

tripadvisor data scraping and analytics project

SOI-2020-2020-0007

Date of Submission: 02-FEB-2021

Submitted By:

SOI-2020-0005

|  |  |
| --- | --- |
| 16044144 | ALAN LAU TAI LIN |
| 18047993 | LIEW CHIA YEE |
| 18033629 | YEE LAI MOI |
|  |  |

# ACKNOWLEDGEMENT

The team would like to thank our supervisor, Dr Tan Lee Kiat, for his supervision throughout the semester. He has given us the best guidance and careful monitoring throughout the project, constantly encouraging the team to do better from week to week to show a better result.

We would also like to thank the WebHarvy company for the support by providing an educational license for each of us in the team to access the WebHarvy scraping tools. The educational license allows the team to scrap the data needed for the project without any limitations.

Last but not least, the team would like to thank Tableau company for providing an educational license for the team to access the Tableau software. Without the license, the team would not be able to do analysis on the project data.

**TABLE OF CONTENTS**

[ACKNOWLEDGEMENTS 2](#_30j0zll)

[ABSTRACT 4](#_3znysh7)

[1](#_2et92p0) Introduction 5

[2](#_tyjcwt) Project Specification and Plan 6

[2.1](#_3dy6vkm) Project Overview 7

[2.2](#_1t3h5sf) Functional Requirements 8

[2.3](#_4d34og8) Project Plan 8

[3](#_2s8eyo1) Business Analysis 9

[3.1](#_17dp8vu) Business Issues 9

[3.2](#_3rdcrjn) Market Analysis 9

[3.3](#_26in1rg) Business Solutions 9

[4](#_lnxbz9) System Design and Implementation 10

[4.1](#_35nkun2) System Architecture 10

[4.2](#_1ksv4uv) Detailed System Design 10

[5](#_44sinio) System Testing 11

[6](#_z337ya) User and Technical Documentations 12

[6.1](#_3j2qqm3) User Documentation/Guide/Manual 12

[6.2](#_1y810tw) Technical Documentation (Installation guide/Manual) 17

[7](#_4i7ojhp) Conclusions 18

[References](#_2xcytpi) 19

[Appendices](#_1ci93xb) 20

[Project Poster](#_3whwml4) 21

# ABSTRACT

The Project is about Data Scraping & Analytics where we are required to learn how to scrape data online which include exploring of the Data Scraping tools online, exploring of the scraped data using visualization tools and then learn data analytics and apply on the scraped data.

The project is implemented by figuring out which data scraping tools are most suitable for scraping data on TripAdvisor website. With that, we are able to gain knowledge of how the tools work and how data is scraped. Also, with the scraped data from TripAdvisor, we are required to use visualization tools to learn data analytics and apply on the data we retrieved.

This project has given us an opportunity to learn how to scrape data on TripAdvisor using data scraping tools online and given us a chance to work on data visualization and analytics which enables us to learn new skills and gain new knowledge.

# Introduction

Our Data Scraping & Analytics on TripAdvisor Data Project Consist of Two Parts:

We are required to explore free data scraping tools online and use it to scrape data on Singapore Hotel, Restaurant and Attraction. Using the scraped data, we are required to carry out data analytics to derive each domain guest profile and build a model to find out which are the key factors (based on the rating data) that different domain guest profiles are concerned about their choice of selection.

We are to adopt commonly used data mining methodology and include any data exploration and data cleansing steps.We need to produce meaningful insights from the data before the analytics using Tableau or any visualization tool as part of data understanding. Minimally, we must demonstrate data visualization knowledge and skills and exploration of the analytics tools.

# Project Specification and Plan

Some deliverables that our supervisor and the team members agreed upon before the start of the project are:

* Each of the team members are to explore free data scraping tools online and decide on one tool to be used for data scraping.
* Team members are required to demonstrate good knowledge of the data scraping tool chosen.
* Each of the team members are to pick on one domain each instead of working on the same domain (Hotel, Restaurant, Attraction).
* Team members is to adopt commonly used data mining methodology and include any data exploration and data cleansing steps
* Team members are to produce meaningful insights from the data before the analytics using Tableau or any visualization tool as part of data understanding.
* Team members must demonstrate data visualization knowledge and skills and exploration of the analytics tool (dashboard, a predictive model and Tableau Story).

## Project Overview

We will learn how to scrape data online, explore the data using visualization tools and then learn data analytics and apply on the scraped data. As these 3 skills are in demand in the job market, we are able to gain new knowledge and learn new skills by doing this project.

