The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 98 – 105

Smart Data Analysis
(Housing Data)

Dr. B. Sanjai Prasada Rao
Associate Professor
Department of Computer Science and Engineering
MLR Institute of Technology
Hyderabad, India-500043
Gayathri Rachabathuni
Department of CSE
MLR Institute of Technology
Hyderabad, India-500043
Rajdeep D. C
Department of CSE
MLR Institute of Technology
Hyderabad, India-500043
K. Poorn Anand
Department of CSE
MLR Institute of Technology
Hyderabad, India-500043

Abstract—In the modern era, the immense value which data holds goes into a waste as much of this information is not prominently visible to the common user, which the internet hosts. One of such areas is Renting/Buying houses in India. Today many people are still relying on brokers who are dishonest for buying/renting their homes. Even though many websites host good amount of information for the same purpose, many go unnoticed. This project aims to simplify the process of collecting and understanding the hosing data to use it effectively. The first module of the project deals with web scraping which allows the users to collect any number of data which internet hosts on different websites in any format and converting the same into a structured data set. The data extraction happens by extracting the HTML code which is embedded to the website and using a scraping tool to break the content on the website into useful data we require. This structured data set is available in the form of an excel sheet (.xlsx format or .csv format). Following it, the second module aims at querying the user and applying reverse indexing to the structured data to find relevant results based on the query raised by the user, to find the results.

Index Terms—Web Scrapping, HTML, Data collection, Structured Data, Unstructured Data, Inverted Index, Hash Map.

Introduction

Today, Internet has become one of the most used inventions in the human history, everything we deal in our daily loves has some part of involved in the internet, let it be for entertainment, research or practicing. Housing situation is something of great importance to the common people, people buying their first house in a far land or renting a house for temporary reasons, whatever it be, the options available to people ae huge, but are very tiresome. Either you approach a Broker who shows you all the properties with his huge margins or you explore the different options in the internet. Many websites like makaan.com, 99acres.com, housing.com, are designed in a manner to help you solve the needs, but these broker agents offer so many varieties, its impossible to find out the lucrative deal for you, as the results offered are very huge. To sum it up, it is a tiresome work to look through all the results even if you wish to filter out based on your needs from all these websites. Web scraping (Web harvesting or Data Harvesting) is the process of collecting data from a single website or multiple websites at once suing a crawler. It can be used to collect information about houses from, property listings, or other online sources. Here are some steps that can be taken to automatically collect house data using web scraping:

Inspect the website, Write a scraping script, Handle errors and exceptions, Ensure compliance with website policies. It's important to note that web scraping can be a complex process and may require technical expertise, so it's advisable to consult with a professional or legal advisor before proceeding. Additionally, scraping data without permission or in violation of website policies can have legal consequences, so it's important to exercise caution and respect website owners' rights.

The next important tool is Inverted indexing. Reverse Indexing (also referred as Inverted indexing) is a data structure which allows the user to store the indices or the location of the token we consider into a HashMap. A Hash Map is a mapping tool available in all programming languages which helps in mapping the root word (in our case the token) to a number or location. This aids us for easy retrieval of the data when we need and can be quickly access given, we also have the index of the token available with us. Reverse indexing has 2 parts namely, The record level inverted indexing and the word level inverted indexing. This project aims at using word level inverted index approach for retrieving data from the documents, as this technique is used when you a word in a document to mapped to its address (i.e.) where exactly it's situated or available. Reverse indexing also helps us in finding the results when multiple attributes are involved at once, narrowing down the options to a few which represents our actual solution. Both these tools when combined have immense potential to bridge the need for being able to analyse the data these websites hold and search for the requirements accordingly. It will help people to make accurate decisions by having access to all the critical data they need when buying a house, like the cost, living area size, the location etc. All the modules and critical functionalities will be built using python 3.6 using the scraping tools like (Scrapy or Beautiful soup) and the implementation of hash maps and arrays for analysing and reporting on the structured data will be implemented using python.

The need for collecting House Data

A. Real estate transactions:

Real estate agents and brokers need data on houses to help buyers and sellers make informed decisions about buying, selling, or renting a property. This data may include information about the size of the property, the number of rooms, its location, and the condition of the property.

B. Home improvement:

Homeowners and contractors need data on houses to plan home improvement projects. This data may include information about the age of the house, the condition of its systems and appliances, and any previous renovations or repairs.

C. Property assessment:

Local governments may collect data on houses to assess property taxes. This data may include information about the size of the property, its location, and its estimated value.

