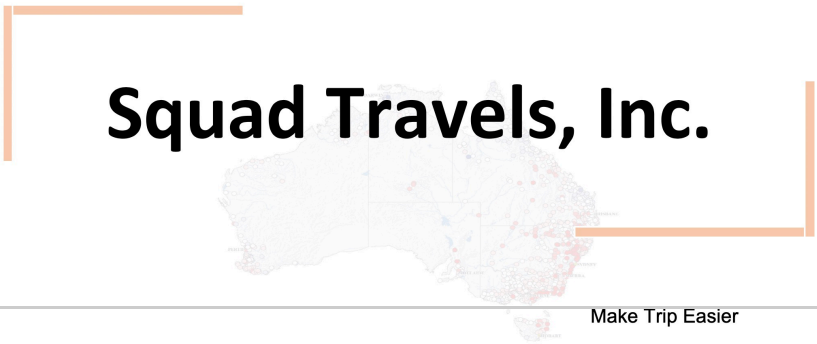
Squad Travels, Inc.

STARS 2



0. Group Members

Group name	Andrew ID	Email
Liyuan Gu	liyuang	liyuang@andrew.cmu.edu
Yuou Lei	yuoul	yuoul@andrew.cmu.edu
Xinrui Zheng	xinruiz	xinruiz@andrew.cmu.edu
Shanyue Wan	shanyuew	shanyuew@andrew.cmu.edu
Rhea-Luz Valbuena	rvalbuen	rvalbuen@andrew.cmu.edu

1. Criteria Examples

1.1. Overall Complexity and Scope of Project

We integrate multiple datasources, including geography data, social media comments, weather website, hotel data in our project. Based on our project, we provide a clear and comprehensive traveling recommendation for Chinese travellers.

1.2. Source code comments

Each file has been documented with comments.

1.3. What needed to be run,and in what order

Please follow the instructions below.

1.4. Descriptive statistics, tabular visualization, cross tabular visualization, and graphical visualization to achieve the scope of your project

Multiple visualization: weather, hotel, airline, map route recommendation and word cloud. We combined geography data with other scraped data to provide a personalized traveling plan. Also, we have completed

data analysis on comments frequency and generate respective word clouds, cost and traveling distance analysis.

1.5 Python Language Basics

Different python basics have been used in the files: modulation, main, etc.

1.6 Built-in Data Structures, Functions, and Files

Data Structures like, list, dict have been used in the files:

list --main.py

```
place_list.append(airline)
hotel, hotel_price = get_suitable_hotel(input_des, input_mode)
place_list.append(hotel)
tourism_list = get_five_top_tourism_attraction(input_des)
place_list.extend(tourism_list)
place_list.append(hotel)
```

dict --ctrip_comment.py

```
city_dict = {}
city_list = father_page.find_all("div", {"class":"list_mod1"})
for city in city_list:
    city_name=city.find("dl").find("dt").text
    city_detail = {}
    place_list=[]
    url_list=[]
    for a in city.find("dd").find_all("a"):
        if(not a.attrs["href"].startswith("http") and ("sight" in a.attrs["href"]
\
                                                        and (not "sightlist" in a.a
ttrs["href"]))):
            place_list.append(a.text)
            url_list.append(ctrip_url+a.attrs["href"])
    city_detail["place"]=place_list
    city_detail["url"]=url_list
    city_dict[city_name]=city_detail
```

1.7 Data Loading, Web Scraping, Storage, and File Formats

During the scrape part, we use selelium, beautiful modules to scrape data from different websites. Aditionally, in the ctrip scrape part, we simulate the browser behavior of the views and simulate the clicking behavior to gain more comment data.

