LEVERAGING WEB SCRAPING FOR COLLECTING COMPETITIVE MARKET DATA

Case: A case study of an Airbnb rental unit in Helsinki

LAB UNIVERSITY OF APPLIED
SCIENCES LTD
Bachelor of Business Administration
Degree programme in Business
Information Technology
Autumn 2020
Thao Ho

Abstract

Author(s) Ho, Hoang Phuong Thao	Type of publication Bachelor's thesis	Published Autumn 2020
	Number of pages 45	
Title of publication **Leveraging Web Scraping For Collecting Competitive Market Data** Case: A case study of an Airbnb rental unit in Helsinki		
Name of Degree Bachelor of Business Administration		

Abstract

In recent years, web scraping has become a popular technique to gather large amounts of data from many sources.

The study focused on the exploration advantages of web scraping in the market research process for a small rental business. The business has existing market research that still contains several manual steps and limitations inaccuracy. The objective of this thesis was to introduce and examine how web scraping could support a small rental business in the market research process. The research framework for this thesis was a constructive research approach that helped to fill the gap between theoretical background and practical problem. The definition of web scraping was applied in the market research process, and the implementation process used both qualitative and quantitative research methods.

The thesis was carried out within the scope of a small rental business. The research applied qualitative research by having interviews with potential users of the case study, reviewed the existing market research process, and then analyzed the process's performance with web scraping. The interviews were adopted to collect the views of participants about the web scraper and their feedbacks.

By applying web scraping, the author was able to identify the advantages gained in time productivity and efficiency in the market research process. Moreover, the study presented the implementation of a simple web scraper with Python for small rental businesses. The implementation process was reviewed by potential users and adjusted to meet the case study's needs better.

Keywords

web scraping, web scraper, data, market research, Beautiful Soup

CONTENTS

LIST OF ABBREVIATIONS
1 INTRODUCTION 1
2 RESEARCH DESIGN 2
2.1 Thesis motivation 2
2.2 Research questions 2
2.3 Research limitations 4
2.4 Research approach 5
2.5 Research methods 6
2.6 Data collection and analysis 7
2.7 Thesis structure 8
3 THEORETICAL BACKGROUND 10
3.1 Description of existing market research 10
3.1.1 Define objectives 10
3.1.2 Develop a market research plan 10
3.1.3 Collect relevant data and information 11
3.1.4 Analyze data 11
3.2 Introduction to web scraping 11
3.3 Web scraping components 12
3.3.1 Crawler 13
3.3.2 Parse and extract 13
3.3.3 Format 13
3.3.4 Storage module 13
3.4 Web scraping frameworks 13
3.4.1 BeautifulSoup 14
3.4.2 Scrapy 14
3.4.3 Lxml 14
3.4.4 Other frameworks 15
3.5 Web scraping as a service 15
3.6 Website structure 15
3.7 Legal Issues 17
3.8 Web scraping in the market research process 18
4 IMPLEMENTATION 19
4.1 Introduction to case study 19
4.2 Project goals 19

4.3 Project requirements 20
4.4 Implementation 21
4.4.1 Application design 21
4.4.2 Target objectives 22
4.4.3 Development process 22
4.5 Figures of implemented web scraper 25
4.6 Results and testing 27
5 COLLECTION OF DATA 30
5.1 Interview 30
5.2 Design the structure of the interview. 30
5.3 Interview results 31
6 CONCLUSIONS 36
6.1 Answering research questions 36
6.2 Demonstration of web scraping solution working 37
6.3 Examination of applicability scope 38
7 SUMMARY 39
7.1 Reliability and validity 39
7.2 Suggestions for further development 40
LIST OF REFERENCES 41
APPENDICES 45

LIST OF ABBREVIATIONS

AI – Artificial Intelligence

CRA – Constructive Research Approach

CSS – Cascading Style Sheets

HTML – Hypertext Markup Language

URL – Uniform Resource Locator

UI – User Interface

1 INTRODUCTION

According to IBM, we create 2.5 billion gigabytes of data every day, and about 75 percent of all the data is unstructured from different sources such as text, voice, and video (IBM 2014). Hence the internet becomes an essential source of information. The explosion of data is transforming the industry that we know today. One of the key data processing techniques is web scraping. Web scraping can be used to explore valuable information from the internet, such as news, online products, or the power of product price to increase purchasing ability, understand customer behavior, and improve marketing strategy.

With the massive amount of data on the internet, web scraping has become one of the best approaches to draw out data from different resources. Web scraping is a technique for extracting unstructured data from websites to a structured format that is ready for analysis (Koshy 2018). Business owners and organizations could use web scraping to gather information from online publications and social media platforms.

In addition, during the Covid-19 pandemic, web scraping is more crucial than ever before. Many businesses are trying to survive, web scraping and data analysis could be seen as a necessary tool for supporting the decision-making process and make it through the crisis. McKinsey (2017) states that businesses with advanced market research techniques can take advantage of 85 percent in sales growth and more than 25 percent in gross margin. Data can strengthen business decision and strategy planning so that businesses around the world determine data as their key competitive advantage to catch up with competitors.

This thesis is carried out to explore how web scraping supports the market research for Airbnb hosts that is highly relevant to the decision-making process in order to overcome difficulties in the Covid-19 situation. Also, the thesis aims to build the web scraper as a suitable solution to meet the case study’s requirement. To find out the benefits of web scraping in research landscape, the research analyzes the performance of this technology in the market research.

2 RESEARCH DESIGN

This chapter presents the research's motivation, research questions, the limitations of the study, the research approach, and the structure of the thesis. Also, the data collection and analysis of this thesis are discussed in this chapter.

2.1 Thesis motivation

Data analytics is now widely used in various industries, also in the hospitality industry. It is becoming one of the most critical aspects for companies to take a competitive advantage in the market. For instance, from 2013 to 2014, Red Roof Inn, a hotel chain in the United States, could predict flight cancellations and launched an effective marketing campaign by gathering public weather data and flight information. (Bhattacharjee et al. 2017)

The hospitality industry includes hotels and relative services with the same concept, for example Airbnb could benefit from using web data. There are currently many aggregation websites that offer all information regarding accommodation, such as Tripadvisor.com, Kayak.com, Airbnb.com, and Booking.com. Technology likes web scraping can help business owners gather and analyze data from these sources.

Python is considered the most popular programming language for web scraping when it could smoothly handle the data extraction process and provide a wide selection of web scraping frameworks (Koshy 2020). Scrapy and BeautifulSoup are two frameworks that are usually chosen to scrape data because of outstanding performance. This thesis reviews existing web scraping frameworks and selects the most suitable framework for creating the case study's web scraper application.

2.2 Research questions

The thesis discusses the traditional market research of rental businesses, then finds some effective ways to increase the efficiency of the process. The primary goal of this thesis is to explore the advantages of web scraping to the market research process through a case study of building a web scraping application for Airbnb hosts to collect data in the market. In specific, collecting competitors' data is helpful for one business owner to adjust the price, improve the customer experience as well as marketing campaign planning. This research aims to create a solution for the case study by answering the main research question as stated below:

- **How can the web scraping technologies support the market research process for rental property business?**

In order to solve the main research question and provide an easy-to-follow path for the research, sub-research questions have been formulated:

- **How does the existing market research process of the case study look like?**
- **How does a web scraper service help to collect competitive market data?**
- **What are the stages of building a simple web scraper as a service?**

The first question aims at reviewing the current market research, its process, and existing problems for the case study. The second question explores how one web scraper works and how to apply web scraping to extract competitive data from public sources. Based on the case study's requirements and together with the theoretical background, the third question finds out needed steps to create a web scraping as a service from scratch with the programming language. With the web scraper built for the case study, the research may apply the solution to the existing process and detect benefits that web scraping can support the market research process for rental property businesses.

Currently, there are different available technologies of web scraping that could be applied in business models. The functions and capabilities of web scraping tools depend on business requirements. Emerson (2019) defines some of the potential web scraping tools that could be applied in the market research process:

- Browser plug-in tools: Plug-ins could be installed and used directly in browsers such as Chrome, Firefox, Microsoft Edge. However, these tools require manual works from users that will select the location of information that they want to scrape.

- Desktop applications: Web scraping desktop applications are famous for the friendly user interface (UI) that users could easily get familiar with functions and working flows. With scraping applications, it does not require knowledge of HTML and CSS. Hence, web scraping desktop applications are widely used by enterprises with paid subscription fees.
- Programming languages: a web scraper could be built with programming languages such as Python, R, or C#. By developing web scraping solution with programming languages, it requires investments of time and learning as well as could be scalable.

As the thesis aims to find a possible solution for a small rental business with a limited budget and the author had a technical background, the study will implement the web scraping solution with programming languages and measure how the built-up web scraper could affect the market research process.

2.3 Research limitations

The topic could be seen from different point of views. Therefore, this thesis only focuses on the rental business context and the Finnish market. Moreover, the research is targeted at the specified case study and may not apply appropriately to other cases. However, it can be applied to other rental businesses in the same market segment.

The data used for evaluating the supports of web scraping to the market research process is mainly collected from the case study. The research analyzes the performance of having web scraping in the existing market research to find out the benefits of this technology. Therefore, the data analysis of this research will not be compared to other studies.

The thesis also introduces fundamental steps to build the web scraper from scratch that are required to have basic knowledge of HTML and CSS. However, web scraping services are also mentioned in the research. These services are easy for the business owner to purchase and quickly start working with web scraping technologies.

2.4 Research approach

This thesis aims to apply web scraping technologies in the market research process to gather data for Airbnb hosts. Hence, the author decided to implement a constructive research method that is used to fill the gap between academic theory and practical problems.

An alternative to a constructive research approach could be design science research, but as CRA includes market testing in the implementation stage (Lindholm, 2008), CRA was chosen. Kasanen (1993) presents that the constructive research approach should always start with a practical research problem and be followed by the below steps:

1. Obtain the necessary knowledge of the study field.
2. Develop applicable solution ideas for the problem.
3. Implement the solution idea and test its feasibility.
4. Demonstrate the relationship between the theoretical foundation and solution contribution.
5. Evaluate the applicable results.

According to Pasian (2015), neither deductive nor inductive reasoning could cover the constructive research approach fully. Therefore, abductive reasoning is used to replace inductive and deductive processes by adopting a pragmatist perspective. Image 1 describes the formulation of the abductive reasoning approach.

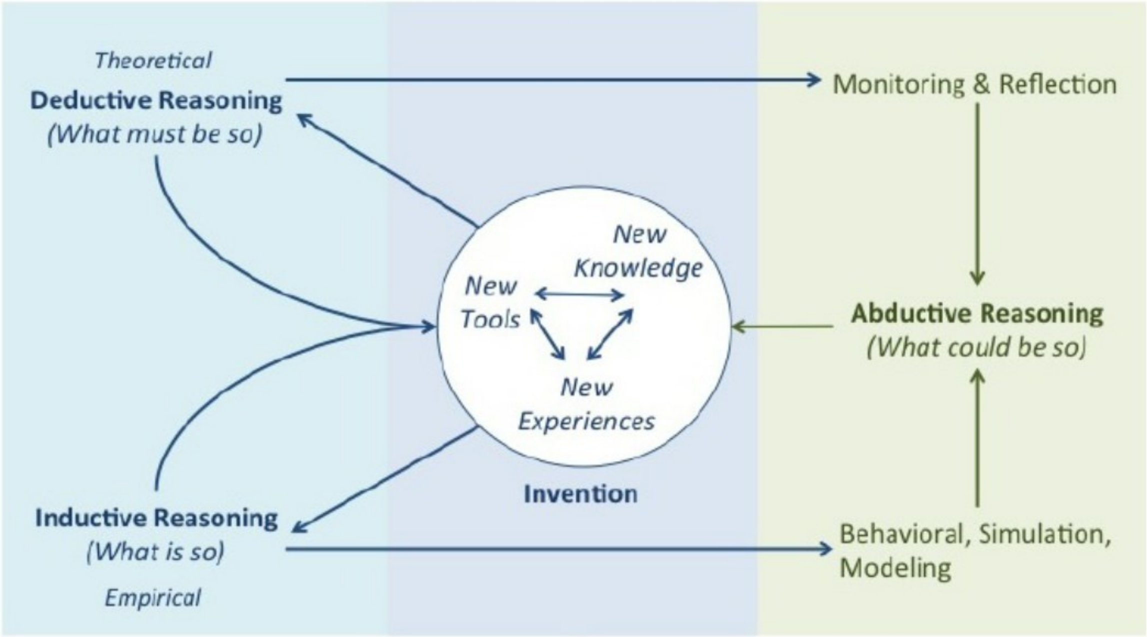


Image 1 The formulation of abductive reasoning approach (Samim 2020)

In this paper, the thesis starts with reviewing existing web scraping frameworks, then develops and tests these options to build a scraper for a case study that was mentioned in the thesis motivation part. As several concerns arise during the study case, abductive is used to explain these redirections. Following the abductive reasoning approach, the research aims to evaluate applicable results as well as reflect on the relationship between theories and practical work.

