Vivekanand Education Society's
Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai, Approved by AICTE & Recognised by Govt. of Maharashtra)
NAAC accredited with 'A' grade

PROJECT REPORT

ON

WEB SCRAPING
For Hotels

SUBMITTED IN FULFILLMENT OF THE REQUIREMENT FOR SEMESTER IV OF

-S.E. (Information Technology)

SUBMITTED BY

Mr. BHUSHAN MALPANI

Mr.SAHIL MOTIRAMANI

Mr.ATHARVA SHINDE

Mr.SHIVPRATIK HANDE

UNDER THE GUIDANCE OF

PROF. BINCY IVIN

DEPARTMENT OF INFORMATION TECHNOLOGY
V.E.S. INSTITUTE OF TECHNOLOGY
2023-24

Vivekanand Education Society's
Institute of Technology

(Autonomous Institute Affiliated to University of Mumbai, Approved by AICTE & Recognised by Govt. of Maharashtra)
NAAC accredited with 'A' grade

Certificate

This is to certify that project entitled

Web Scraping For Hotels

Group Members Names

Mr.BHUSHAN MALPANI(Roll No.34)
Mr.SAHIL MOTIRAMANI(Roll No.37)
Mr.ATHARVA SHINDE(Roll No.57)
Mr.SHIVPRATIK HANDE(Roll No.16)

In fulfillment of degree of BE. (Sem. IV) in Information Technology for Projectis approved.

Prof. BINCY IVIN
Project Mentor

External Examiner

Dr.(Mrs.)Shalu Chopra
H.O.D

Dr.(Mrs.)J.M.Nair
Principal

Date: / /2024
Place: VESIT, Chembur

College Seal

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

- - - - - - - - - - -

(Signature)

BHUSHAN MALPANI (34)
SAHIL MOTIRAMANI (37)
ATHARVA SHINDE (57)
SHIVPRATIK HANDE (16)

Abstract

Web scraping, a technique for automated data extraction from websites, has garnered significant attention in various industries for its role in data collection, analysis, and decision-making. This report presents a comprehensive study on web scraping, focusing on its applications, challenges, and best practices. The literature survey delves into existing research on web scraping methodologies, legal and ethical considerations, and practical implementations, encompassing both theoretical frameworks and methodological approaches. Key findings from the survey highlight the effectiveness of hybrid scraping techniques, the importance of legal compliance, and ethical practices in web scraping activities. The report also includes a mini project on web scraping, demonstrating the practical application of scraping techniques for data collection and analysis. Overall, this report provides valuable insights and recommendations for leveraging web scraping effectively in data-driven processes, with a focus on publication-worthy contributions to the field.

Contents

List of Figures

ACKNOWLEDGEMENT

The project report on ”web scraping of hotels ” is the outcome of the guidance, moral support and devotion bestowed on our group throughout our work. For this we acknowledge and express our profound sense of gratitude to everybody who has been the sourceof inspiration throughout project preparation. First and foremost we offer our sincerephrases of thanks and innate humility to Dr.(Mrs.)Shalu Chopra and HOD, Dr.(Mr.)Monoj Sabnis and Deputy HOD , Prof. Bincy Ivin and Project Mentor for providing the valuable inputs and the consistent guidance and support provided by them.We can say in words that we must at outset tender our intimacy for receipt of affectionate care to Vivekanand Education Society’s Institute of Technology for providing such a stimulating atmosphere and conducive work environment.

Chapter 1

Introduction

1.1. Introduction

- Provide an overview and the significance of hotel data in market analysis, pricing strategies, and customer experience.
- Explain the purpose of your web scraping project, such as gathering competitive intelligence, monitoring price fluctuations, or analyzing customer reviews.

1.2. Objectives

- **Data Acquisition**: Gather comprehensive information about hotels from various online travel platforms to create a rich dataset.
- **Comparison and Analysis**: Enable users to compare hotels based on factors such as prices, ratings, amenities, and reviews to make informed booking decisions.
- **Automation**: Automate the process of collecting hotel data to save time and resources for both users and businesses in the travel sector.
- **Customization**: Allow users to customize their search criteria and preferences to find hotels that best suit their needs and preferences
- **Decision Support**: Assist users in selecting the most suitable accommodations by presenting them with relevant and up-to-date information about available options.
- **Monitoring Competitors**: Enable businesses in the hospitality industry to monitor competitors' offerings and pricing strategies for strategic decision-making.

1.3. Motivation

- Industry Relevance: Discuss why hotel web scraping is relevant in today's hospitality industry, including its role in competitive analysis, pricing optimization, customer sentiment analysis, and market trend identification.
- Benefits: Highlight the benefits of using web scraping techniques for hotel data collection, such as efficiency, scalability, accuracy, and real-time monitoring capabilities

1.4. Scope of the Work

- Identify the primary data sources for hotel web scraping, such as online booking platforms ,hotel chain websites, or review aggregators.
- Specify the types of data to be extracted, such as room rates, property descriptions, images, customer reviews, ratings, and amenities.
- Define the geographical scope of the project, including the regions, cities, or specific properties targeted for data collection.

1.5. Feasibility Study

- Evaluate the technical feasibility of web scraping for hotel data, considering factors such as website structure, data volume, frequency of updates, anti-scraping measures, and available scraping tools/libraries (e.g., BeautifulSoup).
- Address legal aspects, including compliance with website terms of service, data privacy regulations, intellectual property rights, and ethical considerations related to data scraping and usage.

1.6. Organization of the report

-Introduction

The introduction section provides an overview of the project, outlining its purpose,objectives, motivation, scope, feasibility study, and the organization of the report.

-**Literature Survey**

The literature survey section reviews existing research and scholarly work related to webscraping and data extraction, including an introduction, problem definition, and a comprehensive review of relevant literature.

-Design Implementation

The design implementation section details the process of translating conceptual ideas into practical solutions, covering requirement gathering, proposed design, proposed algorithm,architectural diagrams, hardware requirements, and software requirements.

-**Results and Discussion**

The results and discussion section presents the findings and analysis of the web scrapingproject, including an introduction, cost estimation, feasibility study, results

of implementation, result analysis, and observations/remarks.

