

A Novel Approach for News Extraction Using Webscraping Technique

Shreesha M, Srikara S B , Manjesh R*

Department of CSE, Srinivas Institute of Technology, Mangaluru,

DOI: <https://doi.org/10.21467/proceedings.1.56>

* Corresponding author email: manjuis024@gmail.com

Abstract

News's take major role in human life, Intelligent Smart Parser a web based application which display's the top news's and regional news from the different news sites on a single location using the web scrapping technology. Intelligent Smart parser also summarizes the entire news into three to four lines of code using the natural language processing technique. Intelligent Smart Parser also display's the news based on user location and Trending news based on views of the news.

Index Terms- Smart Parser, Beautiful soup, Natural Language Processing.

1 INTRODUCTION

Intelligent Smart parser is useful system which can be used to get the top stories and latest news from the top news sites. The system will display's top stories from the different news site's and regional news based on user's location from these sites. This system also provides live updating cricket score summary. User can select any site from option and read the top news and latest news from that site and even visit actual page also. System also provides the regional news on the basis of user location. User can register into the system and also make any site as default which is completely optional. Web scrapping is the technology which is used to get the content from those news sites. This system will restructures the data and displays to the user.

Currently someone needs to read the news then they must visit the particular websites such as Times of India, India Today, Deccan Herald, and The Hindu. Since the website contains huge amount of data it takes some time to load. In advance to the websites are popped up with plenty of advertisements. That is the some of the issues with the current system. Even if someone wants to read news from two websites then they must visit two websites which is time consuming and consumes the internet. In addition, if regional news is to found then they must search regional news which are again popped up by plenty of advertisements. So, we propose a system which display's the news from multiple news paper in single application, also this system provides the news which is regional to the user.



© 2018 Copyright held by the author(s). Published by AIJR Publisher in Proceedings of the 3rd National Conference on Image Processing, Computing, Communication, Networking and Data Analytics (NCICCND 2018), April 28, 2018. This is an open access article under [Creative Commons Attribution-NonCommercial 4.0 International](https://creativecommons.org/licenses/by-nc/4.0/) (CC BY-NC 4.0) license, which permits any non-commercial use, distribution, adaptation, and reproduction in any medium, as long as the original work is properly cited. ISBN: 978-81-936820-0-5

2 LITERATURE SURVEY

Web scrapping is the very useful technology in the field of getting the content from the different websites. The best features of the web scrapping technology are that can scrape the content which is required. Web scrapping is used by the many companies for business. One of the example of the web scrapping in the real estate listing gathering, It is a huge and growing web scrapping area. This is an area where the businesses are using web scraping to gather already listed. All the machine learning companies are using web scrapping to get the data. Email address gathering is another field of the application where once the emails are collected bulk emails are sent.

Website creators also uses the web scrapping where collecting data from the different social media websites, what is trending and what is in etc. Web scrapping is used in the one of the project in which it is used to scrape the content of particular category of book in the Amazon store. In another project web scrapping is used to scrape the contents from the Twitter on the basis of hash tag or by searching the keywords in the twitter. In the field of machine learning web scrapping is used in sentiment analysis field, where the data is scrapped from the websites.

Web scrapping is used in technologies such as Market research using web data in any of the industry. Even web scrapping technology is used in price comparing sites where it compares the price of item or room from different websites. In advance these applications use the web scraping to the scrape the content from the dependant websites. Various government and private watch dogs uses the web scrapping to monitor the malicious activities going on the internet.

Netucon Company based at Ahmadabad provides ultimate solutions to its customers and software development services with innovation and creativity. Netucon understands the requirements from the customer's and clients and produces the software. They developed a LinkedIn connection creator this LCC is useful for scrapping CEOs, creating B2C contacts, Lead Generators, Digital marketers, Blogger who post their blogs on LinkedIn and so on. News hunt is another application which displays the news from different websites. News hunt uses the web scraping as one of the backend technology to hunt the news. News hunt provide news's in 17 different languages.