2.5 Research methods

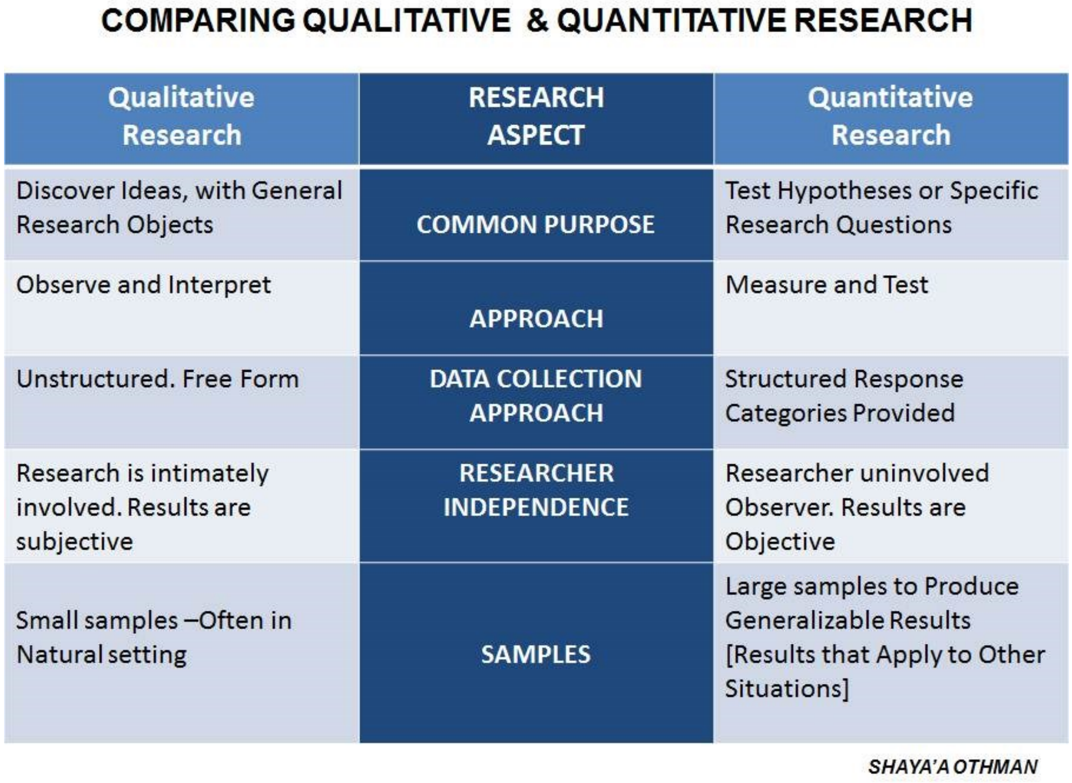
There are three main research methods to demonstrate the solution in the constructive research approach: qualitative research, quantitative research, or the combination of two of them. (Oyegoke 2011)

There are two common research methodologies: qualitative and quantitative. They depend on ways of collecting data and analysis as well as they help to answer different kinds of research questions. Qualitative research is presented under words to understand single concepts or phenomena, and interviews play an important role in executing this research method. Miles and Huberman (1994) also mention that qualitative research provides relevant ways to draw conclusions and verify it.

Alternatively, quantitative research expresses data with numbers and graphs to verify

theories and assumptions. While quantitative research is used to generalize facts about a topic, qualitative research help researchers to collect in-depth insights on topics that are not well understood. The different methods in qualitative research and quantitative research are clarified in the chart below.

Table 1. The differences between Qualitative and Quantitative Research (Othman 2011)



COMPARING QUALITATIVE & QUANTITATIVE RESEARCH

Qualitative Research	RESEARCH ASPECT	Quantitative Research
Discover Ideas, with General Research Objects	COMMON PURPOSE	Test Hypotheses or Specific Research Questions
Observe and Interpret	APPROACH	Measure and Test
Unstructured. Free Form	DATA COLLECTION APPROACH	Structured Response Categories Provided
Research is intimately involved. Results are subjective	RESEARCHER INDEPENDENCE	Researcher uninvolved Observer. Results are Objective
Small samples –Often in Natural setting	SAMPLES	Large samples to Produce Generalizable Results [Results that Apply to Other Situations]

SHAYA'A OTHMAN

2.6 Data collection and analysis

This research presents a case study of applying a web scraper to the existing market research process for Airbnb hosts to collect competitive market data. The goal is to evaluate the support of the web scraping technology to the case study as well as generate a list of suggestions for further improvement.

After collecting the necessary knowledge for the research, the research gathered data through interviews with Airbnb hosts in the Helsinki region who are potential users of the case study. The main purpose of the one-to-one interview was to gather the user's opinions on the implementation of web scraping to the market research process. The interview section consists of different open-ended questions that require specific responses. Therefore, the qualitative research method was adopted to collect and analyze these open-ended questions. Also, the quantitative research method was applied to compare the performance of the existing market process and the new one with web scraping. The performance could be estimated based on data volume, time, and speed. These collected data aimed to help answer the thesis' research questions and provided ideas to improve the case study.

2.7 Thesis structure

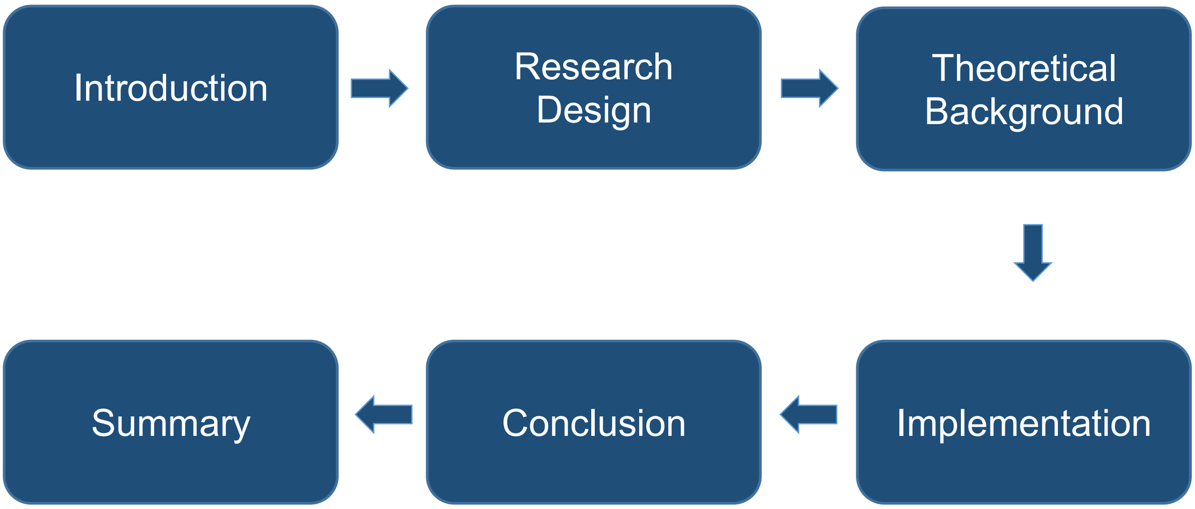


Figure 1. Thesis Structure

Figure 1 presents the structure of this thesis. The thesis is divided into six chapters. In Chapter 1, the thesis introduces the topic and research background. Chapter 2 demonstrates the motivation why the thesis is conducted, identifies the main objectives, and the foundation of research questions. Also, the research approach, research method, data collection, and the limitations of this study are explained in this chapter. Next, Chapter 3 explores the theory of the research. This section focuses on web scraping, its framework (Scrapy and BeautifulSoup), and legal issues. These information helps build the theoretical foundation of how web scraping could support the market search process for small rental businesses. Chapter

4 shows the case study, the implementation, the result, and explains how the performance of the implementation could help answer research questions. Chapter 5 presents the data collection as well as presents the results of the analysis. Chapter 6 finalizes the whole research and answers initial research questions. At the end of this chapter, several suggestions for further development are also given. Chapter 7 summarizes the whole research and provides research conclusions.

3 THEORETICAL BACKGROUND

This chapter introduces technical terminologies and fundamental components related to web scraping, including techniques, components, existing frameworks, and processes that are applied to this thesis as well as how they are combined to execute.

3.1 Description of existing market research

Currently, market research is one of the crucial processes in the rental business. Market research is a term of the process includes gathering, analyzing, and communicating with information about a product or a service that is offered in the market (Entrepreneur Europe). This process helps the Airbnb owners to keep an eye on the market and understand how their competitors are doing. Figure 2 presents the market research process is divided into four steps below.

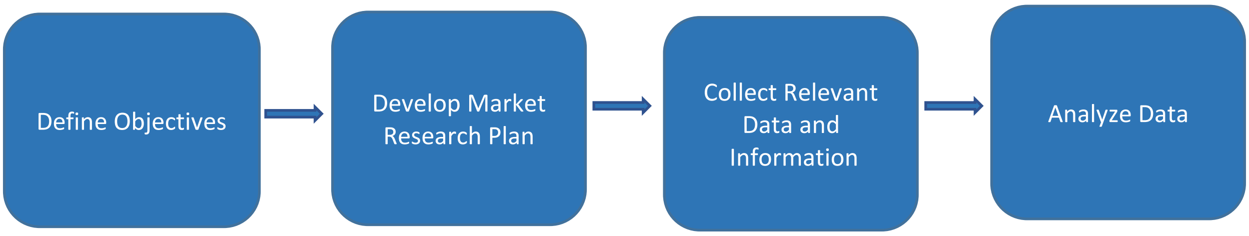

Define Objectives
Develop Market Research Plan
Collect Relevant Data and Information
Analyze Data

Figure 2. Market Research Process

3.1.1 Define objectives

In order to start the market research process, it is important to develop questions that allow defining problems or opportunities. Regarding rental businesses, some of their common concerns are location, pricing, level of occupancy, customers' review, and ratings. By understanding the objective clearly, it helps to keep the research process focused and effective.

3.1.2 Develop a market research plan

The market research plan defines the necessary steps and sources for carrying out the research. This step consists of a written plan that outlines the management problem, defines specific research approaches, required data, sampling plans, the budget for the research, and how the results support the decision-making.

3.1.3 Collect relevant data and information

This is the core of the market research, which is used to collect data through public sources, surveys, conducting interviews, field testing, etc. Most of the collected data will be answers, choices, observations that are then stored in spreadsheet format for further analysis (Market Research Guy 2020). The collection of data plays an important role in decisions that businesses will make later. However, it will take a lot of time if one business wants to extract a large amount of competitors' data from public sources by copying and pasting data manually.

3.1.4 Analyze data

After gathering all relevant data, the last step aims to make the collected data speak through structured tables, visualizations with the help of Excel or business intelligence tools such as Power BI, Tableau, Google Analytics. Analyzing data aims to look for trends of the market, specific conclusions that support the last decision.

3.2 Introduction to web scraping

Web Scraping is known as a technique used to extract data from multiple sources to simple flat files or databases that make further analysis and visualization easier (Sirisuriya 2015, 135). In the past, the only method to collect data from websites is copy-and-paste data manually. Good practical examples of web scraping could be competitive market research, price intelligence, and content monitoring.

Web scraping is also known as web harvesting, web data extraction or web data mining can be defined as a technique used to extract data from websites through an automated process. In other words, this process aims to copy data automatically from websites and deliver them to a database that could be structured or unstructured database. Generally, web scraping could be executed by writing an automated program which is a web scraper. The web scraper queries a web server and request data under format like HTML and XML format, then parses that data to extract needed information. Figure 3 illustrates the structure of the web scraping process.