-Conclusion

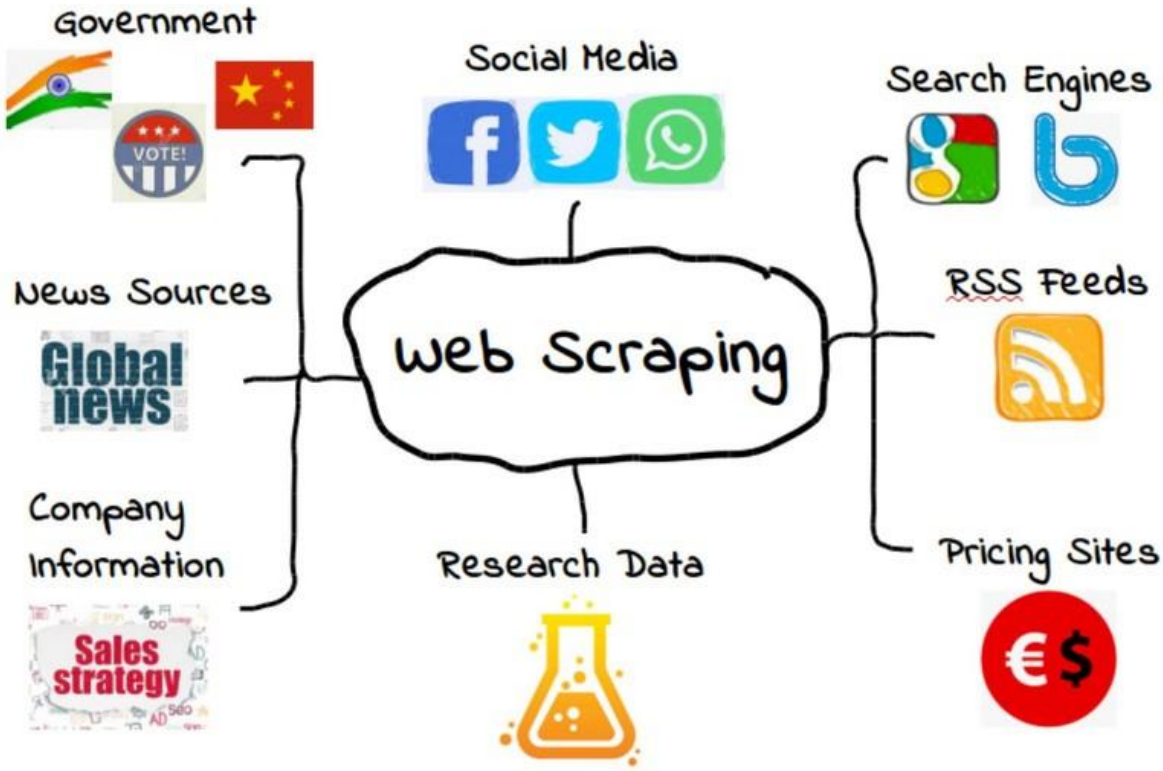
The conclusion section summarizes the key findings and outcomes of the project, discussesfuture scope, and acknowledges any published papers resulting from the project.

Chapter 2

Literature Survey

2.1 Introduction

In this section, we provide an introductory overview of the literature surveyed related to web scraping. We discuss the importance of web scraping in data collection, analysis, and decision-making processes.



2.2 Problem Definition

The subsection on problem definition outlines the key challenges and issues addressed in the literature regarding web scraping. This includes legal, ethical, technical, and practical considerations that researchers and practitioners encounter when conducting web scraping activities.

2.3 Review of Literature Survey

In this subsection, we summarize and review relevant papers and studies related to web scraping. Each paper is analyzed based on its objective, problem statement, proposed system or methodology, and conclusions. The review provides insights into the current state of research and best practices in web scraping.

Paper 1: [Web Scraping Techniques and Applications: A Literature Review]

Objective:
The objective of this literature review paper is to explore and analyze the various web scraping techniques and their applications across different domains. The paper aims to provide a comprehensive overview of the state-of-the-art in web scraping methodologies, highlighting their effectiveness, challenges, and potential applications in data-driven decision-making processes.
Findings:

- This paper provides an updated review of advanced web scraping techniques.
- Equipping scholars and managers with knowledge on effectively mining online data.

Drawbacks:

- The study does not delve into specific challenges related to hotel-specific web scraping1.

-[paper1]

Paper 2: [Comprehensive Review on Techniques for Scraping Hotel Information from Online Sources]

Objective:
The objective of this comprehensive review paper is to analyze and evaluate the various techniques and methodologies used for scraping hotel information from online sources. The paper aims to provide an in-depth examination of the challenges, best practices, and emerging trends in web scraping specifically focused on hotel-related data, including pricing, availability, amenities, reviews, and other relevant information.
Findings:

- The survey offers structured and unstructured data extraction methods.
- Use of natural language processing for sentiment analysis of hotel reviews.

Drawbacks:

- Lack of discussion on the scalability and efficiency of different scraping approaches.

-[paper2]

Paper 3: [Web Scraping Methods for Hotel Price Monitoring]

Objective:
The objective of this paper is to evaluate and compare different web scraping

methods used for hotel price monitoring. The paper aims to provide insights into the effectiveness, accuracy, and efficiency of various scraping techniques in retrieving and analyzing hotel pricing data from online sources. Additionally, the paper seeks to identify best practices and challenges associated with web scraping for price monitoring purposes.

Findings:

- This paper compares various web scraping methods specifically for hotel price monitoring applications.

Drawbacks:

- Limited focus on the adaptability of scraping techniques to different hotel booking platforms.

-[paper3]

Paper 4:[State-of-the-Art Techniques in Web Scraping for Hotel Booking Data]

Objective:

The objective of this paper is to explore and analyze state-of-the-art techniques in web scraping specifically tailored for extracting hotel booking data from online platforms. The paper aims to evaluate the effectiveness, accuracy, and scalability of advanced web scraping methods and technologies used for collecting real-time booking information, including prices, availability, room types, and customer reviews.

Findings:

- Advancements in automated data collection, such as headless browsing and proxy rotation.
- addresses challenges related to data quality assurance and privacy concerns.