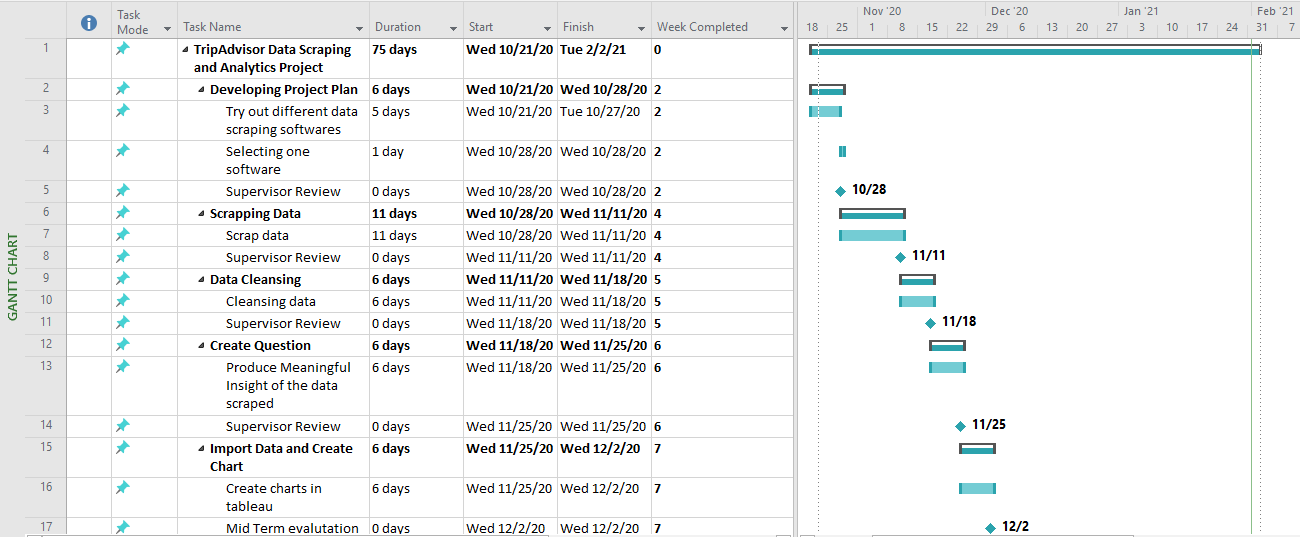
Assuming by the end of the project, we will be able to gain the knowledge and skills on:

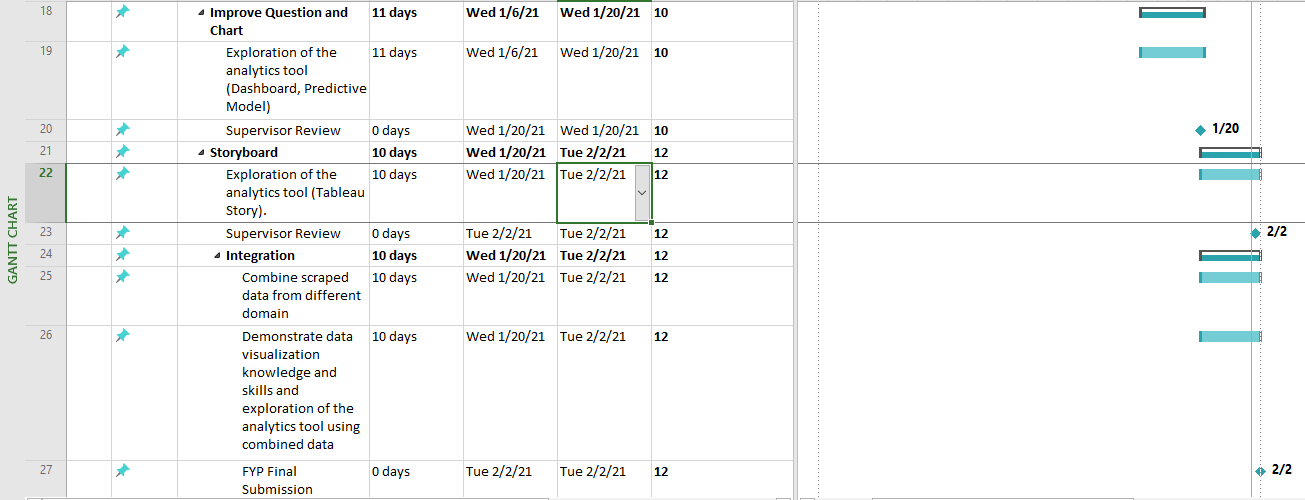
* How to use the data scraping tool we have chosen works
* How to apply data cleansing skills to remove unnecessary data that are not required for data visualisation.
* How to produce meaningful insight from the data we retrieved.
* How we apply data visualisation knowledge and skills on the data we have.
* How to create a predictive model, dashboard and story.
* How the interpretation will be applied.