1.8 NumPy

Numpy is used combined with pandas to do data analysis. Following is one example:

nlp_analysis.py

```
def get_five_top_tourism_attraction(city):
    city = translate(city)
    city_df = ctrip[ctrip["city"]==city]
    city_df = city_df.groupby("place")
    city_df = city_df.aggregate(np.mean)
    city_df = city_df.sort_values(by="rating", ascending=False)
    return list(city_df.index)
```

1.9 Pandas

Either in software part or in the scrape part, we use dataframe to process and store data. Also, we use several dataframe advanced techniques, like apply to make our analysis more efficient.

```
def add_range_hotel(price,hotel_down,hotel_up):
    price = int(price)
    if(price<=(int)(hotel_up) and price>=(int)(hotel_down)):
        return 1
    else:
        return 0

hotel_data["price_within"] = hotel_data["price"].apply(lambda x:add_range_hotel(x,
hotel_down,hotel_up))
```

1.10 Plotting & Visualization: Join, Combine, Reshape

We combine data from weather, geography, hotel and tourism attractions to give integrated route plans. Map could be viewed in map folder.

2. Software

2.1. Abstract

Squad Travels, Inc is a software to provide comprehensive personalized suggestions for travellers from China. We scraped data including, geography data, social media comments, weather website, hotel data. Our suggestions are based on the integration of those information. Our version is to be the premier choice of Chinese travelers who want to travel to Australia for travel information needs.

Following is our core functions:

- Route Guidance: Map and Trip Plans
- Airline Information Recommend
- Hotel Information Recommend
- Weather Information Display
- Calculate Distance and Total Costs
- Word Cloud Generation

2.2. Dependence

Library	Version
pandas	0.25.1
numpy	1.16.4
matplotlib	3.1.0
folium	0.10.0
wordcloud	1.5.0
prettytable	0.7.2
jieba	0.39

2.3. File Structure

> main.py
>
> graph_generator.py
>
> map_generator.py
>
> filter*suitable*service.py
>
> weather_analysis.py
>
> NLP
>
> > nlp_analytics.py
>
> dataset
>
> > AdelaideMetroStops_GDA2020.json
> >
> > AdelaideMetroStops_GDA94.json

> > airline_data.xlsx
>
> > ctrip*cleaned*data.csv
>
> > hotel_data.xlsx
>
> > weather_data.xlsx

2.4. Library Install

Use command line: `pip install -r requirements.txt` to install all the python modules that you need.

2.5. Running

Running the main modle.

Input Mode: `Economy` / `Luxury`

Input City: `Adelaide`

Input Beginning Time: `2018-01-01`

Input Ending Time: `2018-04-01`

Input Down-Bound Hotel Price: `50`

Input Up-Bound Hotel Price: `500`

Input Down-Bound Flight Price: `500`

Input Up-Bound Flight Price: `1000`

2.6. Live Deomo

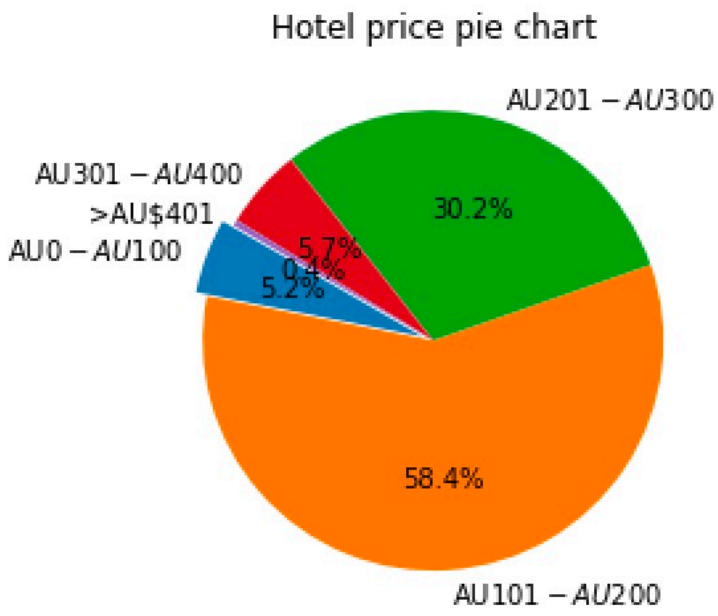
a). Get Input

```
Input your destination:

Adelaide
Input your travel time down bound: (Format yyyy-mm-dd)

2015-01-01
Input your travel time up bound: (Format yyyy-mm-dd)

2015-03-01
Input your hotel money down bound:

50
Input your hotel money up bound:

500
Input your flight money down bound:

500
Input your flight money up bound:

1000
```