Drawbacks:

- Limited discussion on the impact of website layout changes on scraping accuracy.

-[paper4]

Paper 5:[An Overview of Web Scraping Methods for Hotel Information Retrieval]

Objective:

The objective of this paper is to provide an overview and analysis of web scraping methods specifically focused on retrieving hotel information from online sources. The paper aims to explore the various techniques, tools, and best practices used for extracting hotel-related data, including pricing, availability, amenities, reviews, and other relevant information. Additionally, the paper seeks to evaluate the effectiveness, accuracy, and ethical considerations of web scraping in the hospitality industry.

Findings:

- Covers techniques such as DOM parsing, web crawling, and API integration.

Drawbacks:

- Limited analysis of the legal and ethical considerations associated with web scraping practices.

-[paper3]

This structured approach to the Literature Survey section of your report provides a clear overview of the existing research landscape, challenges, proposed solutions, and insights gained from the literature on web scraping. Adjust the details and examples based on the actual papers you've surveyed and the specific focus of your mini project.

Chapter 3

Design Implementation

3.1. Introduction

The design implementation phase represents a pivotal stage in the project lifecycle, where theoretical concepts are translated into tangible solutions. This section serves as a blueprint for the practical execution of the proposed design for the hotel web scraping project. It encompasses the methodologies, tools, and techniques utilized to transform abstract ideas into functional systems that meetstakeholder requirements and objectives.

3.2. Requirement Gathering

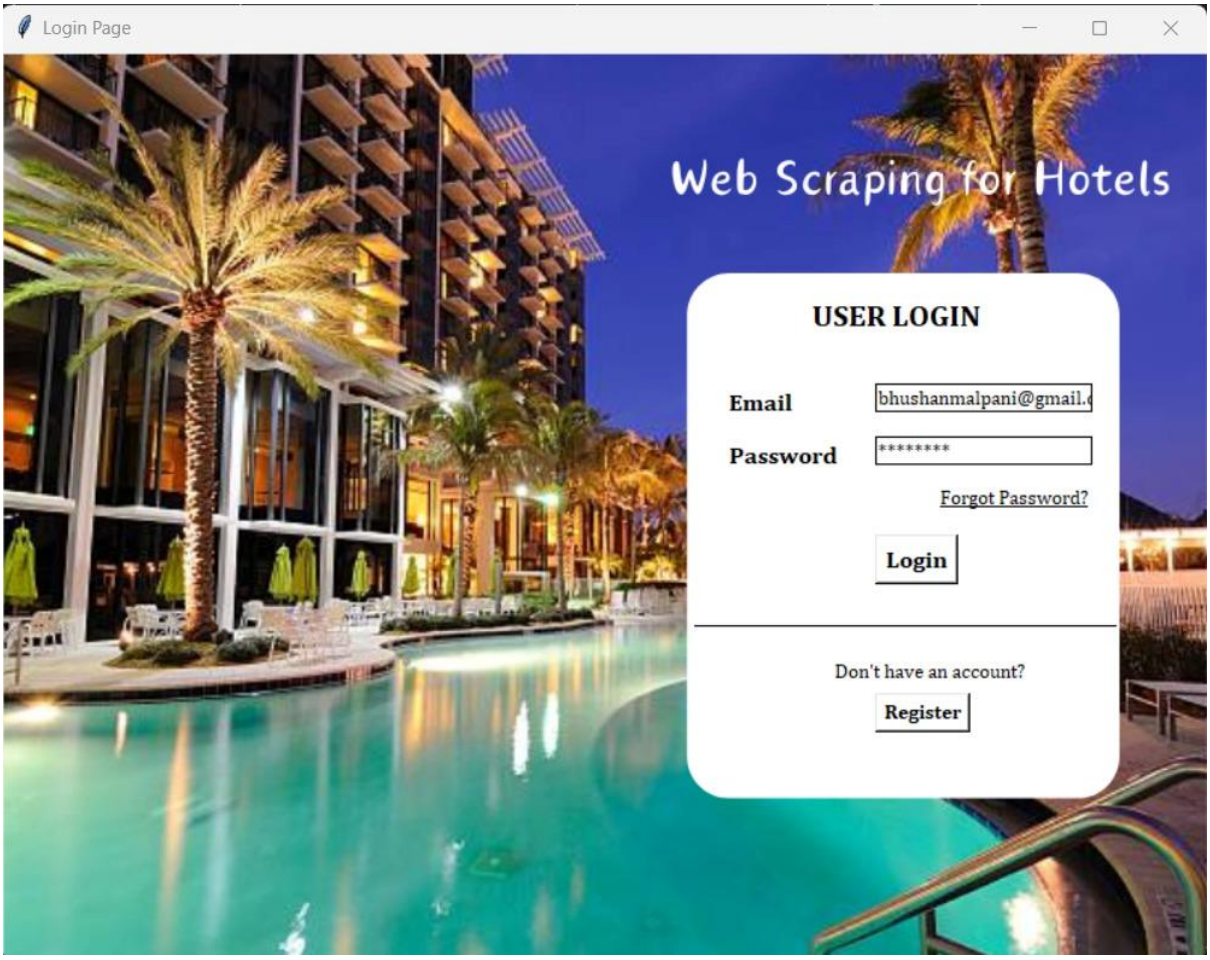
A comprehensive requirement gathering process was conducted to elicit and document the needs, expectations, and constraints of stakeholders. This involved engaging with key stakeholders, including project sponsors, end-users, and domainexperts, to understand their perspectives and gather insights into the desired functionality and scope of the web scraping solution. Various techniques, such as interviews, surveys, and analysis of existing systems, were employed to ensure a thorough understanding of both functional and non-functional requirements.