D. Research:

Researchers may collect data on houses for various purposes, such as studying housing trends, understanding housing markets, or examining the impact of housing policies on different populations.

Literature Survey

Many papers have been reviewed in the field of Housing data and House rent management in general. The project [1] was helpful to find out the purpose of developing our project and also to channel on what direction out project should flow.

According to the paper [1], there is greater need for connecting the people directly with the owner of the house, eliminating the involvement of a broker, so that their shady practices to earn more money by duping us can be eliminated. The writers of this paper aim at implementing a Deep learning Technique which scans the history of the users on the application to recommend them houses based on what they saw. The recommendation system worked on the basis of a CNN for classifying the house images. The paper [1] concluded with the idea of creating a personalized, easy to use application that will reduce the effort and time required for people in the renting market to find housing.

Project [2] and Project [3] focused more on the cost perspective for the rental options. The authors were focused on implementing a tool which would predict the price of the homes based on the different months in a city. The model was trained with previous years data where clustering tools have been used to group those months together where the property prices would experience a sharp drop on sharp increase.

Independent websites like makaan.com, housing.com have existed since a long time in India, hosting hundreds of properties online for the viewers to choose from. The magnitude of the data available with these sites is so bug, it becomes tiresome to keep scrolling through results looking for the perfect conditions while also navigating through different sites for the same purpose. It was also discovered, the prices of certain properties being lower on certain sites like 99acres.com than the traditional players.

Reviewing papers [5] and [6] was beneficial to understand how to start the idea of this project. The concept of web scraping, the tools involved, the legal issues linked with the concept and the practical application of the same. Some internet articles and YouTube videos were watched to grasp the concept and understanding its importance of the same to deliver this project. The authors of paper [6] were also clear in stating the convenience and simplicity linked with using python and python tools for the process of scraping, thus the tools and programming procedures were decided to be made entirely using PYTHON programming language.

Information retrieving was understood upon querying the same on Google, which gave us an idea about reverse/inverted indexing. The authors of paper [7] give us a complete detailed idea/guide on the use of reverse Indexing

Limitations Of Existing Systems

1) Neither of the projects reviewed so far had collected the data from multiple sources, all relied on a single broker/agent, thereby decreasing the effectiveness of the project
2) New technologies were used, but all of the systems are volatile, as the property market can change instantly due to various reasons and the machines designed might have to be trained again for accuracy
3) Papers [1] and [4] aims at establishing a direct connection between the owner and the customer wishing to buy the house, while the intention may seem good, they seriously lack in the number of options provided to the customers. (Limited study data available with them)
4) Some websites display prices which are less when compared to others, if proper research has not been done, a lucrative deal can easily slip through the hands of the customer.
5) All the solutions above still are dependent on users to scan through different results with different aspects, which is tiresome on its own.

Proposed System

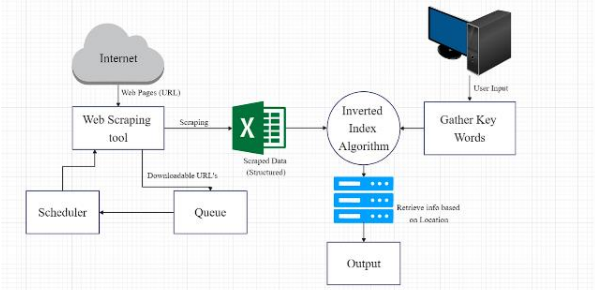
The proposed system involved around the idea of data collection and querying. Data collection is possible with the help of web scraping. Multiples websites can be scanned at once to get relevant information. Then information retrieval will be used in order to get the answers to user queries

A. Objectives of Proposed systems

- To collect data from multiple sites.
- Convert the unstructured data in net, to structured data set.
- Prepare a system which allows you to query the dataset.
- Build an efficient system which allows for information retrieval

B. Proposed System Architecture

The below depicted System architecture depicts the system's components, their relationship and functionality. The Data is absorbed from the internet, scraped by the web scraping tools, linked with the IR model to generate the final results.



The components involved in the architecture are:

Internet: The Internet is a of billions of computers and other electronic devices. With the Internet in the hand, it makes accessing any part of the data easy and convenient

Web Scraping tool: This is a tool consisting of set methods in a library which allows you to collect information from the various websites. We intend to use Scrapy or Beautiful Soup, 2 renowned libraries of python which allows for collection/extraction of data with ease from websites. Scrapy is a framework for more diverse and broader tool covering more components of web pages simultaneously.