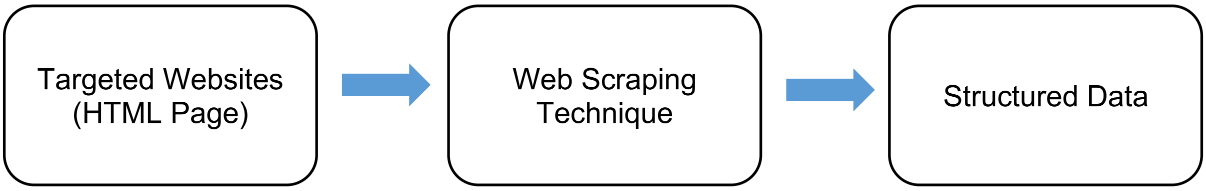


Figure 3. Web scraping process

Web scraping has been a popular technique in the business world that includes not only big corporations but also startups. As data is a priceless asset for companies and organizations, web scraping allows the business owner to quickly gather vast amounts of data from online sources. This is the core of the market research and business strategy, for instance, gathering product reviews, social networking posts, or comparing competitive prices.

3.3 Web scraping components

There are four main components included in the web scraping: Crawler; Parse and Extract; Format; and Storage Module. Figure 4 presents how web scraping's components are used in the scraping process under simple steps.

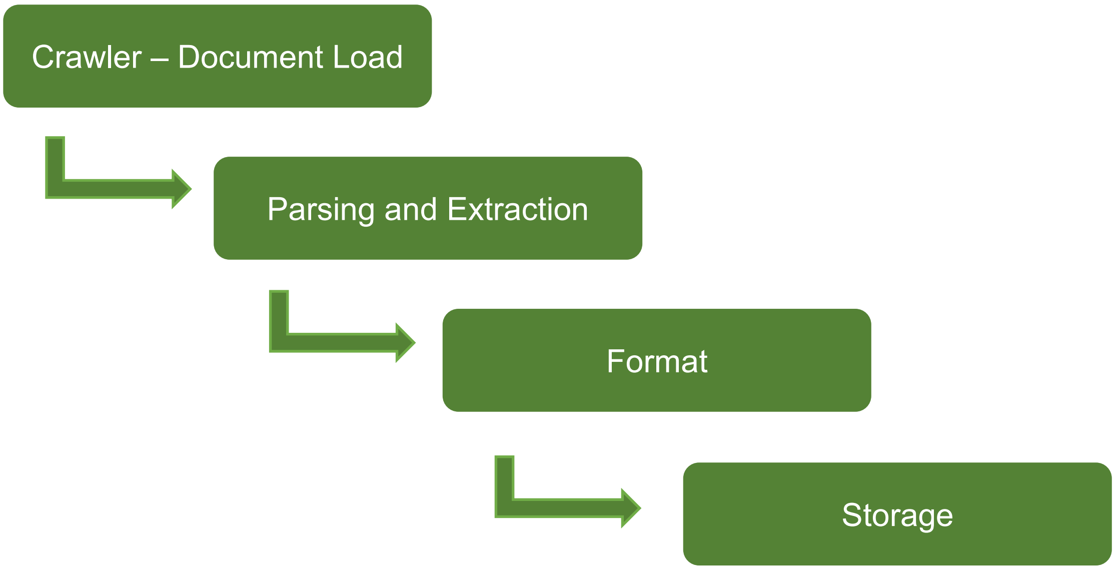


Figure 4. Web scraping working flow

3.3.1 Crawler

Web Scraping always firstly starts with a data source that navigates the target website by making HTTP Request to the URLs and downloads the unstructured data, for example, HTML content (Tutorials Point 2020). The goal of the crawler is to visit the web page and handle the discovery process (Scrape Hero 2018). One good example of the crawler is Turnitin.com’s robot that searches and gathers content information from the internet and helps higher education institutions prevent plagiarism. (Turnitin 2020)

3.3.2 Parse and extract

Extracting data is known as the process of extracting raw data in HTML format and parsing the relevant data element. Regular Expression, HTML Parsing, DOM Parsing are the main parsing techniques. This process could help users extract and parse data as their demands.

3.3.3 Format

The extracted data then needs to be transformed into a human-readable format. These data could be saved as CSV, JSON, XML form that depends on user purposes.

3.3.4 Storage module

In order to complete the web scraping process, a storage module is required. Once the data has been extracted and formatted, it could be stored in a standard format such as JSON and CSV files or a database such as MongoDB.

3.4 Web scraping frameworks

Web Scraping frameworks address most web scraper common tasks to achieve specific goals such as Site access; HTML parsing and contents extraction; Output building. Scraping frameworks present a more integrative solution. (Daniel, Anália, Hugo, Miguel & Florentino 2013, 791)

Web Scraping comprises many programming technologies, frameworks, and techniques. Some of the best options for each programming language are Selenium, Scrapy (Python), BeautifuSoup (Python), Goutte (PHP), etc. In this section, the thesis focuses on discussing Scrapy, BeautifulSoup, and how they can be applied in variety situations.

3.4.1 BeautifulSoup

BeautifulSoup is a pure open-source Python library developed by Leonard Richardson, which is mainly used for HTML and XML file parsing. It provides methods to search elements and extract structured data from a website and saves programmers a lot of time since it could collaborate with common parsers such as lxml or html5lib to provide a solution to navigate, search and modify the parse tree.

Beautiful Soup is a more streamlined version than its brother, which is Scrapy and could be used for simpler web scraping projects. It can parse HTML and XML documents and bring simple methods to work with DOM (Document Object Model). In addition, Beautiful Soup has the capacity of exporting fetched documents to Unicode or UTF-8.

3.4.2 Scrapy

Scrapy is a web scraping framework for Python and useful for various of purposes such as data mining, information processing, or historical archival. Scrapy is more powerful than Beautiful Soup as it is usually used for large scale web scraping.

Scrapy offers a complete package for downloading web pages, processing inside content, and converting to readable files or databases. Developing web scraper with Scrapy is quite easy to follow, the important thing is to create the parse() function to deal with downloaded data and return scraped data.

3.4.3 Lxml

Lxml is a Python library and is one of the fastest libraries for processing XML and HTML. This library is built on top of the libxml2 XML parsing library written in C which combines the speed of the native C library and the simplicity of Python (Fatenaite 2020). Python lxml can be used to create XML and HTML documents, find specific elements.

3.4.4 Other frameworks

Selenium

In the beginning, Selenium is widely used for website testing (Mottet 2018). Also, this library could be used as a web scraping library and allows users to control a web browser and automate different events such as clicking and scrolling. While Scrapy or Requests cannot work well with websites written in JavaScript, Selenium has good performances with JavaScript events and can control web browsers such as Chrome, Safari, or Firefox by combining with WebDriver protocol.

3.5 Web scraping as a service

Web scraping as a service is a term for executing the data extraction without required coding knowledge. From big corporations to startups and tiny businesses, web scraping as a service provides several solutions such as price tracking, social media monitoring, information aggregation. The service allows users to collect data from online sources within a few clicks in the same way as using softwares. Some of the common web scraping services are Octoparse.com, Parsehub.com, and Scrapinghub.com.

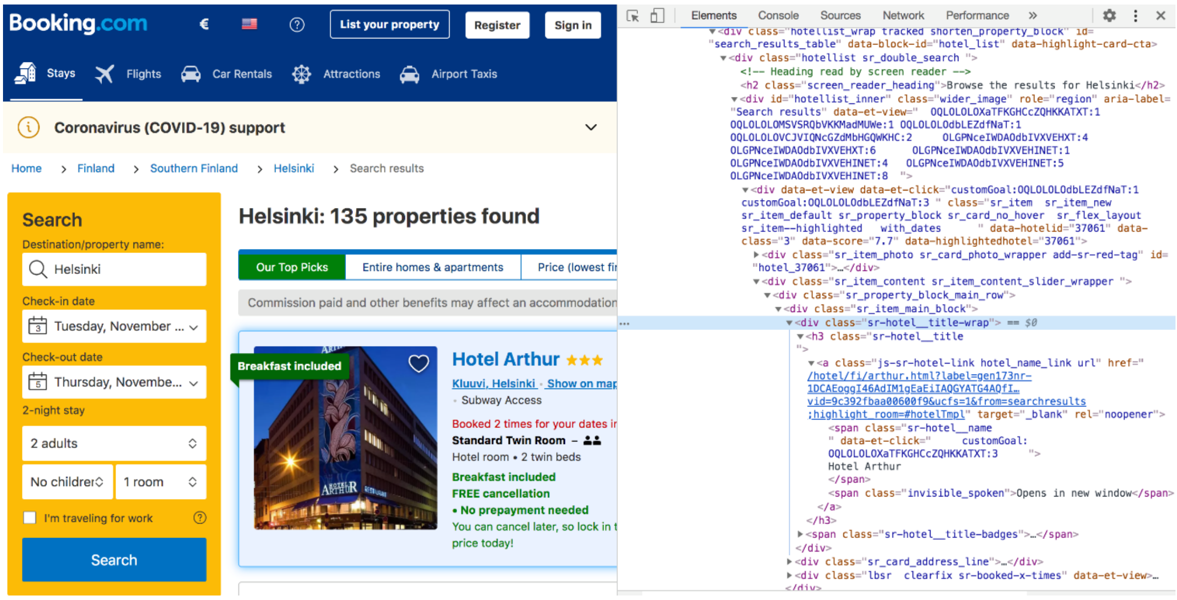
One of the big limitations of web scraping service is pricing. The cost depends on the numbers of websites to be scraped, complexity, frequency of scraping, maintenance as well as customer support service. While big companies could consider web scraping service to be a good option to replace humans, small businesses need to work on the budget. For these reasons, this thesis aims to create a simple web scraper self-service that shares the difficulties with small businesses in the hospitality industry.

3.6 Website structure

Web scraping approaches depend on website structures. Websites store data in two main ways that are Hypertext Markup Language (HTML) elements and Application Programming Interface (API).

Hypertext Markup Language (HTML)

Hypertext Markup Language (HTML) could be seen as a backbone that describes the structure of the web page(w3schools). As data is stored under HTML elements, it is important to have a basic understanding of how the web page is structured to extract desired information. The structure could be checked by inspecting the website. Figure 5 presents the structure of the Booking.com website that is the targeted website used in the case study.



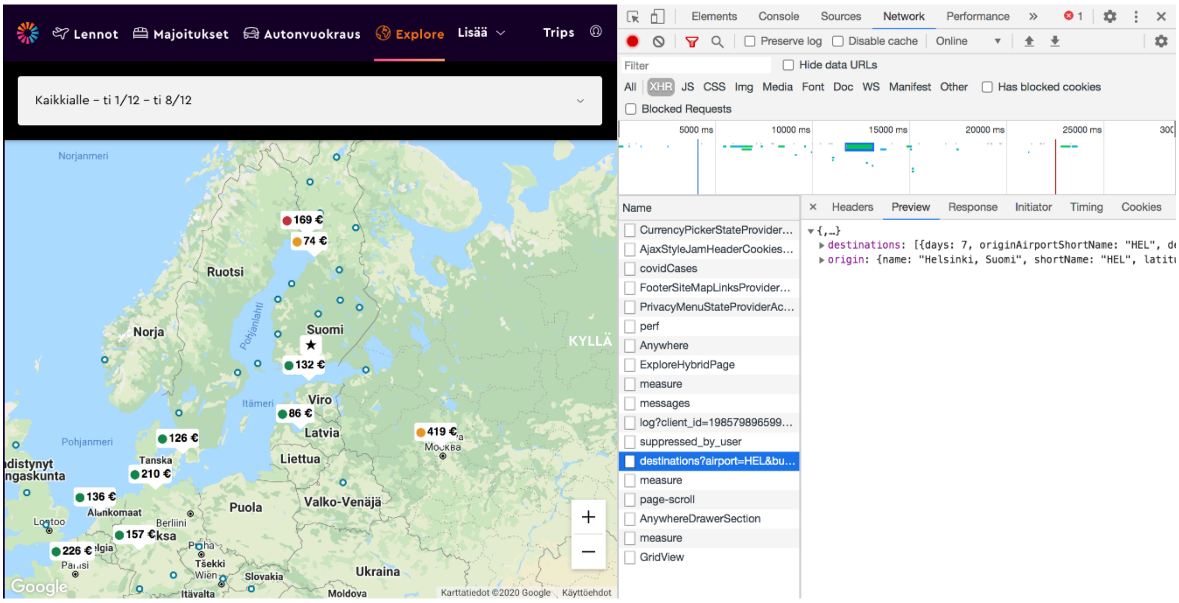
```
▼<div class="hotellist_wrap tracked shorten_property_block" id=
"search_results_table" data-block-id="hotel_list" data-highlight-card-cta>
  ▼<div class="hotellist sr_double_search ">
     <!-- Heading read by screen reader -->
     <h2 class="screen_reader_heading">Browse the results for Helsinki</h2>
    ▼<div id="hotellist_inner" class="wider_image" role="region" aria-label=
    "Search results" data-et-view="  OQLOLOLOXaTFKGHCcZQHKKATXT:1
    OQLOLOLOMSVSRQbVKKMadMUWe:1 OQLOLOLOdbLEZdfNaT:1
    OQLOLOLOVCJVIQNcGZdMbHGQWKHC:2     OLGPNceIWDAOdbIVXVEHXT:4
    OLGPNceIWDAOdbIVXVEHXT:6      OLGPNceIWDAOdbIVXVEHINET:1
    OLGPNceIWDAOdbIVXVEHINET:4   OLGPNceIWDAOdbIVXVEHINET:5
    OLGPNceIWDAOdbIVXVEHINET:8  ">
      ▼<div data-et-view data-et-click="customGoal:OQLOLOLOdbLEZdfNaT:1
      customGoal:OQLOLOLOdbLEZdfNaT:3 " class="sr_item  sr_item_new
      sr_item_default sr_property_block sr_card_no_hover  sr_flex_layout
      sr_item--highlighted   with_dates      " data-hotelid="37061" data-
      class="3" data-score="7.7" data-highlightedhotel="37061">
        ▶<div class="sr_item_photo sr_card_photo_wrapper add-sr-red-tag" id=
        "hotel_37061">…</div>
        ▼<div class="sr_item_content sr_item_content_slider_wrapper ">
          ▼<div class="sr_property_block_main_row">
            ▼<div class="sr_item_main_block">
...             ▼<div class="sr-hotel__title-wrap"> == $0
                ▼<h3 class="sr-hotel__title
                ">
                  ▼<a class="js-sr-hotel-link hotel_name_link url" href="
                  /hotel/fi/arthur.html?label=gen173nr-
                  1DCAEoggI46AdIM1gEaEiIAQGYATG4AQfI…
                  vid=9c392fbaa00600f9&ucfs=1&from=searchresults
                  ;highlight_room=#hotelTmpl" target="_blank" rel="noopener">
                      <span class="sr-hotel__name
                      " data-et-click="     customGoal:
                      OQLOLOLOXaTFKGHCcZQHKKATXT:3     ">
                      Hotel Arthur
                      </span>
                      <span class="invisible_spoken">Opens in new window</span>
                    </a>
                  </h3>
                ▶<span class="sr-hotel__title-badges">…</span>
                </div>
              ▶<div class="sr_card_address_line">…</div>
              ▶<div class="lbsr  clearfix sr-booked-x-times" data-et-view>…
              </div>
```

