$Recall = \frac{tp}{tp + fn}$$

$$Precision = \frac{tp}{tp + fp}$$

$$sin(q,d) = \frac{V(q).V(d)}{|V(q)|vV(d)v}$$

X. SYSTEM ARCHITECTURE

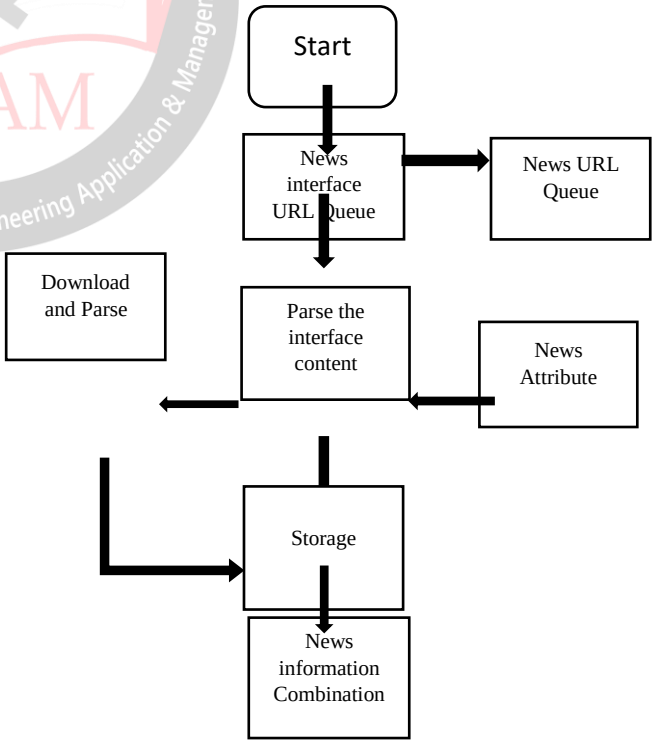

Fig no.1: System Architecture

Description:

1) Take the news URL given to scrap a crawler entry URL queue, start the crawler process.

2) Parse the HTML data obtained from the given URL. The parsing content includes the headline of the latest top trending news

3) The news body URL obtained from interface parsing will be stored in the URL queue, then the crawl will continue.

4) extract the HTML document from the news body URL and render the data such as news headlines

5) Combine all the news information obtained in the above steps to a containing allnews attributes and store in local.

XI. ADVANTAGES

Users are able to the get all the required and necessary information with from the multiple sources only at a single place. In this methodology a working model of system is provided. Users can experience advertisement free news and consumption it in efficient manner in very easily way .

XII. DESIGN DETAILS

Fig no.2: Webpage (TOI)

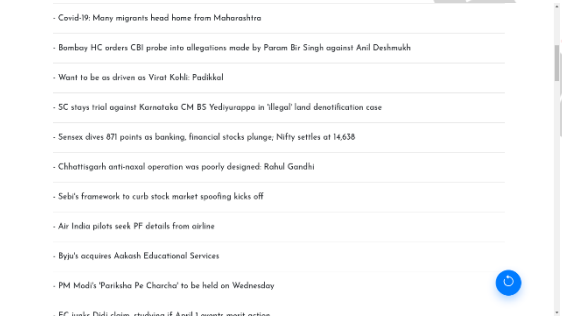

Fig no.3: News Collection

Fig no .4: Indian EXPRESS

XIII. CONCLUSION

Thus we have tried to implement the paper "Haixia Lv", "Design And Implementation Of Domestic News Collection System Based On Python" published in IEEE 2020 Web crawler is an important way to obtain data from the Internet. This paper designs a set of configurable news collection system based on web crawler, which can crawl news from target news website. It can crawl a variety of multi-source data and the crawler is customized highly. In addition, it can do corresponding processing to crawled news content in accordance with need. The system not only reduces the difficulty of news editors, but also updates the news content in database real-time, improving the efficiency of news gathering and publishing.

REFERENCE

[1] J. L. Zhang, "Design and Implementation of Intelligent News Collection and Processing System," Shandong University, 2017.

[2] G. M. Yu, "Big data method and innovation in news communication:

From theoretical definition to operational route," JAC Forum, vol. 266,

No. 4, pp. 5-7, 2014.

[3] S. Q. Long, Z. W. Zhao, H. Tang, "Chinese word segmentation

Algorithm review," Computer Knowledge and Technology, vol.5, no. 10, pp. 2605-2607, 2009.

[4] J. F. Hu, Y. B. Shen, "Web-based news gathering system," Computer

Knowledge and Technology, vol.5, no.19, pp. 5111-5113, 2009.

[5] H. C. He, "Research and Implementation of Information Collection

Technology in web mining

[6] H. Zhang, "Keyword extraction algorithm based on automatic text Classification. Computer Engineering," vol. 35, no. 12, pp. 145-147,2009.

[7] L. W. Sun, G. H. He, L. F. Wu, "Research on Web Crawler Technology," Computer Knowledge and Technology, vol. 6, no. 15,pp. 4112-4115, 2010.

[8] Sharma Kartik; Aggarwal Ashutosh; Singhania Tanay; Gupta Deepak;

Khanna Ashish (2019). Hiding Data in Images Using Cryptography And Deep Neural Network. Journal of Artificial Intelligence and Systems, 1, 143–162.

[9]M.Saravanan and A. Priya (2019). An Algorithm for Security Enhancement in Image Transmission Using Steganography. Journal Of the Institute Of Electronics and Computer,1,1-8.