3 PROPOSED IDEA

In the smart parser top news and latest news from the top news sites displayed using web scrapping technologies. The system scrapes the top news and latest news from the top news sites and displays it to the user. The home page of the application will be displaying the top news from different sites. A user menu will be given and from the menu user can select any of the site and read the news from that site and also in addition to the this user can register to the system and select any one of the news website as default and in case the top news and latest news from that sites are displayed on home page is completely optional. Web scrapping is the

major technology used to develop the system. The special attraction of the system is that of regional news that is news based on user's location is also scraped and displayed in the application. For example consider the fig 1. , which contains all the information of the given website including top news. From this page the system takes only the top news by using web scraping technique.



3.1 Implementation

The implementation part of the system starts by taking the url of the website. System scraps all the links of the top stories. After taking the link of each top stories system scraps the content of those links. System keeps monitoring in the change in the contents of the site. System can be implemented using a python library known as beautiful soup. Beautiful soup can also be used to implement the monitoring of changes. System scraps the contents from the site using beautiful soup library, html tags, css classes and ids. Finally, the system restructures the data and displays it to the user. The summary of the scraped content is generated using natural language processing toolkit.

4 EXPERIMENTAL RESULT:

System scraps the top news from the site and displays in a restructured format. The fig 2. Shows the result of the proposed system. The system updates automatically for any changes in the sites. Also the summary of the each news is displayed.

Top news

At least 10 govt websites hit by glitches in hardware, data storage india_today 2018-04-07 11:39:36 ago	Amit Shah sounds 2019 election bugle, calls Opposition animals who unite during storm india_today 2018-04-07 11:39:35 ago	Indian Air Force all set to show its mettle in war exercise Gagan Shakti india_today 2018-04-07 11:39:35 ago	Man catches wife cheating, uses meat chopper to dice her lovers body india_today 2018-04-07 11:39:34 ago	#Vacay: The one treatment for your March stress india_today 2018-04-07 11:39:33 ago	Indrani Mukerjea rushed from Jail to JJ Hospital, admitted in ICU india_today 2018-04-07 11:39:33 ago
Salman Khan given VVIP treatment in Jodhpur jail india_today 2018-04-07 11:39:32 ago	Rohtak girl repeatedly slaps traffic cop who harassed her in shared auto india_today 2018-04-07 11:39:32 ago	After 2 nights in jail, will Salman Khan get bail today? Top 5 updates india_today 2018-04-07 11:39:31 ago	Judge who sent Salman Khan to jail in blackbuck case among 87 transferred in Rajasthan india_today 2018-04-07 11:39:31 ago	When Shah Rukh Khan backed Salman over his criminal accusations. Watch viral throwback india_today 2018-04-07 11:39:30 ago	CBI alerts airports. Chanda, Deepak, Venugopal cant leave India without informing authorities india_today 2018-04-07 11:39:30 ago

5 CONCLUSIONS AND FUTURE WORK

The system is able to fetch only the required contents among the huge data. The study of each website structures is a major part of this scraping technique. The system can also be modified to scrap the content based on the type of the news. The system can also update to display the news based on the user's location. Trending news can be highlighted according to the number of views.

References

- [1] Ram Sharan Chaulagain, Santosh Pandey, Sadhu Ram Basnet Cloud Based Web Scraping for Big Data Applications, *Smart Cloud (SmartCloud) IEEE conference*, 2017
- [2] The information from glowing python -<https://glowingpython.blogspot.in/2014/09/text-summarization-with-nltk.html>
- [3] The beautiful soup related information is by crummy -<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
- [4] Edureka video tutorial for webscraping using beautifulsoup. <https://youtu.be/CWf0OtM3bbw>
- [5] Webscraping video tutorial by Coding entrepreneur - <https://youtu.be/3xQTJi2tqgk>