Figure 5. The structure of the Booking.com website

On the console dialogue, it presents the HTML structure with basic components such as div, class, and their hierarchy. From the website surface, the user can hover over the information they want to extract data and the div in the console gets highlighted.

Application Programming Interface (API)

In the case that the website stores data in API, it will call the API every time a user comes to the page. Then, the web scraper needs to create the request and extract data from the API. First, the XHR network needs to be inspected from the URL that would be scrapped. The next step is to detect the requested response that provides the required data. Figure 6 shows the structure of the XHR network that contains the list of request responses of destinations.



Kaikkialle – ti 1/12 – ti 8/12

```
Name
□ CurrencyPickerStateProvider…
□ AjaxStyleJamHeaderCookies…
□ covidCases
□ FooterSiteMapLinksProvider…
□ PrivacyMenuStateProviderAc…
□ perf
□ Anywhere
□ ExploreHybridPage
□ measure
□ messages
□ log?client_id=198579896599…
□ suppressed_by_user
□ destinations?airport=HEL&bu…
□ measure
□ page-scroll
□ AnywhereDrawerSection
□ measure
□ GridView
```

```
▼{,…}
 ▶destinations: [{days: 7, originAirportShortName: "HEL", de
 ▶origin: {name: "Helsinki, Suomi", shortName: "HEL", latitu
```

Figure 6. The XHR section under the network of Momondo.com

3.7 Legal Issues

Web scraping is not an illegal activity. Google is a good example for the legalization of web scraping, the group built the business based on web scraping, and it is still one of the biggest corporations in the world. However, there are important things that need to be carefully considered before starting the process. Firstly, it is essential to check the Robots.txt file. This file is widely used by websites to communicate with web scrapers and inform areas that the web scraper has the right to access (Narizhnykh 2018). The next thing is to respect existing law acts such as the General Data Protection Regulation (GDPR) in Europe and the US Privacy Act in the United States. Recently, the GDPR sets the law for protecting personal information which is used to identify a person. Specific personal data could be a real name, date of birth, email address, social security number, gender, and living address. Unlike European countries, the US does not set the rules for consumer data privacy at the federal level (DataOx 2020). However, the US is working on the preparation phase for data privacy and California recently passed state laws to protect personal data.

3.8 Web scraping in the market research process

As presented in the sub-chapter 3.1, the traditional market research process consists of four main stages: defining objectives, developing a market research plan, data collection, and analyzing data. In the process, manual data collection is a time-consuming and inefficient step in which web scraping replace repetitive tasks and automate the process. Figure 7 shows the integration of web scraping into the market research process.

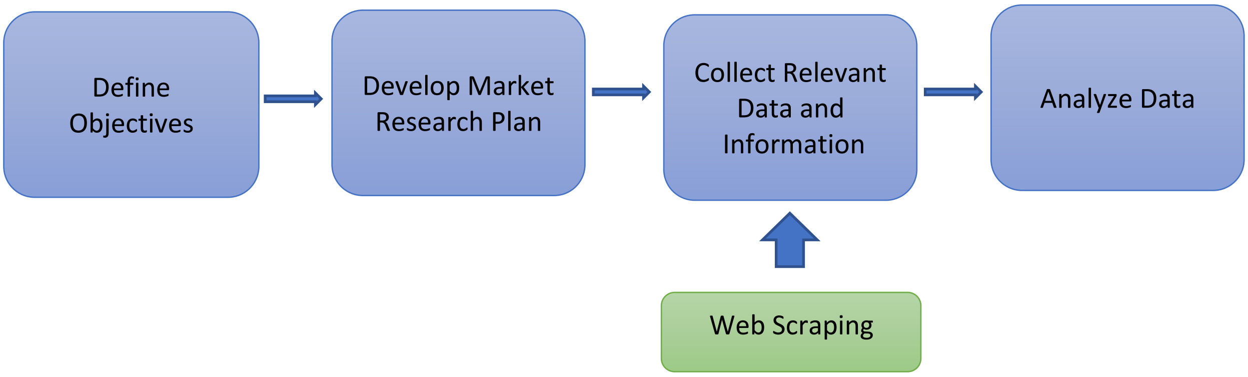

Define Objectives
Develop Market Research Plan
Collect Relevant Data and Information
Analyze Data
Web Scraping

Figure 7. Web scraping in market research process.

The web scraping could be integrated into collecting relevant data and information stage. Web scraping automates the process of data collection and allows users to save data in a format such as CSV, XLSX.

4 IMPLEMENTATION

As mentioned in Chapter 2, the constructive research was conducted for this thesis. This chapter outlines an overview of the case study, project's goals, its demands, followed by the discussion of the selected integration tools that are necessary parts of phase 2 and phase 3 of the constructive research approach to construct and implement solution ideas.

4.1 Introduction to case study

Abby Corner is an Airbnb listing based in Kamppi, Helsinki which was operated by two founders – the author and her partner. The listing has been running since April 2019 with the objective is to provide accommodation to tourists. Abby Corner was chosen as a case study of this research because the author is one of the founders of this listing. Through this study, the web scraping technology was applied to research the Airbnb market in Helsinki and expect the results could be beneficial for Abby Corner as well as other listings with the same concept.

Due to COVID-19 (coronavirus), the Airbnb business has been hit hard, and 80% of Abby Corner bookings have been canceled from May to September 2020. According to McKinsey (2020), it will take until 2023 or even later for the hotel industry to recover to the post-COVID-19 level. To overcome the pandemic situation and speed up the recovery, the author decided to create a web scraper to collect a large amount of public information about the hotel industry to gain a competitive advantage.

4.2 Project goals

The project aims to build web scraping as a simple service for two hosts to extract hotel listings data from the Booking.com website and transfer the results back to users. In other words, the web scraper automates steps such as searching, clicks, and scrolls on the website that could help minimize human interactions and reduce time consumption spent on market research. Below is the diagram of the scraper architecture:

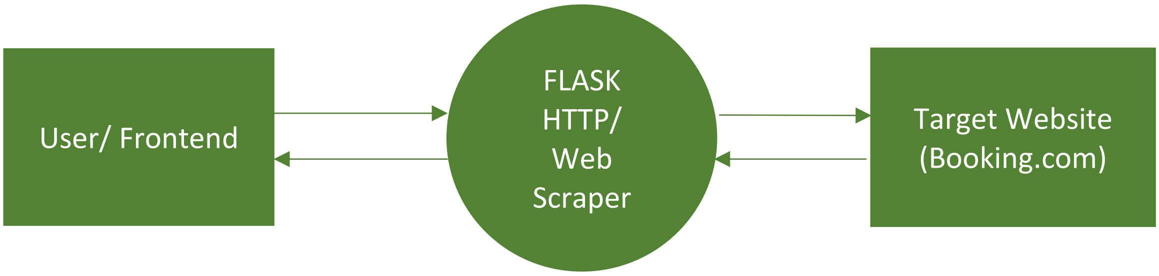


Figure 8. Basic architecture of web scraper application

4.3 Project requirements

The project collects functional requirements mainly from two hosts of Abby Corner that will be primary web scraper users. Besides, the requirements are also gathered from the first interview group. By implementing all functionalities below, the web scraper can be used by main users and other businesses that have the same difficulties in the market research process. Table 2 illustrates the necessary functional requirements for the web scraper application.

Table 2. Functional requirements

No.	Description	Priority (1 is highest)
1	The users can paste the website URL where they want to extract data.	1
2	The web scraper can detect website contents.	1
3	The user can add HTML tags and CSS classes to select data that need to be scrapped.	1
4	The user can export the result to CSV format.	1
5	The user can see the preview window of the result.	2

4.4 Implementation

This chapter describes the implementation of a custom web-scraper-as-a-service that is used to automate the collecting market data process to replace manual steps in the process and reduce repetitive tasks day-by-day. There are three main stages in this implementation part, which are designing, implementation, and testing.

4.4.1 Application design

The project will integrate the BeautifulSoup for scraping the site with Flask to build up a web form. Below is the first sketch-up of the scraper:

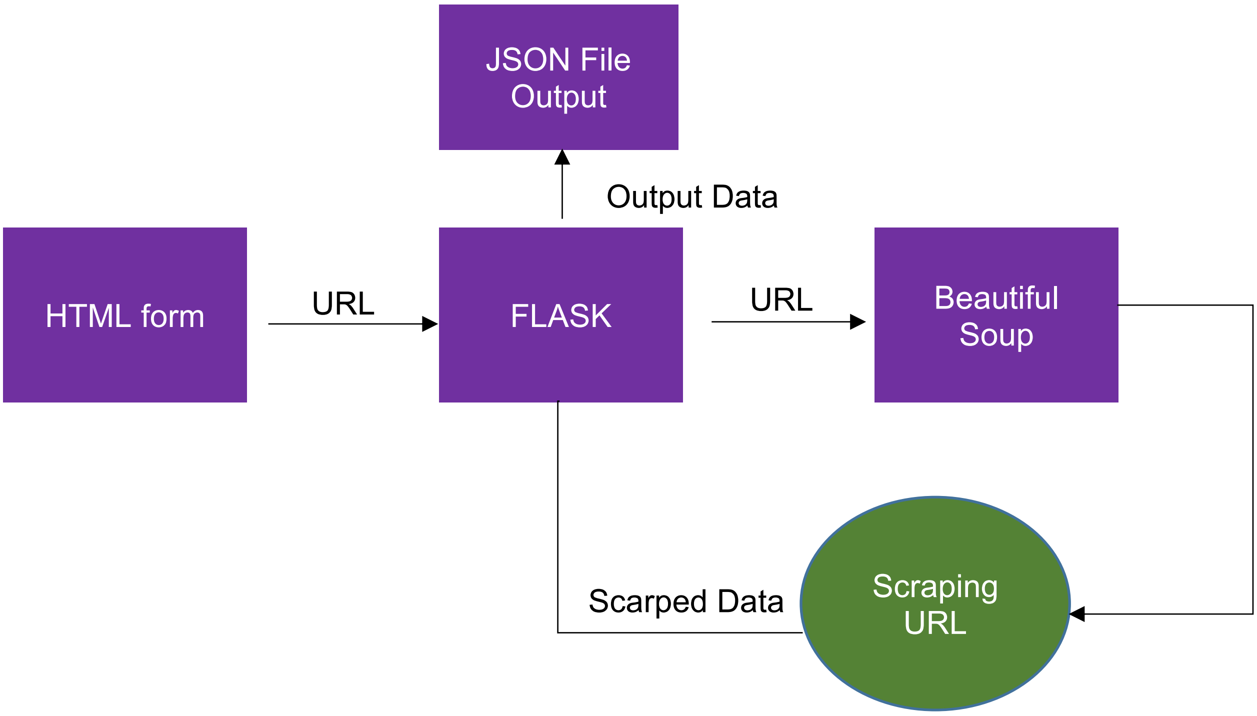
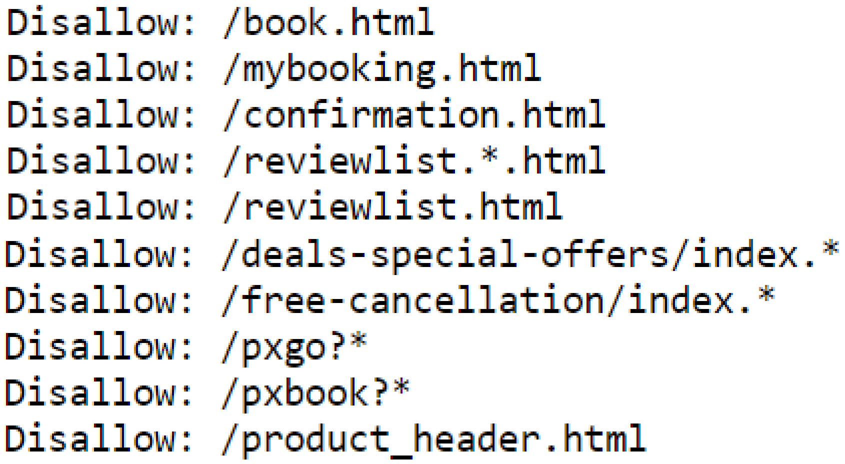


Figure 9. Sketch-up of web scraper

As could be seen from Figure 9, the application has a simple User Interface (UI) for users to enter the website URL contains data that needs to be extracted. Then the requested link is sent to a web scraping service, and the triggered script will handle searching, extracting steps, and send searched results back to the user. Finally, the user could export collected data to a flat file in CSV format.

4.4.2 Target objectives

The scraping target in this project is the Booking.com website. Booking.com is hotels, rentals, and accommodations aggregation website with the mission of providing easier experiences to travelers around the world (Booking 2020). According to the Booking.com website, it provides more than 28 accommodation listings together with over 6.2 million listings of individual places, apartments, homes, and currently available in 43 languages.



```
Disallow: /book.html
Disallow: /mybooking.html
Disallow: /confirmation.html
Disallow: /reviewlist.*.html
Disallow: /reviewlist.html
Disallow: /deals-special-offers/index.*
Disallow: /free-cancellation/index.*
Disallow: /pxgo?*
Disallow: /pxbook?*
Disallow: /product_header.html
```

Figure 10. A part of robots.txt from Booking.com

Before starting web scraping, it is essential to discover the targeted website’s structure and robots.txt to check any restrictions in order to minimize the chance of being blocked. As Booking.com had a long list of robots.txt, this thesis only focused on important sections are shown in Figure 10.

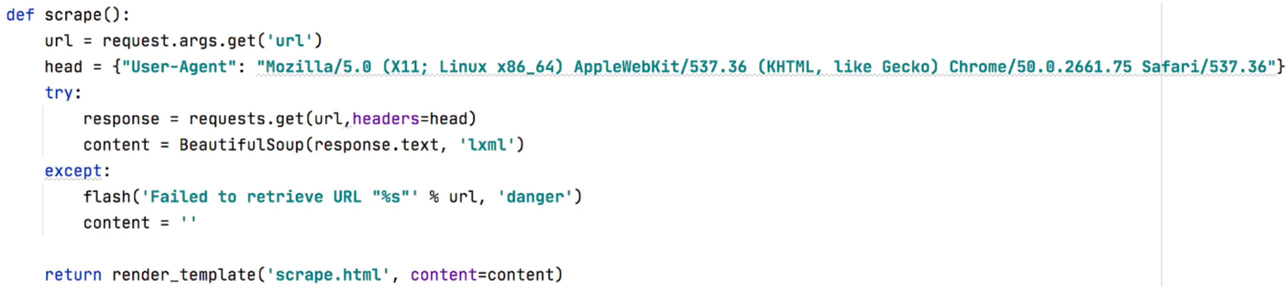
It could be seen from Figure 10 that Booking.com seems does not allow web scraping to work at /product_header.html and reviewlist.html. However, this case study aims to gather results at /searchresults.*.html which is not noted in the list of Booking.com’s robots.txt.

4.4.3 Development process

After completing the initial investigation, selected tools or frameworks are applied to build the web scraper application. The development stage is divided into two parts: scraping site and web service.

In terms of the scraping part, many existing frameworks that could perform web scraping that reviewed and summarized in Chapter 3. However, the difficulty of implementing a web scraping framework depends on the knowledge and skills of the user. To make this easier for scale-up or maintenance in the future, BeautifulSoup was selected as it is very friendly for beginners to learn and execute.

Building a web scraper with BeautifulSoup is simple and it can quickly extract data from websites. The most important parts of BeautifulSoup is to access the website content and find desired data that is contained inside HTML tags.

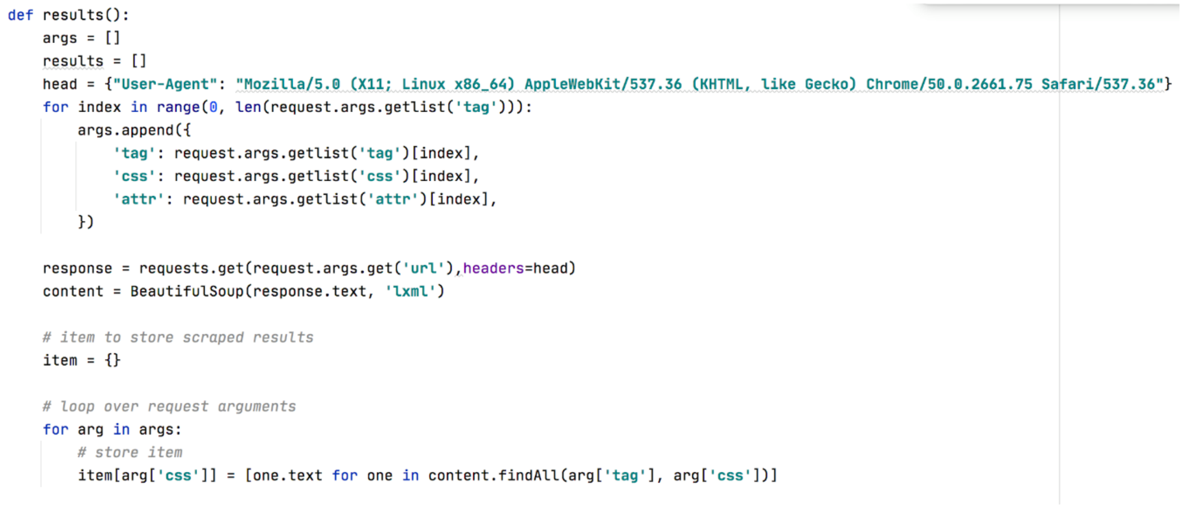


```
def scrape():
    url = request.args.get('url')
    head = {"User-Agent": "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/50.0.2661.75 Safari/537.36"}
    try:
        response = requests.get(url,headers=head)
        content = BeautifulSoup(response.text, 'lxml')
    except:
        flash('Failed to retrieve URL "%s"' % url, 'danger')
        content = ''

    return render_template('scrape.html', content=content)
```

Figure 11. Scrape function of the web scraper

Figure 11 shows the way to get access to the website with requests library and applying the BeautifulSoup method to parse it into HTML structure with the support of lxml.

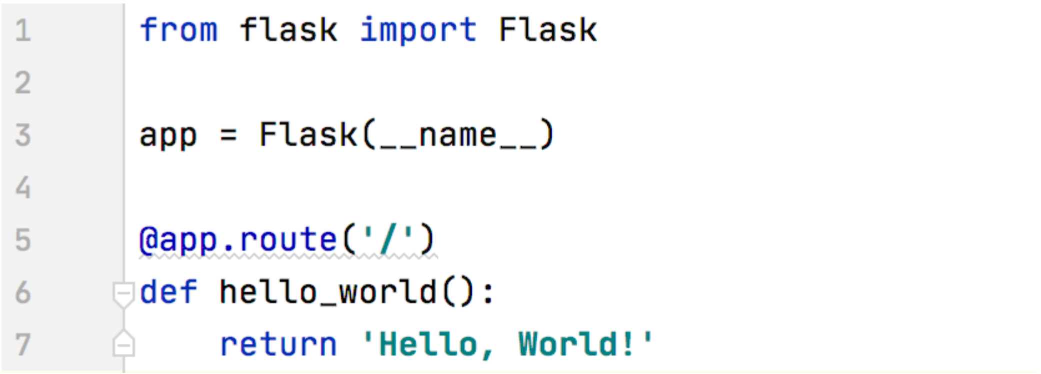
Next, the results function is built to search for HTML elements containing the desired data. It could be seen from figure 12 that the web scraper used findAll() method to go through all selected elements and extract the data.



```python
def results():
    args = []
    results = []
    head = {"User-Agent": "Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/50.0.2661.75 Safari/537.36"}
    for index in range(0, len(request.args.getlist('tag'))):
        args.append({
            'tag': request.args.getlist('tag')[index],
            'css': request.args.getlist('css')[index],
            'attr': request.args.getlist('attr')[index],
        })

    response = requests.get(request.args.get('url'),headers=head)
    content = BeautifulSoup(response.text, 'lxml')

    # item to store scraped results
    item = {}

    # loop over request arguments
    for arg in args:
        # store item
        item[arg['css']] = [one.text for one in content.findAll(arg['tag'], arg['css'])]
```

Figure 12. Results function of the web scraper.

For the web service, the project chooses Flask as a web framework as it is explicit and easy to get started. Flask is a web framework written in Python. Although Flask does not have any components that integrate third-party libraries, Flask supports extensions that easily include features (Kraczkowsky 2019). Some examples of extension include opening authentication, validation, upload handling. Flask allows user to develop and deploy a simple web application promptly with URLs and REST. Figure 13 shows an example of a simple application created with Flask.

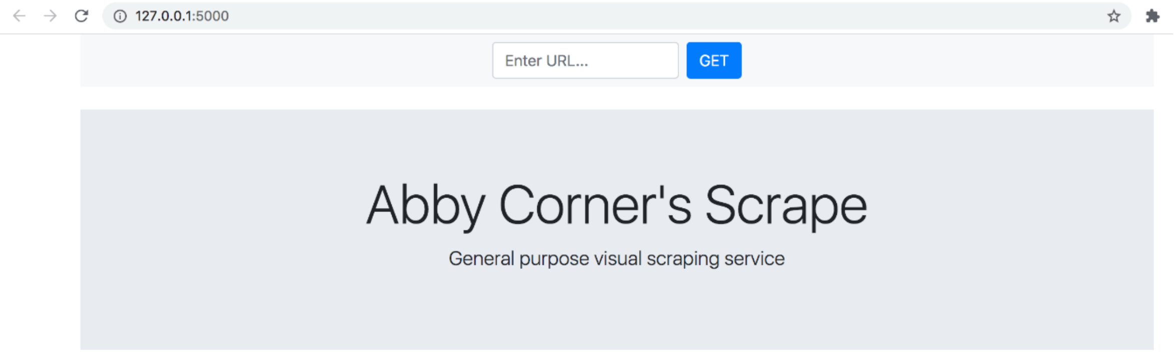


```python
from flask import Flask

app = Flask(__name__)

@app.route('/')
def hello_world():
    return 'Hello, World!'
```

Figure 13. A complete Flask application

4.5 Figures of implemented web scraper

The home page of Abby Corner’s web scraper provides a form for users to define the scraping objective. In order to start the web scraper, the user pastes the Uniform Resource Locator (URL) as a website address that user want to extract data to the box, then click on “Get” button.