4

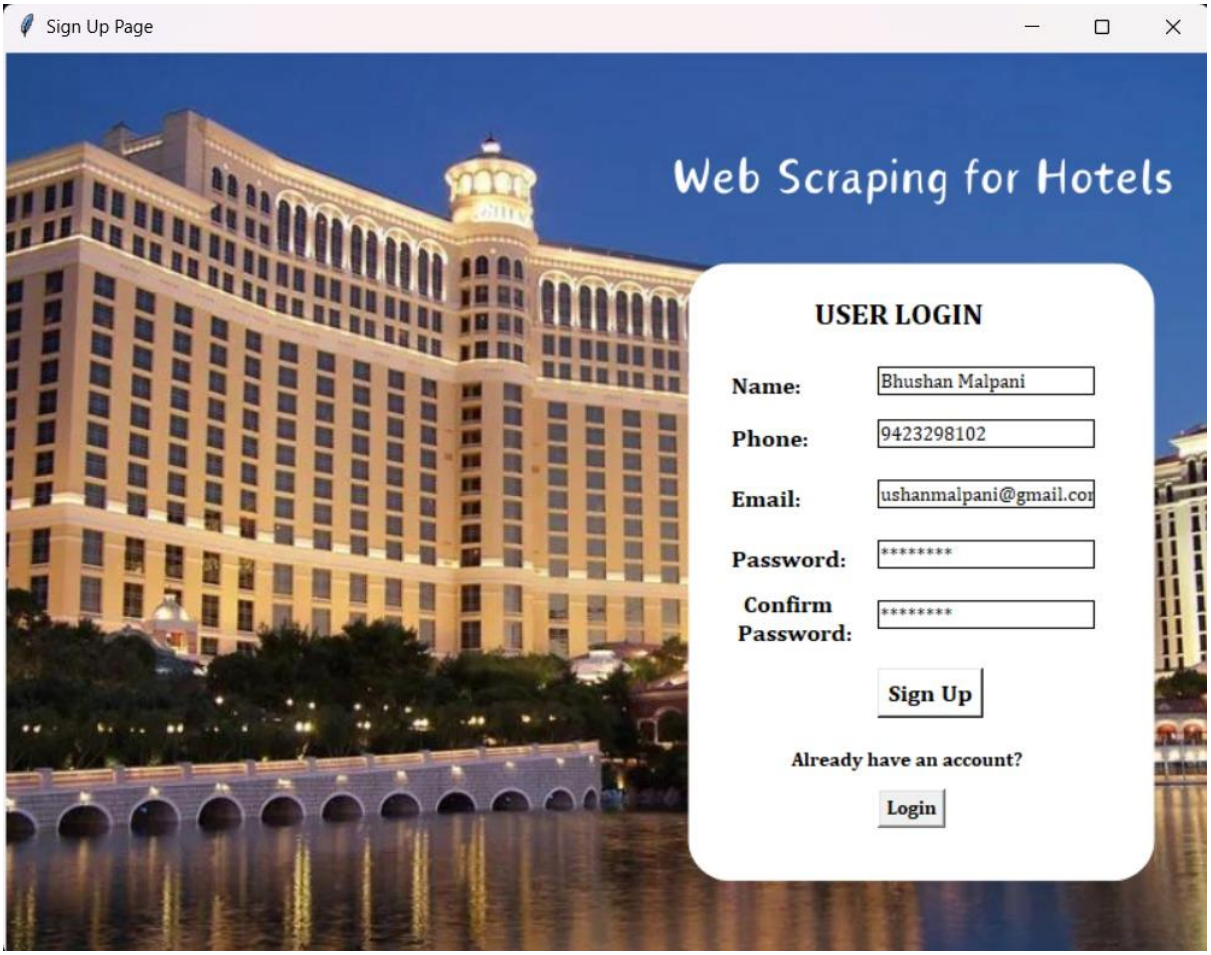
3.3. Proposed Design

Building upon the insights gleaned from the requirement gathering phase, a proposed design was formulated to guide the development of the hotel web scraping solution. This design encapsulates the architectural framework, systemcomponents, and data flow mechanisms necessary to achieve the project objectives. It serves as a roadmap for implementation, providing a clear directionfor developers and stakeholders alike. The proposed design emphasizes modularity, scalability, and maintainability, facilitating iterative development andfuture enhancements.

