Search Engine Scrapper

Somnath Dudhat[1], Dnyaneshwar Nawale[2], Atharva Dhotre[3], Vinayak Rahate[4], Priyanka Halle[5], Prof. Priyanka halle[6]

[1,2,3,4,5] BE Students, Department of Computer Science & Engineering, SKN Sinhgad Institute of Technology And Science, Lonavala, Pune, Maharashtra, India
[6]Assistent Professor, Department of Computer Science & Engineering, SKN Sinhgad Institute of Technology And Science, Lonavala, Pune, Maharashtra, India

---***---

Abstract –

search engine scrapper is a set of processes which allows the user to collect relevant information presented on the World Wide Web (WWW) similar technology is used by search engines (Browsers like Chrome, Firefox Mozilla). This article covers the processes involves to extraction of data from different website content so user can get relevant and necessary information of their query.

Key Words: Web Scrapping, Web Crawling, Search Engine, Natural Language Processing

1. INTRODUCTION

Nowadays people are facing so much problem to search relevant information of their query, so to make them easy we are going to create search engine scrapper which applies web scrapping, natural language processing (NLP) and pointwise mutual information (PMI) and the use of SERP extraction API which help to extract and analyse the website content. Basically, our project model extracts the data from website content and then it summarized the data which is useful for the user and at the end summarized data is made grammatically correct with the help of machine learning module so that user can get the relevant and essential information of their query in meaningful way.

2.METHODOLOGY

1. <u>Extraction of Data:</u> -

It gathers the publicly available web data from different search engines through SERP Data Extractor APIs.

2. <u>Summarisation of Data</u>: -

SERP Extracted data get summarize through **NLTK** (Natural language Toolkit) processing libraries
Input document → sentences similarity → weight sentences → select sentences with higher rank.

3. <u>Conversion to relevant result</u>: -

In this service using ML model summarised data is getting converted to relevant result.

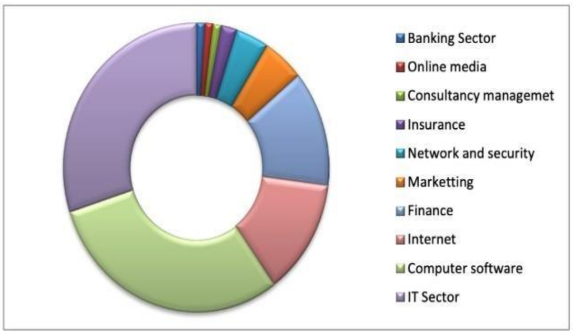


Figure 3: Fields using Web scraping

Search engine scrapping done in three stages:

1.extraction: This includes extraction of data from the different website content.

2.summarization: This deals with the summarization of extracted data, which is relevant to the user.

3.conversion to the relevant result: This includes the conversion of summarized data into meaningful manner so that user can understood the context.

3. LITERATURE SURVEY

Information is the most important asset in the world, but for retrieving it we need data. Data being the second important asset is not accessible to all the people around. Everyone can't get access to data which they require, for this purpose web scraping come up to the surface. Web Scraping has entirely shifted the way we used to see this world with less amount of data. Analysis and Retrieval have become so easy as of now.

Our life wouldn't have been the same without web scraping.

Many businesses have been able to skyrocket because of Web scraping as the collection of leads was made possible with it. The process of gathering unstructured data on the web is an interesting area within many contexts whether it be for business, scientific or personal usage. According to Mazlin etal, "selecting optimum significant features from high dimensional data may produce a challenge especially to an overfitted data that consequently resulting to data dimensionality issues." The advertisement business relies on the directed advertisement which is distributed throughout several pages, for the service to understand its current context, web data extraction, and web wrappers can be used in conjunction with content analysis tools for contextual analyses of the current page. In science, data sets are shared and used by several researchers and often publicly publicized. In some cases, the data sets are provided through a structured

API, but often data is only accessible through search forms and HTML documents which call for web wrapping methodologies to be used. Personal use has also grown as services have started to emerge which provides users with tools to mashup components from different web pages into own collection web pages.

Data scraping is a term used to describe the extraction of data from an electronic file using a computer program. Web scraping describes the use of a program to extract data from HTML files on the internet. Typically this data is in the form of patterned data, particularly lists or tables. Programs that interact with web pages and extract data use sets of commands known as application programming interfaces (APIs). These APIs can be 'taught' to extract patterned data from single web pages or from all similar pages across an entire web site. Alternatively, automated interactions with websites can be built into APIs, such that links within a page can be 'clicked' and data extracted from subsequent pages.

4. CONCLUSION

Like web scrapping, search engine scrapping is the process by which you can collect data from the websites and processing it and save it for further research or preserved it for over time.

Search engine scrapping is mostly use for gathering textual data which allows you to structure the data as you collect it, so instead of massive unstructured text, you can transform your scrap data into structured and well organised results that allows you to analyse and use it in machine learning model applications.

5. REFERENCES

1.Osmar Castrillo-Fernández, "Web Scraping: Applications and Tools", European Public Sector Information Platform Topic Report, no. 2015, December 2015.

2.M Kanehisa, S Goto, Y Sato et al., "KEGG for integration and interpretation of large-scale molecular data sets", Nucleic Acids Res, no. 40, pp. D109-14, 2012.

3.Glez-Pen‹a et al., Web scraping technologies in an API world, April 2013.

4.[online] Available: http://jsoup.org/.

5.[online] Available: http://adamsoft.sourceforge.net/.