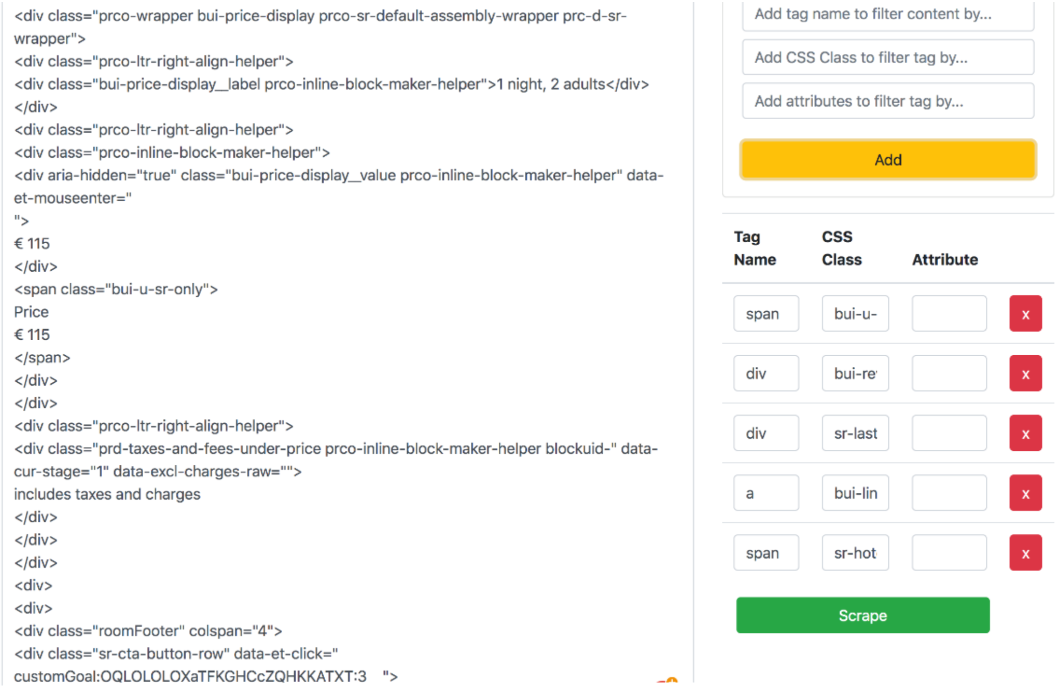
127.0.0.1:5000

Abby Corner's Scrape

General purpose visual scraping service

Figure 14. Home page of Abby's Corner scrape

Then the scraper detects web page data and shows a window like figure 15 below. The box on the right side contains setting options for data selection. Data is scraped from HTML structure through the tag and class. Therefore, the user needs to add inputs for the HTML tag name, CSS class, or other attributes and click the “Scrape” button to trigger the extraction.



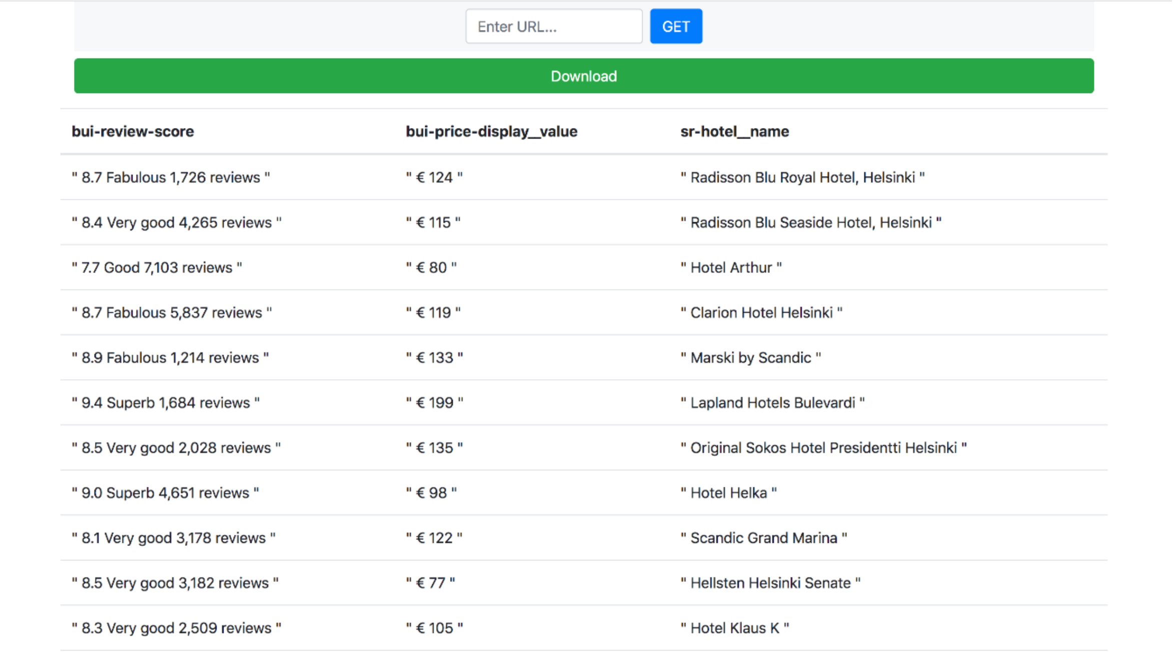
```
<div class="prco-wrapper bui-price-display prco-sr-default-assembly-wrapper prc-d-sr-
wrapper">
<div class="prco-ltr-right-align-helper">
<div class="bui-price-display__label prco-inline-block-maker-helper">1 night, 2 adults</div>
</div>
<div class="prco-ltr-right-align-helper">
<div class="prco-inline-block-maker-helper">
<div aria-hidden="true" class="bui-price-display__value prco-inline-block-maker-helper" data-
et-mouseenter="
">
€ 115
</div>
<span class="bui-u-sr-only">
Price
€ 115
</span>
</div>
</div>
<div class="prco-ltr-right-align-helper">
<div class="prd-taxes-and-fees-under-price prco-inline-block-maker-helper blockuid-" data-
cur-stage="1" data-excl-charges-raw="">
includes taxes and charges
</div>
</div>
</div>
<div>
<div>
<div class="roomFooter" colspan="4">
<div class="sr-cta-button-row" data-et-click="
customGoal:OQLOLOLOXaTFKGHCcZQHKKATXT:3    ">
```

Add tag name to filter content by...
Add CSS Class to filter tag by...
Add attributes to filter tag by...
Add
Scrape

Tag Name	CSS Class	Attribute	
span	bui-u-		x
div	bui-re		x
div	sr-last		x
a	bui-lin		x
span	sr-hot		x

Figure 15. Configuring page

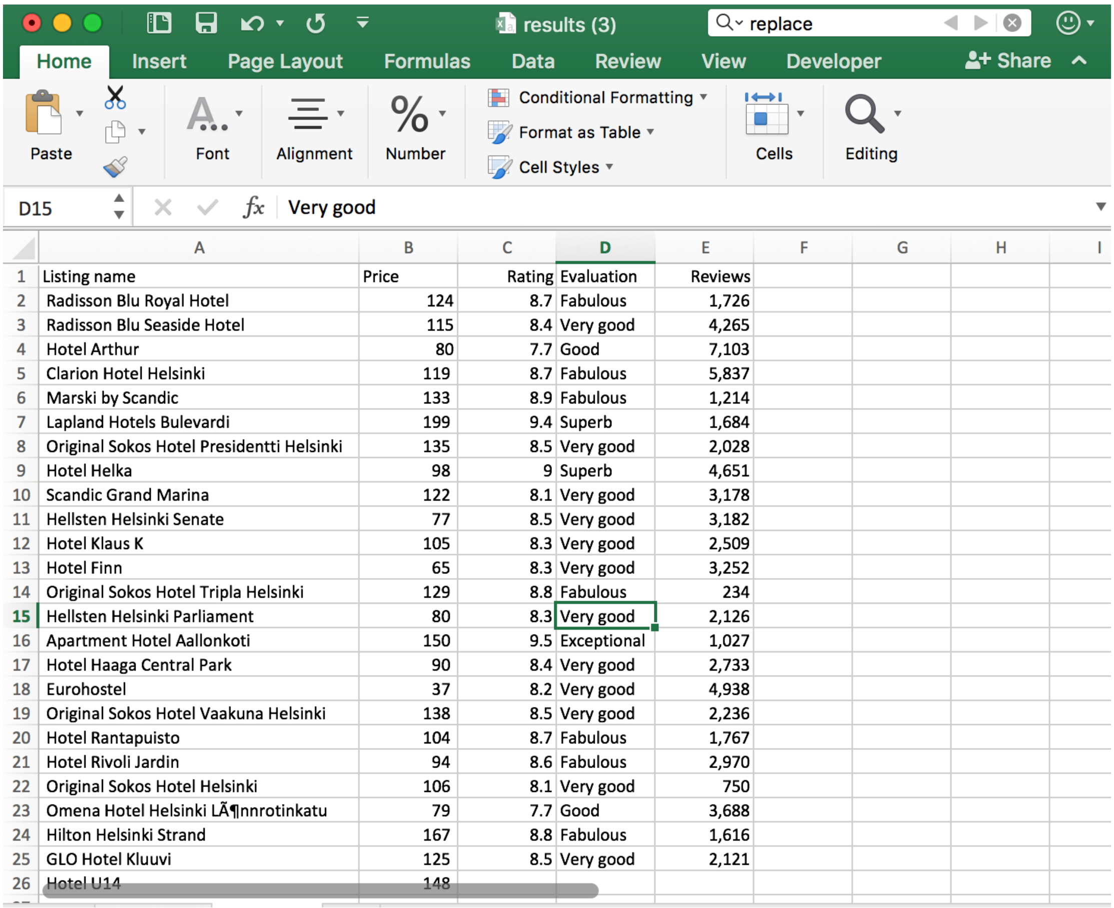
In the next stage, the scraper displays the preview window of scraping results that the user could review the information before exporting data to CSV format.



bui-review-score	bui-price-display__value	sr-hotel__name
" 8.7 Fabulous 1,726 reviews "	" € 124 "	" Radisson Blu Royal Hotel, Helsinki "
" 8.4 Very good 4,265 reviews "	" € 115 "	" Radisson Blu Seaside Hotel, Helsinki "
" 7.7 Good 7,103 reviews "	" € 80 "	" Hotel Arthur "
" 8.7 Fabulous 5,837 reviews "	" € 119 "	" Clarion Hotel Helsinki "
" 8.9 Fabulous 1,214 reviews "	" € 133 "	" Marski by Scandic "
" 9.4 Superb 1,684 reviews "	" € 199 "	" Lapland Hotels Bulevardi "
" 8.5 Very good 2,028 reviews "	" € 135 "	" Original Sokos Hotel Presidentti Helsinki "
" 9.0 Superb 4,651 reviews "	" € 98 "	" Hotel Helka "
" 8.1 Very good 3,178 reviews "	" € 122 "	" Scandic Grand Marina "
" 8.5 Very good 3,182 reviews "	" € 77 "	" Hellsten Helsinki Senate "
" 8.3 Very good 2,509 reviews "	" € 105 "	" Hotel Klaus K "

Figure 16. Preview window of scraping result

Collecting data with Abby Corner's web scraper is successful in the end. The results are exported to the XLSX file format that displays the number of accommodations listed on Booking.com based on user input. The output is illustrated in Figure 17.



	A	B	C	D	E
1	Listing name	Price	Rating	Evaluation	Reviews
2	Radisson Blu Royal Hotel	124	8.7	Fabulous	1,726
3	Radisson Blu Seaside Hotel	115	8.4	Very good	4,265
4	Hotel Arthur	80	7.7	Good	7,103
5	Clarion Hotel Helsinki	119	8.7	Fabulous	5,837
6	Marski by Scandic	133	8.9	Fabulous	1,214
7	Lapland Hotels Bulevardi	199	9.4	Superb	1,684
8	Original Sokos Hotel Presidentti Helsinki	135	8.5	Very good	2,028
9	Hotel Helka	98	9	Superb	4,651
10	Scandic Grand Marina	122	8.1	Very good	3,178
11	Hellsten Helsinki Senate	77	8.5	Very good	3,182
12	Hotel Klaus K	105	8.3	Very good	2,509
13	Hotel Finn	65	8.3	Very good	3,252
14	Original Sokos Hotel Tripla Helsinki	129	8.8	Fabulous	234
15	Hellsten Helsinki Parliament	80	8.3	Very good	2,126
16	Apartment Hotel Aallonkoti	150	9.5	Exceptional	1,027
17	Hotel Haaga Central Park	90	8.4	Very good	2,733
18	Eurohostel	37	8.2	Very good	4,938
19	Original Sokos Hotel Vaakuna Helsinki	138	8.5	Very good	2,236
20	Hotel Rantapuisto	104	8.7	Fabulous	1,767
21	Hotel Rivoli Jardin	94	8.6	Fabulous	2,970
22	Original Sokos Hotel Helsinki	106	8.1	Very good	750
23	Omena Hotel Helsinki LÃ¶nnrotinkatu	79	7.7	Good	3,688
24	Hilton Helsinki Strand	167	8.8	Fabulous	1,616
25	GLO Hotel Kluuvi	125	8.5	Very good	2,121
26	Hotel U14	148			

Figure 17. Output of the web scraper

4.6 Results and testing

By adding the web scraper to the market research process, data can be automatically run from a web page to end-users with only a few clicks. Also, data scraped from the webpage is transformed into an appropriate form and ready for further analysis steps.