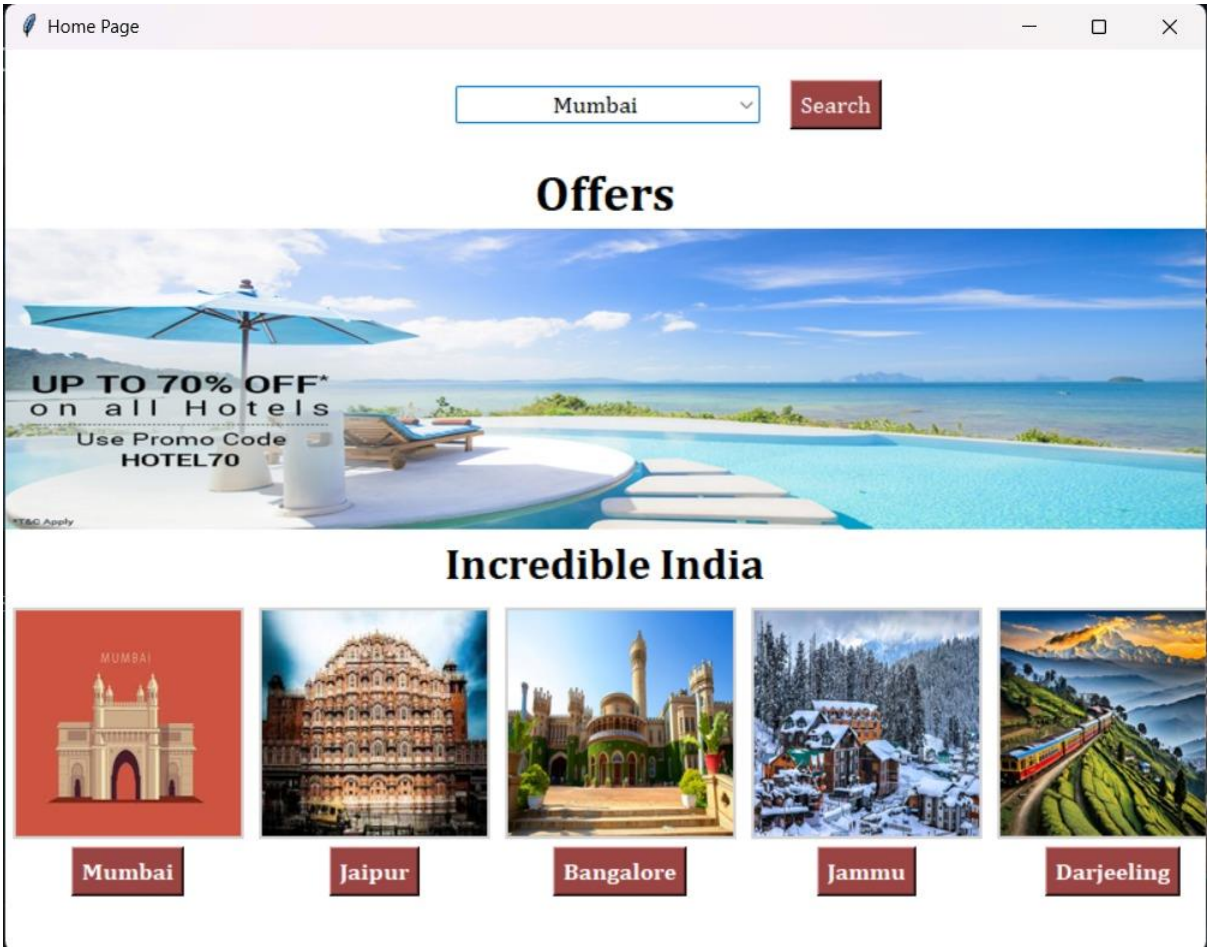
Login page


Login Page
Web Scraping for Hotels
USER LOGIN
Email
bhushanmalpani@gmail.c
Password
****
Forgot Password?
Login
Don't have an account?
Register

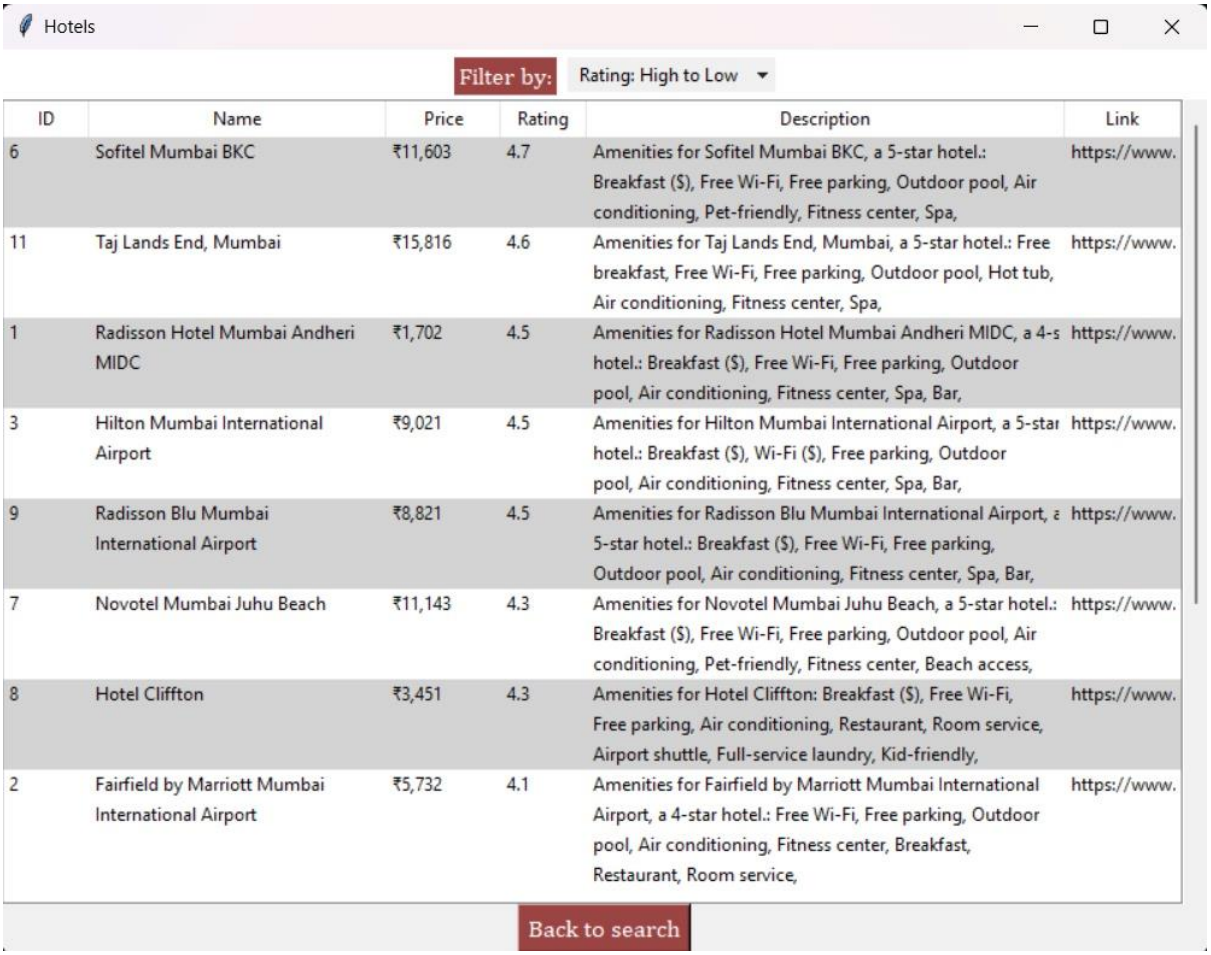
Sign up page


Sign Up Page
Web Scraping for Hotels
USER LOGIN
Name:
Bhushan Malpani
Phone:
9423298102
Email:
ushanmalpani@gmail.cor
Password:
****
Confirm Password:
****
Sign Up
Already have an account?
Login

Home page


Home Page
Mumbai
Search
Offers
UP TO 70% OFF*
on all Hotels
Use Promo Code
HOTEL70
*T&C Apply
Incredible India
MUMBAI
Mumbai
Jaipur
Bangalore
Jammu
Darjeeling