6.[online] Available: https://nutch.apache.org/.

7.[online] Available: https://lucene.apache.org/solr/.

8.G. James, D. Witten, T. Hastie and R. Tibshirani, "An Introduction to Statistical Learning with Applications in R", Springer Texts in Statistics, 2013.

9.Giulio Barcaroli et Al, "Use of web scraping and text mining techniques in the Istat survey", European Conference on Quality in Official Statistics (Wien 2014), June 20.

10.Justin. Grimmer, Representational Style in Congress: What Legislators Say and Why It Matters, 2013.

11.William Marble, Web Scraping With R, stanford.edu, August 2016.

12.A. Carlos, Iglesias Mercedes Garijo Jose Ignacio Fernandez-Villamor and Jacobo Blasco-Garcia, "A Semantic Scraping Model for Web Resources", Applying Linked Data to Web Page Screen Scraping.

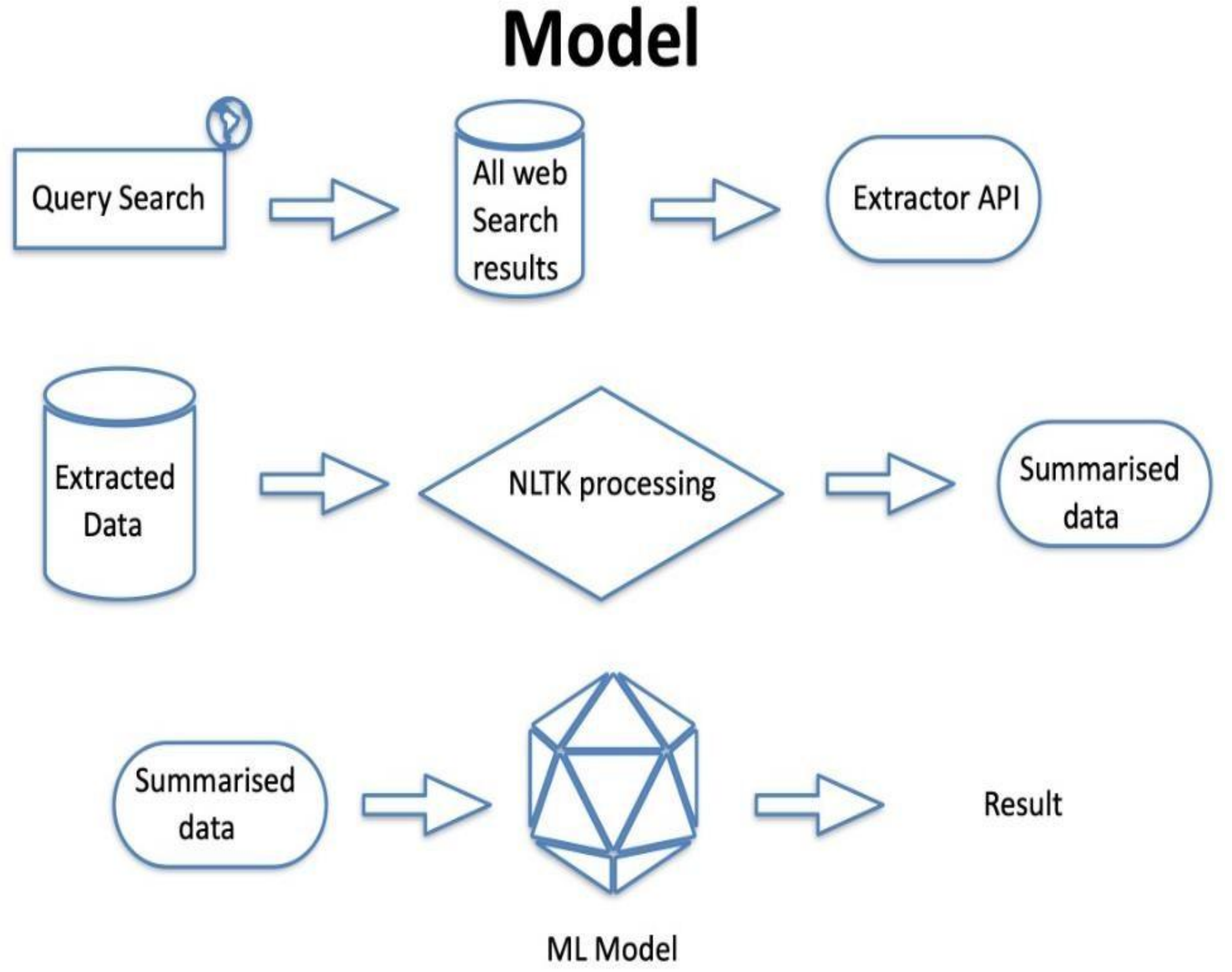
13.Muntasir Mashuq MichelZiyan Zhou, Web Content Extraction Through Machine Learning.

14.Diffbot: Extract content from standard page types: articles/blog posts front pages image and product pages, [online] Available: http://www.diffbot.com/.

15.Alex Gimson, A Data Journalism Webinar with BeaSchofield, [online] Available: http://blog.import.io/post/this-just-in-a-datajournalismwebinar-with-bea-schofield.

16.Daan Krijnen, Automated Web Scraping APIs, mediatechnology.leiden.e.

6.ARCHITECTURE DESIGN


Model
Query Search
All web Search results
Extractor API
Extracted Data
NLTK processing
Summarised data
Summarised data
ML Model
Result