In order to check whether the web scraper could decrease the time spent on the data collection step, the time consumption of each step in the market research process was recorded. Two hosts conducted the process parallelly. For comparing execution time on collecting market data step, while defining objectives, developing a market research plan, and analyzing data were completed together by two hosts, the collecting data step was conducted individually. The duration of each step is presented in Figure 18 with four colors represent step defining objectives, developing a market research plan, collecting relevant data and information, analyzing data respectively. The first row illustrates the timeline of manual market data research, the second line is conducting the time of market research process with web scraping.

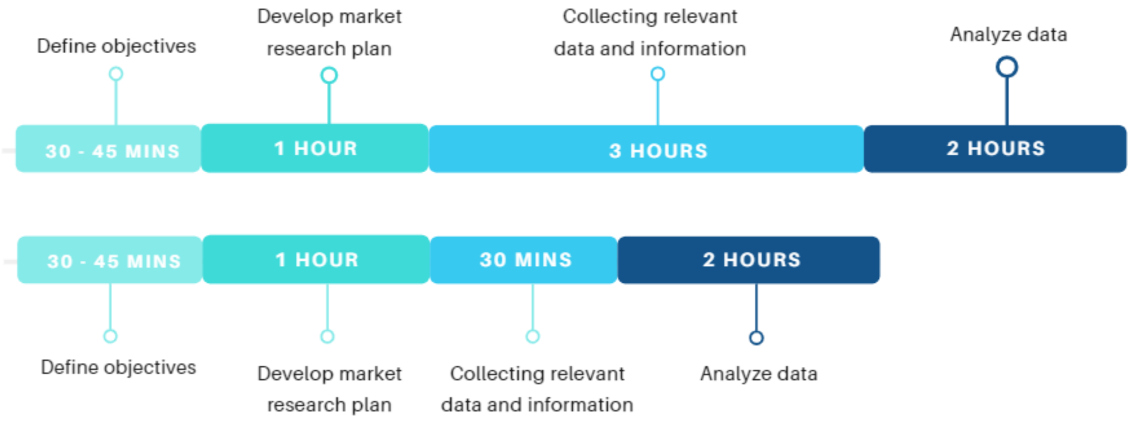


Figure 18. Conducting time of market research process

Both hosts started the data collection process at the same time, while the first host gathered relevant data for the research in 3 hours, it only took the second host 30 mins with the help of Abby Corner's web scraper. A special note was that the spent time on the data collection of the first host included time consumption for gathering data and fixing typo errors in the record, while the exported file from the automation process was in the appropriate format.

Based on time-consuming records, it could be supposed that the web scraper could save time for Airbnb hosts in the market research process, especially the data collection stage. Also, the web scraper provided the output under structured data quickly which was beneficial for the data analysis stage.

In order to clarify the benefits of applying web scraping technology regarding time productivity and quality of the collected data in the market research process of small rental businesses, face-to-face interviews with other Airbnb hosts will be organized. The interview process and analysis will be presented in the next chapter.

5 COLLECTION OF DATA

This chapter outlines the data collection process, interview setup, analysis, and the results of the study.

5.1 Interview

A face-to-face interview is a staple method used in qualitative research. As the qualitative data collection method collects in-depth answers from a focused small group of people that will manage and use the implemented system, it helps solve research questions. (Wei, Liou & Lee 2008, 607-626.)

The face-to-face interviews were carried out within selected Airbnb's hosts in Helsinki in October 2020. The goal of these interviews was to find out how the web scraper could be beneficial for their current market research processes as well as find out further suggestions to improve the scraper. The same open-ended questions were delivered to all interviewees and mainly operated in conversation format. Also, these interviews were performed in English and documented. The interview' structure is as below:

- General introduction of the web scraper
- Evaluation of market research processes before using the web scraper
- The benefits of web scraper could bring from the point of view of each participant
- Expected functions or features could be used for further improvements.

5.2 Design the structure of the interview

The data were gathered through face-to-face interviews, four Airbnb hosts agreed to participate in the interviews and provided data. The four invited interviewees were divided into two groups that are initial testing and final testing groups. The first group evaluated the first version of the web scraper, and collected data is enforced for continual improvement. The second group tested the fixed version at the final stage of the research and provided further improvement ideas.

The interview consisted of 8 questions that were divided into three important parts. The first part included general questions regarding users' feelings about current market research processes. The second part contained questions about collecting competitive data using implemented web scraper. Finally, the third part mostly focused on users' expectations of additional features and functions of web scrapping solutions. Questions in the interview mostly required detailed responses. Hence, the qualitative research was applied to collect and analyze answers regarding such open-ended questions. The full list of questions is shown in Appendix 1.

All users were given general instructions to use the web scraper as well as the scraper application one day before the interview. Each user had around one hour to examine the web scraper. Figure 19 presents the timetable of the interview process as well as the analysis.

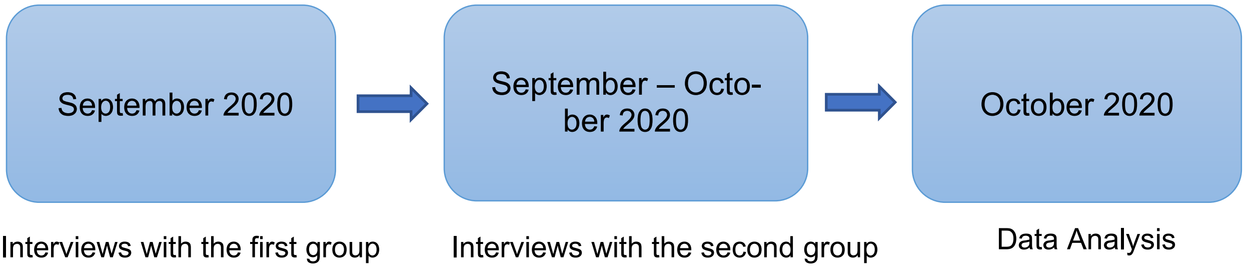

September 2020
September – October 2020
October 2020
Interviews with the first group
Interviews with the second group
Data Analysis

Figure 19. Timetable of data collection and analysis

The purpose of conducting interviews is to evaluate the effects of web scraping in the market research process of Airbnb hosts. Also, the results of the interviews help to solve the initial research questions. The goal of dividing participants into two groups is to raise requirements during the web scraper implementation and have the scraper to be tested twice.

5.3 Interview results

This sub-chapter presents an analysis of the data received from face-to-face interviews, compiles findings, and draws conclusions from interview results. Four interviewees agreed to participate in the conversations. The interview results aim to demonstrate the author's assumptions about applying web scraping to the market research process could help to reduce the execution time. The analysis of each conversation is summarized and presented separately below.

Business owners’ general data

In the first part of the interview, the conversation gathered general data about interviewees’ businesses. It was important to get an understanding of their business, the market segment, and the reliability of data received from participants.

All four participants are Airbnb hosts and have rental listings in the Helsinki region, but the rental businesses were founded in a different time.

Business owners’ current market research process

The second part of the interview collected information about the market research process that participants are currently using. As mentioned earlier in Chapter 3, there are four main steps in market research. Therefore, questions were asked separately for each step of the process

First of all, there was a question about their target objectives that need to be researched for the business. Participants defined four common objectives for researching:

- real-time pricing
- listings’ location
- listings’ capacity
- customers’ reviews
- listing’ rating.

All participants mentioned pricing as their top research objective. There was one participant stated that the pricing is her priority of researching:

> *The price per night depends on many factors like season, day of week and it affects the monthly revenue.*

The reason for this general concern about the pricing system is that it could help Airbnb hosts to optimize their rental listings. Although the Airbnb platform provides a smart pricing tool for hosts, the tool aims to help increase the occupancy rate by suggesting lower prices than the hosts’ expectation (Gollapudi 2020). Hence, Airbnb hosts prefer to research and compare competitive prices against prices of listings in the same market segment.

Secondly, there were questions regarding the development of the research plan as the second step of market research. According to answers, participants usually access the travel aggregator websites from two to four times a week that depends on seasons. There were four websites that four hosts focused on: Booking.com, Kayak.com, Momondo.com, and Airbnb.com. In terms of sampling sizes, participants answered that they usually collected data from 50 to 100 similar listings in the same location. Also, while answering questions about the research plan, interviewees noted that they recorded data and information in Excel:

> *I would like to use an Excel spreadsheet to store collected data as it's easy to use and I can also analyze data in it.*

The market research plan included choosing data storage options. Data extracted are usually saved to flat files such as XLSX, CSV, XML, etc.

Thirdly, there was a section with questions about collecting data and information processes. Questions focused on execution time and data quality. While three participants gathered data manually by searching listings, copying, and pasting data to Excel spreadsheets respectively, the fourth interviewee stated that she collected information with ParseHub which is a web scraping service. Participants specified:

> *As I am running three Airbnb listings at the same time and they are in different locations, I find ParseHub useful for my business.*

Moreover, participants estimated that the execution time was in the range of 3-5 hours for one person to handle the task. The process collected information on listings' names, locations, capacities, price per night, and customer ratings. Also, because of the large amount of manual data entry, there were typo errors in the results, such as punctuation mistakes, spelling, or grammar.

Browsing through many web pages to collect information is overwhelming. In fact, the case study had similar problems with the manual data collection process as interviewees. Therefore, the research was conducted with the goal to minimize time spent on gathering desired data as well as transform collected data to standard database format.

Abby Corner's web scraper usage

Participants were asked to use Abby Corner's web scraper in their data collection process one day before the interview. After that, there were questions about their opinions on the web scraper and whether it could be beneficial for their market research process. At first, the majority of participants mentioned that it was easy to paste the link, and the preview screen for the website's content went well. However, some of the opinions were that it was difficult for them to understand website structure to select the correct HTML and CSS components for executing the scraper. For example, one of the interviewees said:

> *The website structure in the preview screen has been confusing to me.*

Also, another interviewee stated:

> *I could not find the exact place of data I want to extract.*

The reason for these opinions is that the Abby Corner's web scraper requires to have basic knowledge of HTML and CSS to control it. The author herself had a previous technical background so that she could monitor and use the solution in her business model.

Based on the received responses, the author had a quick explanation of HTML and CSS components. After that, together with the author, participants run the web scraper to collect their desired information and examined how it could work. There were questions to follow up with participants regarding their feelings about the solution as well as its features. In terms of the execution time, the majority was happy about the speed of collecting the required information on the web scraper. They evaluated that the web scraper could minimize a three-hour-task to only 2-3 minutes. One participant also said:

> *The scraper runs pretty fast. It is definitely worth learning HTML.*

Also, there were two interviewees mentioned that the exported data files are suitable for them to do further analysis. For instance, one participant answered:

> *The output is under XLSX format is suitable for me to explore data with visualizations in Excel.*

Based on the participants' answers, the interview sections help the research verify the author's assumption in Chapter 3. By applying web scraping to the data collection process, it improves the execution time that decreases the spent time on this step from three hours to three minutes. In addition, the participants declared themselves satisfied with the exported data format.

The last question of the interview section was optional. It aimed to let interviewees give some suggestions for applying the web scraper to the market research process. Most of the ideas were mainly about Abby Corner's features. For other Airbnb hosts could also apply Abby Corner's web scraper to their model, add brief instruction of selecting HTML components and CSS elements, the possibility to select data to scrape by directly choosing on the website interface. Also, participants would like to send the output to the email. That is because Airbnb hosts sometimes want to send the result to their team or partners, and email is a popular tool to exchange information. The participants suggested that:

> *Also, the email button to export results could be useful.*

To sum up, key findings from data collection through interview sections and data analysis part could recommend for further improvement of the web scraper implementation as well as the research. As a suggestion for further development is also a part of the constructive research framework, they are presented in sup-chapter 5.4.