Scraped Data



ID	Name	Price	Rating	Description	Link
6	Sofitel Mumbai BKC	₹11,603	4.7	Amenities for Sofitel Mumbai BKC, a 5-star hotel.: Breakfast ($), Free Wi-Fi, Free parking, Outdoor pool, Air conditioning, Pet-friendly, Fitness center, Spa,	https://www.
11	Taj Lands End, Mumbai	₹15,816	4.6	Amenities for Taj Lands End, Mumbai, a 5-star hotel.: Free breakfast, Free Wi-Fi, Free parking, Outdoor pool, Hot tub, Air conditioning, Fitness center, Spa,	https://www.
1	Radisson Hotel Mumbai Andheri MIDC	₹1,702	4.5	Amenities for Radisson Hotel Mumbai Andheri MIDC, a 4-s hotel.: Breakfast ($), Free Wi-Fi, Free parking, Outdoor pool, Air conditioning, Fitness center, Spa, Bar,	https://www.
3	Hilton Mumbai International Airport	₹9,021	4.5	Amenities for Hilton Mumbai International Airport, a 5-star hotel.: Breakfast ($), Wi-Fi ($), Free parking, Outdoor pool, Air conditioning, Fitness center, Spa, Bar,	https://www.
9	Radisson Blu Mumbai International Airport	₹8,821	4.5	Amenities for Radisson Blu Mumbai International Airport, a 5-star hotel.: Breakfast ($), Free Wi-Fi, Free parking, Outdoor pool, Air conditioning, Fitness center, Spa, Bar,	https://www.
7	Novotel Mumbai Juhu Beach	₹11,143	4.3	Amenities for Novotel Mumbai Juhu Beach, a 5-star hotel.: Breakfast ($), Free Wi-Fi, Free parking, Outdoor pool, Air conditioning, Pet-friendly, Fitness center, Beach access,	https://www.
8	Hotel Cliffton	₹3,451	4.3	Amenities for Hotel Cliffton: Breakfast ($), Free Wi-Fi, Free parking, Air conditioning, Restaurant, Room service, Airport shuttle, Full-service laundry, Kid-friendly,	https://www.
2	Fairfield by Marriott Mumbai International Airport	₹5,732	4.1	Amenities for Fairfield by Marriott Mumbai International Airport, a 4-star hotel.: Free Wi-Fi, Free parking, Outdoor pool, Air conditioning, Fitness center, Breakfast, Restaurant, Room service,	https://www.

3.4. Proposed Algorithm

An algorithmic approach was devised to orchestrate the web scraping process, outlining a systematic procedure for retrieving, parsing, and storing data from hotel websites. This algorithm incorporates best practices in data extraction and processing, ensuring efficiency, accuracy, and robustness. By breaking down thescraping process into sequential steps, the algorithm enables developers to implement each stage methodically, optimizing performance and minimizing errors. Additionally, the algorithm accommodates variations in website structuresand data formats, enhancing the solution's adaptability and versatility