It can also deal with asynchronous data increasing the efficiency of data retrieval. Beautiful Soup on the other hand is a go to version of web scraping tool which is a python library with set feature defined in place already for scarping the data. It provides quick scarping for general loads and is more concise than Scrapy.

The components of web scraping include:

- Web Crawler: It is used to travel from one web page to another based on the URL's provided.
- Web Scraper: It is used to read the HTML code and analyse which tags needs to be targeted to extract data.

Queue: The queue generally holds the downloadable URLs of the website (i.e.) only the HTML codes link or JSON etc. The data is fed into Queue by Crawler.

Scheduler: It is used to dispatch the link stored on the Queue to the Scraper which will extract the data from the HTML code.

Scraped Data: It is the structured data which is generated as output after the website has been scraped based on the parameters specified.

User Input: Asking users of the requirements regarding housing data, like location, cost, size of the home etc.

Gather the key words: It is used if the input has been given in the form of a text. Here the aim is to remove the stop words and give only key words to the system

Inverted Index: It is a data structures which aims are storing the indexes, more specifically the locations of keywords written/available in the structured data set which has been generated. It uses Hash Map for mapping the data in excel sheet with their rows and columns. Thus, Hash Maps are generated for each document.

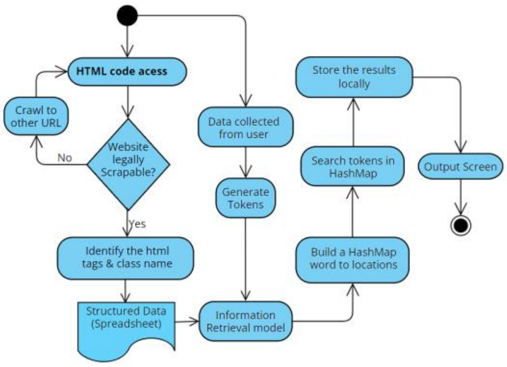
Information retrieval: This is done by narrowing down our available options based on the attributed specified by user. For instance, if the user has given 2 attributes, The hash map will be used to identify which cells in structured data set are matching the requirements.

Based on the results, the remaining cells will be discarded and withing the scope of cells available, the next attribute will be searched. At the end, the whatever cells remain, are the output indicating certain house have been found which matches the need of user.

Output: The output will basically contain a list from which website, what data has been found useful and the link to navigate to the website for further rounds of operations.

C. UML Diagram

The UML diagram shows the interaction of one component with another in a linear fashion and how the systems software receives the user input

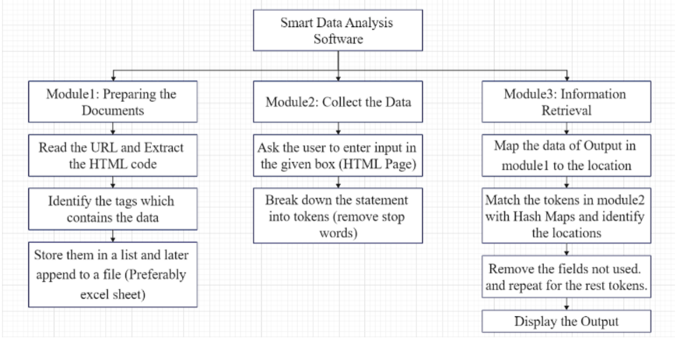

HTML code acess
Crawl to other URL
No
Website legally Scrapable?
Yes
Identify the html tags & class name
Structured Data (Spreadsheet)
Data collected from user
Generate Tokens
Information Retrieval model
Build a HashMap word to locations
Search tokens in HashMap
Store the results locally
Output Screen

The origin of the diagram flows into two states simultaneously as shown in the diagram. From there the paths then converge at Information retrieval model which takes the output of the earlier two phases and takes an inverted indexing approach to solve the query raised by the user. The output is a single node which is displayed on the output screen through the HTML page designed for taking input from the user. The activity diagrams provide a clear and concise representation of the process flow in the proposed system, and they can help developers and stakeholders better understand the system's behaviour and interactions.