## Functional Requirements

Scraping tools software able to scrap the needed data to be imported into the Tableau software for analysis.

## Project Plan



**

# Business Analysis

We study the reviews on Tripadvisor website. Each of us took one domain mainly Hotels, Restaurants and Attractions. Using Tableau to analyse the reviews, Tripadvisor’s reviews have been decreasing since Covid-19 restrictions started. The Singapore Tourism Board is offering a $100 SingapoRediscovers vouchers for all Singaporean age 18 and above to support local tourism businesses.

## Business Issues

With Covid-19 border restrictions and borders closure, many foreigners are not able to travel to Singapore. Also, with circuit breakers, many local tourism businesses are not able to operate. The number of visitors dropped which resulted in local tourism businesses making a huge loss, some even had to cease operations.

## Market Analysis

With only locals to support the businesses of the tourism sector in Singapore, TripAdvisor reviews will be able to give an insight to promote the business to the locals.

## Business Solutions

With the help of the data collected and charts in reviews of the different domains, better insights could be given to support Tripadvisor in promoting the business of the tourism sector in Singapore by recommending different packages to offer to the locals.

# System Design and Implementation

## System Architecture

Scraped data by the WebHarvy software will be exported to an excel worksheets which Tableau software allows the importing of excel worksheets for chart creating and analysis. The team has imported each of the domain data that we have scraped from the website into Tableau software to create charts for analysis to create a solution.

## Detailed System Design

Under hotel domains, there are 2 excel worksheets which consist of the hotel's detail and user reviews. With the hotel field name that the username has reviewed, it will be linked to the Hotel’s detail to match with the hotel name to create a relation between the 2 worksheets.

This will allow the Tableau software to calculate the numbers of reviews, average rating by the user for each hotel to be shown on the charts that have been created.

In the attraction domain, there are 2 excel worksheets which have the attraction’s information and user individual reviews. With the same attraction field name in both excel sheets, it will be a link which allows the Tableau software to perform calculations in the bar graph and pie charts that I created. Also, with the address and postal code of each attraction, it allows Tableau to create a world map with the longitude and latitude.

For the restaurant domain, there is only 1 excel worksheets as the data are all combined in one sheet which contains 14 types of data (Restaurant, Cuisine, Restaurant Rating, Total Review, Address, Username, User Location, User Rating, Reviewed Date, Date of Visit, Review Title, User Review, Cost, PostalCode). With these data, it will allow us to demonstrate data visualisation knowledge, skills and allow us to create predictive models, dashboard and story on Tableau.

# System Testing

In the testing phase, the team has tested 4 scraping software, ScrapStorm, WebHarvy, ParseHub and OctoParse, to see which software will be suitable to scrap the data from the TripAdvisor website.

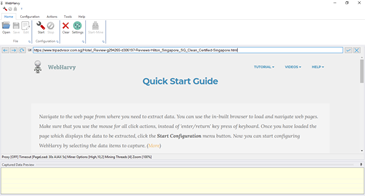
WebHarvy was chosen due to the other 3 software that have many limitations on scraping the data compared to WebHarvy, which has only 1 limitation which is the maximum number of pages to be scrapped is 2 pages each time. WebHarvy also allows us to select the needed field from the website rather than scanning the entire page with fields which we do not need for the project.

With the limitation of maximum 2 pages of scraped data, the teams have decided to approach Webharvy company for the support of providing us the license code to use the full version of the software which allows the team to scrape data with unlimited number of pages.

# User and Technical Documentations

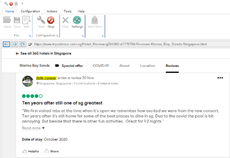
## User Documentation/Guide/Manual

***1. Enter the URL of TripAdvisor review that you want to scrap.***

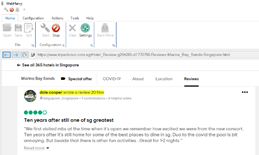
**

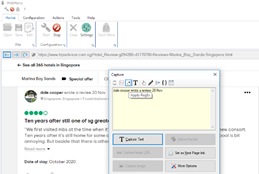
***2.******Scroll down to the first review to scrap the review***