6 CONCLUSIONS

After summarizing theoretical research and case study’s results, this chapter concludes answers to the research questions. The demonstration, applicability scope of the research is also discussed as parts of the constructive research approach.

6.1 Answering research questions

The goal of the study was to explore how the web scraping technologies can support the market research process for the rental property business. In order to achieve the initial purpose, this section answers formulated research questions in Chapter 1. The results are presented below, starting from sub-questions to the main research question:

How do existing market research steps look like?
As stated in Chapter 3, the traditional market research mainly includes four steps: defining objectives and problems; collecting data manually; analyzing data; visualization, and communicating with data. As the amount of data on the internet is unimaginable, data collection is the most time-consuming step. In the interview section, three participants mentioned the challenges of identifying and collecting the relevant data. Moreover, inaccuracy in the manual data collection is another issue that makes the existing market research less efficient.

How does a web scraper service help to collect competitive market data?

According to the web scraper implementation result, the scraper helped to decrease the time spent on the data collection process from 3 hours to 30 minutes for 98 listings’ results. The web scraper service helps the researcher to collect a huge amount of data from multiple websites. One of the biggest advantages of using a web scraper for market research is speed. It replaces repetitive tasks and increases the efficiency of the process. By implementing the web scraper, users only need to select the relevant information and send the order to the service. In addition, employing a web scraper in market research will increase the accuracy of the collected data.

What are the stages of building a simple web scraper as a service?

While purchase web scraping service could be overpriced for small businesses, creating the web scraper service with Python is simple and easy to understand. The implementation part of the thesis presented required frameworks to build one web scraper service from scratch.

The web scraper service is the combination of two parts that are scraping and web application. Creating a simple web scraper service is easy to adopt, follows the common three stages that are detection, developing the scraping site, and building the simple web application.

The detection includes identifying and detecting the web page that the user wants to extract data, selection of relevant information. In order to collect the data in an automatic way, the development of the scraping site could be done with the help of BeautifulSoup or Scrapy. The final stage is to integrate the scraping technology with the web service, and it could be quickly completed by applying Flask.

To conclude, the main research question is answered below:

How can the web scraping technologies support the market research process for rental property business?
Web scraping technologies enable innovative methods of collecting data concisely that support the market research process. Also, web scraping provides service at a low cost, and it is easy for small rental property businesses to implement. In this case study, the web scraper helps Airbnb hosts generate information on rental listings from Booking.com, customer rating, price, and promotions. Employing web scraping helps users save time on the data collection process and focus on analyzing and communicating with data for decision-making.

With the web scraper, Airbnb hosts can collect real-time pricing intelligence data and quickly react to the situation for the competition. Thanks to suitable formats of exported data, business owners can easily analyze the dataset to improve marketing campaigns, understand customers' behavior as well as monitor vacancy rates, which are really important in the COVID-19 pandemic

6.2 Demonstration of web scraping solution working

As a demonstration of the web scraping solution working, especially the supports of the web scraper to the market research process, interviews were carried out within Airbnb hosts in which participants examined the solution in their market research process. The main goal of the study to evaluate whether a web scraper could help Airbnb owners to collect competitive market data in an efficient way. In general, feedback received from interviews was positive. Also, the final outcome of the case study was successful in the end that it decreased the time spent on the gathering data process. Together with the interview's results and the outcome of the case study, it can be concluded that the research has reached its initial objectives.

6.3 Examination of applicability scope

In order to examine the applicability scope of the study, the results of the interviews were taken into consideration. The initial applicability scope of the research was to demonstrate whether the web scraping could support the market research process of small rental businesses. From the interviewee's answers, the web scraping technology could be beneficial for these businesses regarding time productivity and efficiency. However, there were challenges for users that do not have a previous technical background to create and maintain a web scraper by themselves. The reason for this fact was that the case study had the financial limitation, and building a web scraper from scratch could fit the business' current requirement.

7 SUMMARY

In a nutshell, the application of web scraping is the data extraction in only a few minutes to gather entire pieces of information and convert unstructured data into an organized format. This thesis discovered the benefits of having web scraping in the market research, especially the time productivity and efficiency in collecting relevant information stage. Thanks to web scraping, businesses can quickly access and collect competitive market data as well as organize and manage data efficiently.

The objective of this thesis was to find out how web scraping could support a small rental business' market research process. As the case study had a budget limitation, this research developed a web scraper with a programming language, which is Python. This implementation required previous technical background for creating and maintaining the solution. Hence, other companies could look at web scraping services that offer available functionalities with different subscription fee options.

Along with the theoretical background and case study giving practical results, the research accomplished the initial goals and was able to answer the main research question. The web scraping helped Airbnb hosts automate collecting data steps with high data collection quality and sped up the case study's market research process. Based on these key benefits, it leads to the conclusion that companies could consider web scraping as a powerful technology to scrape competitive information in order to make timely and suitable decisions. Lastly, the final sections present the reliability and validity of the research as well as suggestions for further improvements in the future.

7.1 Reliability and validity

Considering the data collection of the research, the interviews were handled with Airbnb hosts who have operated the business from two to four years. They have worked closely with the market research process and had experience in collecting competitive data. Hence, their outputs and assessment are reliable, useful, and practical. Also, this could help increase the reliability of this research.

7.2 Suggestions for further development

In general, the results of the case study and testing sections with interviewees have valid data. However, because of the COVID-19 situation, the research was only able to invite a limited number of interviewees. Moreover, the research was conducted in small rental businesses in the Helsinki region. Hence, the research results may not be applicable to other business concepts and other regions.

Some of the potential topic for further development to continue after this thesis could be:

- Add brief HTML and CSS instructions to select correct components and classes in the web scraper.
- Design different prototypes for the user to monitor the web scraper without knowledge of HTML and CSS.
- Add the option to send results to the email address.
- Interview with bigger numbers of participants from other region’s markets.

LIST OF REFERENCES

Written References

Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., Fdez-Riverola, P. 2013. Web scraping technologies in an API world. The United Kingdom: Oxford University Press.

Hajba, G. 2018. Website Scraping with Python: Using BeautifulSoup and Scrapy. Hungary: Apress.

Lindholm, A. 2008. A constructive study on creating core business relevant CREM strategy and performance measures. Emerald Group Publising 2008, 343-358.

Pasian, B. 2015. Designs, Methods and Practices for Research of Project Management. Gower Publishing company. United States of America: Gower Publishing company.

Sirisuriya, S. 2015. A Comparative Study on Web Scraping. Proceedings of 8th International Research Conference, KDU, 135.

Wei, C.C., Liou, T.S., Lee, K.L. 2008. An ERP performance measurement framework using a fuzzy integral approach. Journal of Manufacturing Technology Management 6/2008, 607-626.

Oyegoke A.S. 2011. The constructive research approach in project management research. International Journal of Managing Projects in Business. Leeds: Leeds Beckett University, 573-595.

Electronic Sources

Bhattacharjee, D., Seeley, J., & Seitzman, N. 2017. Advanced analytics in hospitality. McKinsey Digital [accessed 13 September 2020]. Available at: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/advanced-analytics-in-hospitality

Booking. 2020. About Booking [accessed 19 September 2020]. Available at: https://www.booking.com/content/about.en-gb.html?aid=356980;label=gog235jc-1DCBQoggJCBWFib3V0SDNYA2hIiAEBmAEJuAEXyAEM2AED6AEBiAIBqAIDuA

L9t9f8BcACAdICJDhmZWI0NDZhLTUzZWMtNDZjOC1hMDE2LWE5MDc3YjQzYzNlOdgCBOACAQ;sid=1c1ef993ff25a17dd1c7ce17137d1c96;keep_landi

DataOx. 2020. A Comprehensive Overview of Web Scraping Legality: Frequent Issues, Major Laws, Notable Cases. Medium [accessed 19 September 2020]. Available at: https://dataox.medium.com/dataox-quick-overview-of-the-best-data-scraping-tools-in-2020-a-devils-dozen-everyone-should-f7016fa348a0

Entrepreneur Europe. Market Research [accessed at 30 October 2020]. Available at: https://www.entrepreneur.com/encyclopedia/market-research

Emerson, K. 2019. An Introduction to Web Scraping for Research. University of Wisconsin-Madison Research Data Services [accessed 11 August 2020]. Available at: https://researchdata.wisc.edu/news/an-introduction-to-web-scraping-for-research/

Fatenaite, G. 2020. Oxylabs. XML Processing and Web Scraping With lxml. [accessed at 13 September 2020]. Available at: https://oxylabs.io/blog/lxml-python-library

Gollapudi, A. 2020. Airbnb Pricing Recommender. Towards Data Science [accessed at 10 November 2020]. Available at: https://towardsdatascience.com/Airbnb-pricing-recommender-19225d0f5d1

IBM. 2014. IBM. IBM Delivers New Big Data Capabilities on IBM Cloud Marketplace. [accessed at 7 September 2020]. Available at: https://www-03.ibm.com/press/us/en/pressrelease/44188.wss

Koshy, J. 2017. Should Data Scientists Learn Web Scraping? Prompt Cloud [accessed 13 September 2020]. Available at: https://www.promptcloud.com/blog/should-data-scientists-learn-web-scraping/

Koshy, J. 2020. Best Programming Language For Web Scraping. Prompt Cloud [accessed 15 August 2020]. Available at: https://www.promptcloud.com/blog/best-programming-language-for-web-scraping/

Kraczkowsky, C. 2019. Getting started with the Flask web framework. Medium [accessed 5 September 2020]. Available at:

https://medium.com/@ckraczkowsky/getting-started-with-the-flask-web-framework-a7c2862dfba8

Market Research Guy. 2020. The Market Research Process: 6 Steps to Success. My Market Research Methods [accessed 14 October 2020]. Available at: https://www.mymarketresearchmethods.com/the-market-research-process-6-steps-to-success/

McKinsey. 2017. Capturing value from your customer data [accessed 22 August 2020]. Available at: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/capturing-value-from-your-customer-data#

McKinsey. 2020. Hospitality and Covid 19: How long until ‘no vacancy’ for US hotels? [accessed at 28 September 2020]. Available at: https://www.mckinsey.com/industries/travel-logistics-and-transport-infrastructure/our-insights/hospitality-and-covid-19-how-long-until-no-vacancy-for-us-hotels#

Mottet, J. 2018. Web Scraping Using Python Selenium. Medium [accessed 22 August 2020]. Available at: https://medium.com/@1995818liu/wed-scraping-using-python-selenium-c58e4036d742

Narizhnykh, D. 2018. Is web scraping legal or not? Medium [accessed 3 October 2020]. Available at: https://medium.com/dataflow-kit/is-web-scraping-legal-or-not-f6c26074584

Othman, S. 2011. The comparison of qualitative research and quantitative research [accessed 26 August 2020]. Available at: http://shayaaresearch.blogspot.com/2011/04/qualitative-vs-quantitative-research-v.html

Samim. 2020. The formulation of abductive reasoning approach [accessed at 19 September 2020]. Available at: https://samim.io/p/2019-12-15-abductive-reasoning/

Scrapy. 2020. Scrapy at a glance. Github [accessed 1 June 2020]. Available at: https://docs.scrapy.org/en/latest/intro/overview.html

Scrape Hero. 2018. What is web scraping – Part 1 – Beginner’s guide [accessed 22 September 2020]. Available at: https://www.scrapehero.com/a-beginners-guide-to-web-scraping-part-1-the-basics/

Tutorials Point. 2020. Python Web Scraping Tutorial [accessed 22 September 2020]. Available at: https://www.tutorialspoint.com/python_web_scraping/python_web_scraping_introduction.htm

Turnitin. 2020. Turnitin [accessed 14 October 2020]. Available at: https://www.turnitin.com/robot/crawlerinfo.html

APPENDICES

Interview Question

1. What is your name?
2. How long have you been working as Airbnb hosts?
3. Describe briefly how does the current collecting market data process? How long does it take?
4. How many times do you visit websites such as booking.com, kayak.com, momondo.com a week to review the rental market?
5. Which data sources were extracted?
6. How do you feel about the web scraper? Do you have any difficulties in using the solution?
7. Do you think the web scraper could support your market research process? Are you willing to use it in your weekly market research?
8. Do you have any suggestions for improvement?