D. Module Diagram

The module diagram shows a high-level view on the flow of execution of project. It has been categorised into 3 phases:

- Preparing the Documents
- Collecting the User input
- Information retrieval


Smart Data Analysis Software
Module1: Preparing the Documents
Read the URL and Extract the HTML code
Identify the tags which contains the data
Store them in a list and later append to a file (Preferably excel sheet)
Module2: Collect the Data
Ask the user to enter input in the given box (HTML Page)
Break down the statement into tokens (remove stop words)
Module3: Information Retrieval
Map the data of Output in module1 to the location
Match the tokens in module2 with Hash Maps and identify the locations
Remove the fields not used. and repeat for the rest tokens.
Display the Output

i. Preparing the Documents

The data which is needed to be extracted is loaded in the web scraper by the web crawler. This module has been divided into 3 critical functionalities as shown in the figure 2. Without the completion of previous functions as mention the figure 2, it is not possible to proceed further. After Module1 has been completed, the output is sent to Module3

ii. **Collecting Input for user**

This is a simple step where the data is collected from the user in the form of text and broken down into token removing all the stop words that have been found. This data is then passed onto Module3.

iii. **Information Retrieval Model:**

The last and the crucial step in the process of making final decisions, this module has been identified with 4 functions having their own importance. All the functions are co related to each other. In case of multiple documents, the cycle gets repeated again form start until the particular document has not been analysed, the answer later are stored as pair, containing the doc id and the location of answer (row number) for the query which user had issued, so that a concise answer can be shown to the user by:

- Using the Doc-id to refer to the website that was scraped.
- Using the Location Number to show the exact result stored on the spreadsheet

Simulation Settings

A. *Hardware Requirements*

- Processor (CPU): Intel core i5 2.6GHz (or higher)
- Display: 1920 x 1080
- RAM: 4GB (Minimum)
- Disk Storage: 2GB (Minimum)

B. *Software Requirements*

- Language: Python (3.0 and higher)
- Operating System: Windows 7 (or higher)
- Integrated Development Environment (For ease in coding)

C. *Data Sources*

The data for the project has been collected from mane online brokers. Some of the major brokers include, makan.com, housing.com, 99acre.com and many others. The count of the data being collected is more than 100 from each website, which will be scraped based on the requirement of the user.

Performance Evaluation

The proposed system is evaluated based on the following parameters.

- Time: How quickly can the IR model scan through the document and give the detailed information to us?
- Accuracy: Based on the structured data, how accurately is the model able to predict correct answer?
- Simple UI: The User being able to simply ask a question, and receive the output in same window in couple of seconds.
- Data Gathering: How much amount of data is gathered form a particular website? Is the data source reliable? How many other data sources can be harvested for the same reason?

Conclusion

Internet is a blessing in disguise for the common people. People in India face a lot of issues as the new multiplexes, apartments are comping up in the city, but there is no way for the people to know about them unless they come across them. This limits our capacity of searching and confine us to the information what we have knowledge of. Online House Rent/ownership model helps in tackling the same issue by providing u a curated list at the comfort of your home, by scanning through large sets of Homes. Let it be finding a cheaper flat for rent or buying a dream home for you and family, this project is the perfect answer to the housing solution!

Acknowledgment

It would have been highly difficult for such a project to be carried out without the help of our guide Dr. B. Sanjai Prasada Rao, his guidance and help made the project to be driven in the correct path with minimal scope of error. I would also like to thank all the authors of the IEEE papers we have referred to for doing this project, their results helped us to select the correct choices and actions.

References

[1] "ZenDen - A Personalized House Searching Application" Authors : Kristina Milkovich; Saurabh Shirur; Pratap Kishore Desai; Likhith Manjunath; Wencen Wu 2020 IEEE Sixth International Conference on Big Data Computing Service and Applications

[2] "Rental application to rental service development of advanced ASP framework" Authors: T. Kawamura; T. Hasegawa; A. Ohsuga; S. Honiden Proceedings Fourth International Enterprise Distributed Objects Computing Conference. EDOC2000

[3] "Real-Estate Price Prediction System using Machine Learning" Authors: Veerraju Gampala; Nalajala Yaznitha Sai; Tadikonda Naga Sai Bhavya 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)

[4] "Development of Online Based Smart House Renting Web Application" Authors: Dipta Voumick, Prince Deb, Sourav Sutradhar, Monirujjaman Khan Journal of Software Engineering and Applications, 14, 312-328. doi: 10.4236/jsea.2021.147019.

[5] "A Review on Web Scrapping and its Applications" Authors: Vidhi Singrodia; Anirban Mitra; Subrata Paul 2019 International Conference on Computer Communication and Informatics (ICCCI)

[6] "Data Analysis by Web Scraping using Python" Authors: David Mathew homas; Sandeep Mathur 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)

[7] "Inverted index and interval lists for keyword search" Authors: J Giridharan; S. V. Vairavan 2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)

[8] Web Scraping - Legal or Illegal? - GeeksforGeeks

[9] What Is Web Scraping? [A Complete Step-by-Step Guide] (careerfoundry.com).

[10]. Inverted Index - GeeksforGeeks