***a.*** *Click on the username of the username, then “Capture Text.” To capture the username.*

**

***b.*** *Click on the date of review by the user.*

**

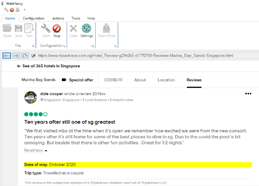
**

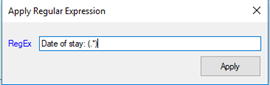
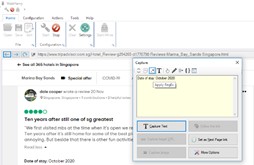
*Applying regular expressions to capture only the date.*

**

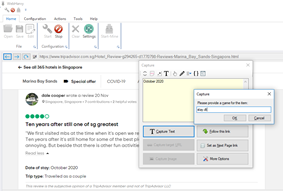
*“Capture text” to capture the Date of review by the user.*

***c.*** *Click on the Date of Stay.*

**

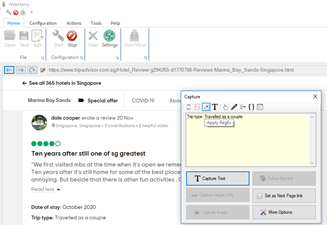
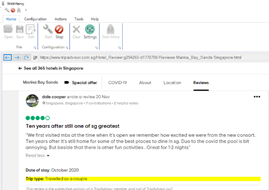
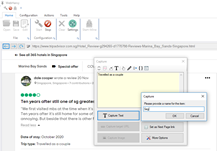
**

*Applying regular expressions to capture only the date.*

**

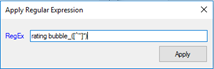
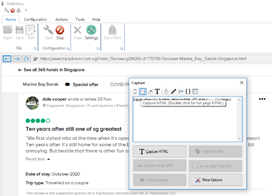
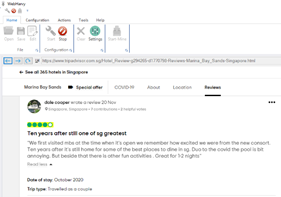
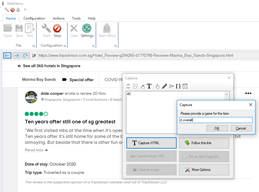
*“Capture text” to capture the Date of Stay by the user.*

***d.*** *Click on the Travel Type*

*Applying regular expressions to capture only the date.*

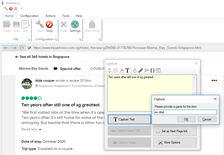
*“Capture text” to capture the Travel type of the user*

***e.*** *Click on the rating bubble*

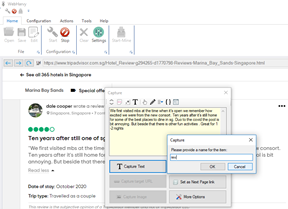
*Applying regular expressions to convert the bubble to numbers.*

*“Capture Text” to capture the rating by the user. (Cleansing to be done in the excel worksheet to convert the rating from e.g., 40 to 4)*

***f.*** *Click on the review title to capture the title of the review.*

**

***g.*** *Click on the review paragraph to capture the review by the user.*



## Technical Documentation (Installation guide/Manual)

1) Go to<https://www.webharvy.com/download.html> to download the Webharvy scraping software.

2) Install the software in your desktop/laptop

3) Email support@webharvy.com for educational license to get the full version of the software. (Unlimited pages of scraping)

# Conclusions

At the start of the project, we were required to explore free online data scraping tool to scrape data from TripAdvisor, the software provided online are free but it’s very limited as most of it requires payment to access the full version of the software and some of the software we tried can’t capture specific data from TripAdvisor. Fortunately, we’ve found WebHarvy and chosen it as our data scraping tool as WebHarvy provides educational licenses for students which we don’t have to pay for accessing their full software and the software allows us to capture the required data for the project.

With the finishing of data scraping, we are required to use the data to produce meaningful insights, demonstrate data visualisation knowledge, skills and create predictive models, dashboard and story on Tableau. We had a difficult time thinking of ideas to produce meaningful insights as we lack creativity, this project also requires lots of effort to design and thinking of creative charts using those data scrapped. Also, solutions on how we demonstrate data visualisation knowledge and skills are very limited. This enables us to look for resources online, within the team or even figuring it out individually. With that, we are able to perform multiple visualisation skills on Tableau by connecting charts to dashboard and to Tableau Story.

After working on this project, we find that most of us are able to apply some of the data analytics skills using Tableau and gained some knowledge of data scraping using the scraping tool we have chosen (WebHarvy) and in the future if there’s any need to perform any data scraping or data analytics skills, this knowledge and skills gained from our Final Year Project would be very helpful.

* References

https://www.webharvy.com/download.html

# Appendices

# Project Poster

**