3.5.Architectural Diagrams

3.5.1 UML Diagrams

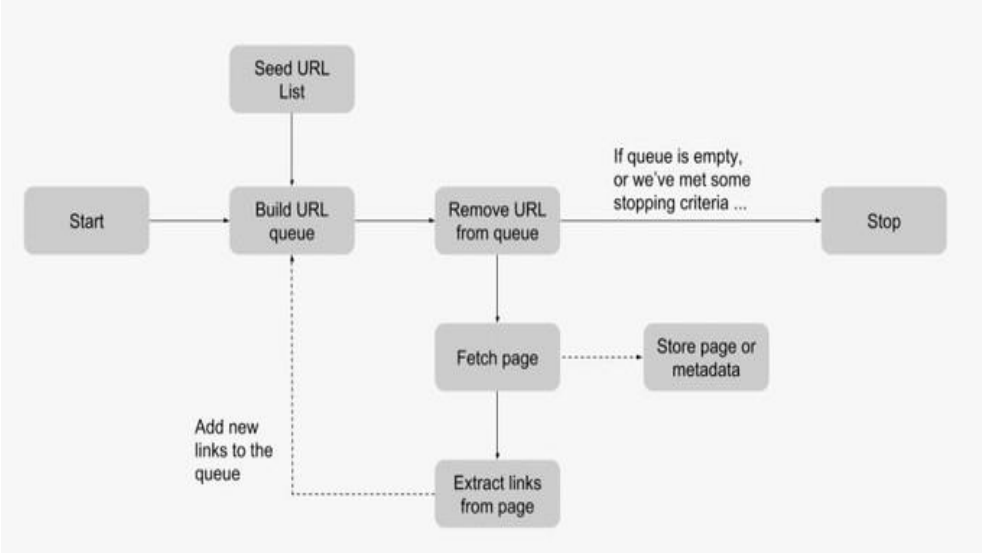


Figure 3.1: UML Diagrams

3.5.2 Block Diagrams

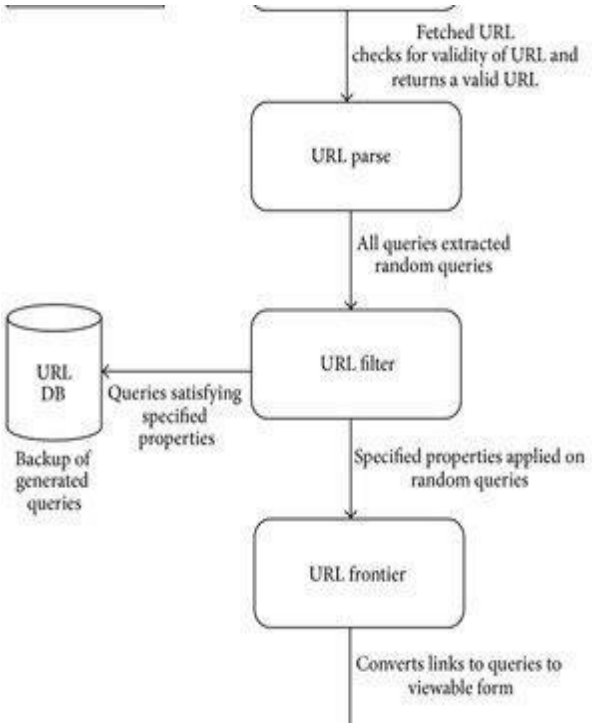


Figure 3.2: Block Diagram

3.5.3 Data Flow Diagrams

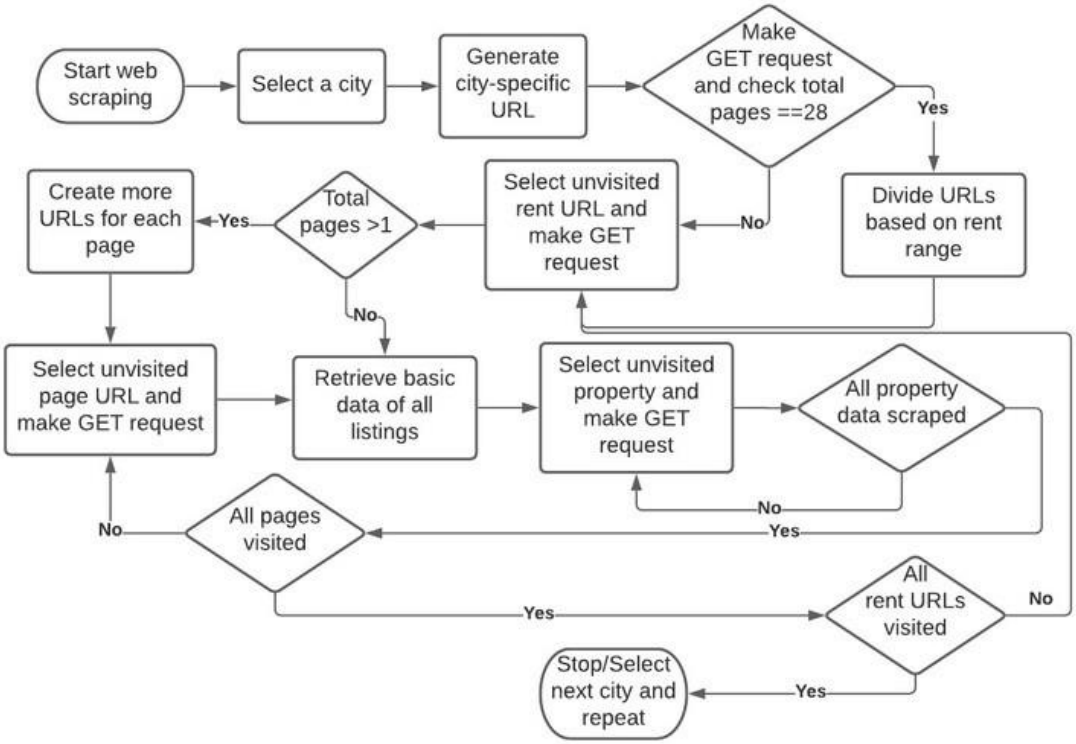
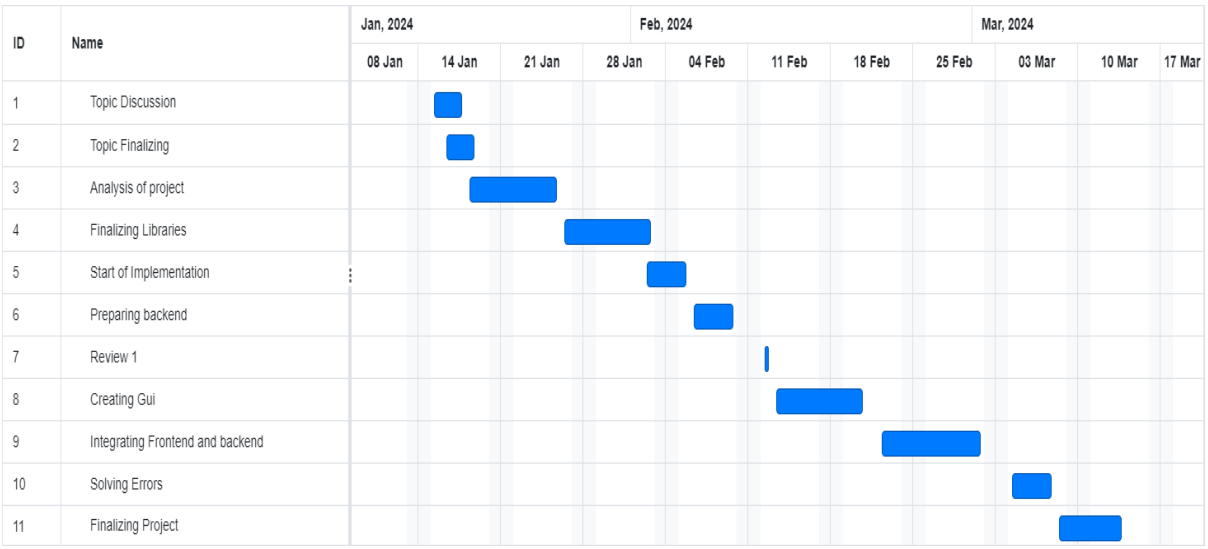


Figure 3.3: Data Flow Diagram

3.5.4 Timeline Chart



3.6 Software Requirements

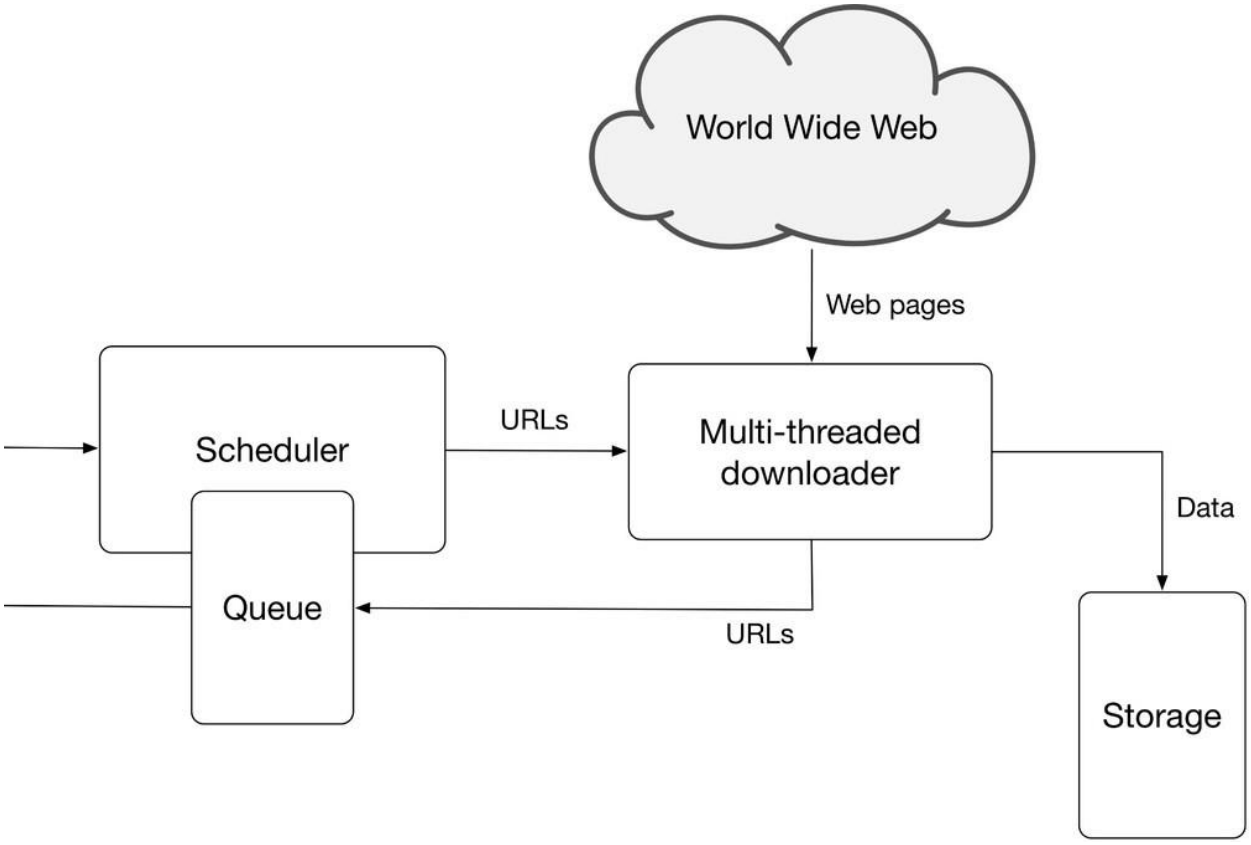
1. **Programming Languages:**
   - Backend : Python
   - Database: Sqlite3
   - Frontend : Python
2. **Libraries:**
   - **BeautifulSoup** Python library for parsing HTML/XML documents
   - **re** Python library for regular expressions.
   - **Pandas** Data manipulation and analysis library in Python.
   - **Requests** Python library for making HTTP requests easily.
   - **Tkinter** Python library for creating graphical user interface

Chapter 4

Results and Discussion

4.1. Introduction

The Results and Discussion section presents a detailed analysis of the web scraping process specifically tailored for hotels. This section encompasses various aspects, including cost estimation, feasibility study, results of implementation, result analysis, and observations/remarks. Each subsection provides valuable insights into the challenges, methodologies, outcomes, and implications of web scraping in the context of hotel information .



4.2. Feasibility Study

- Evaluating Website Structures: Assessing the complexity and variability of hotel websites to determine if scraping is technically feasible.
- Data Availability: Investigating the availability and accessibility of hotel data on target websites.
- Scraping Tools: Testing the suitability and effectiveness of web

scraping tools and libraries for extracting hotel information.

4.3. Results of Implementation

- Data Collection: Details on the data collected, such as hotel prices, room availability, amenities, location information, and customer reviews.
- Scraping Techniques: Explanation of the specific scraping techniques used, such as HTML parsing, API integration, or browser automation.
- Data Processing: Steps taken to clean, preprocess, and organize the scraped data for analysis.
- Data Quality: Assessment of data accuracy, completeness, and reliability.

4.4. Result Analysis

- **Statistical Analysis:** Using statistical methods to identify reviews, trends, rates, and anomalies in the hotel data.

- **Comparison:** Comparing scraped data with external sources, industry standards for more precise data.

4.5. Observation/Remarks

- Insights and Opportunities**:** The analysis of scraped data revealed valuable insights into pricing trends, customer preferences, market dynamics, and competitive landscape in the hotel industry. These insights can inform strategic decision-making, pricing strategies and customer experience enhancements for hotel businesses.

Chapter 5

Conclusion

5.1. Conclusion

- Web scraping for hotels streamlines data access, enabling users to make informed decisions by automating the collection of hotel information.
- It enhances the user experience by simplifying the search and comparison process, improving efficiency in booking accommodations.
- Compliance with legal and ethical standards is crucial for maintaining trust and integrity while utilizing web scraping technology.
- Overall, web scraping offers valuable insights into market trends and customer preferences within the hospitality industry, facilitating better strategic decision-making for businesses.

5.2. Future Scope

- Enhancing data validation and cleaning processes to improve data accuracy.
- Developing advanced scraping algorithms to handle dynamic content and anti-scraping measures.
- Exploring the impact of web scraping on the hospitality industry, hotel industry, including market trends, competitive analysis, and customer behavior insights.

5.3. Reference Paper

- https://www.researchgate.net/publication/367719780_Web_Scraping_Techniques_and_Applications_A_Literature_Review
- https://www.publications.scrs.in/uploads/final_menuscript/863dc5628ae9215e611c22943d061742.pdf
- https://www.publications.scrs.in/chapter/978-93-91842-08-6/38
- https://www.researchgate.net/publication/367719780_Web_Scraping_Techniques_and_Applications_A_Literature_Review
- https://www.researchgate.net/publication/357401723_Web_Scraping_or_Web_Crawling_State_of_Art_Techniques_Approaches_and_Application

Bibliography

1. CHATGPT: https://chat.openai.com
2. CODEWITHHARRY:
   Web Scraping Tutorial Using Python | BeautifulSoup Tutorial
3. GFG: https://youtu.be/O6nnVHPjcJU?si=wPdjZuXREwZt8FWr