At the beginning, to get personalized services, customers should input their preferences. Based on the preference, our program can then give the services, traveling route they want.

b). Hotel Price Pie Chart

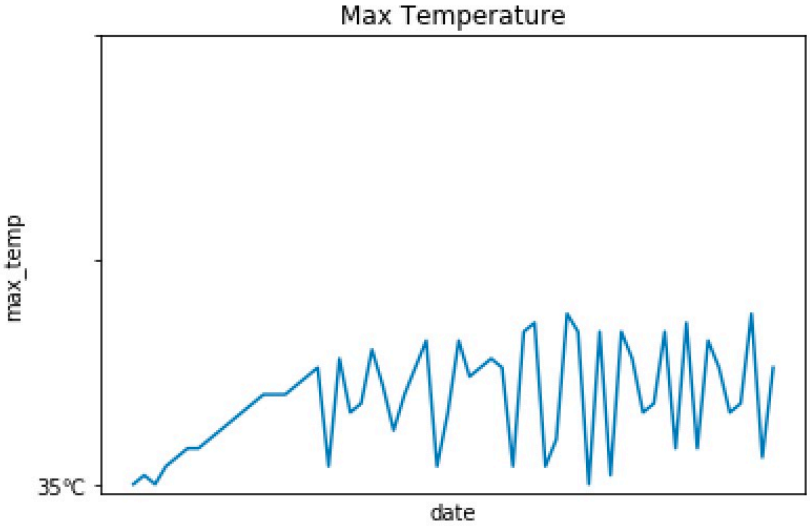
```
Hotel Price in Adelaide
```



Based on the output of the price range of the customer, our program displays the hotel information as a pie chart. Customers can know what percentage of the price range takes in the selected city.

c). Temperature Line Chart

```
Weather Information in Line Chart
```

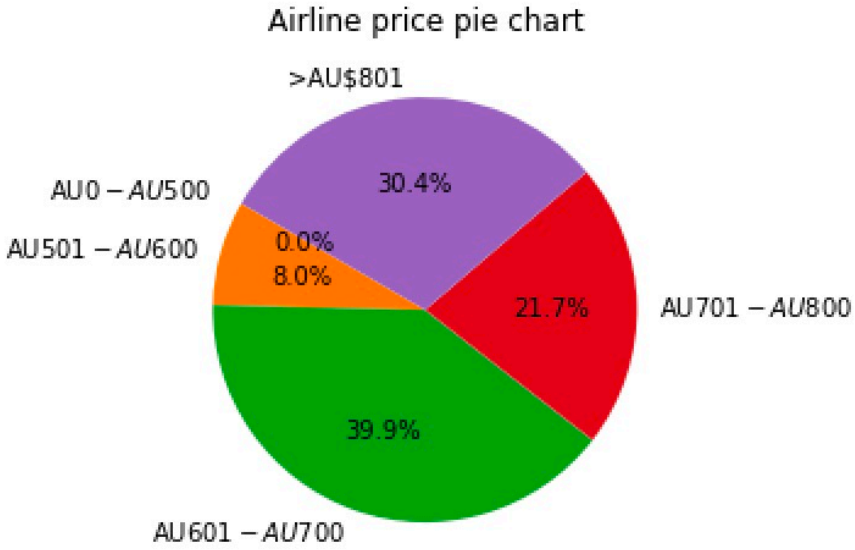



```
Weather detail:
     min_temp max_temp
0101     15.8     28.8
0102     17.4       28
0103     16.2     27.8
0104     17.8       29
0105     20.2     31.4
...       ...      ...
0226     15.6       27
0227     18.4     30.8
0228     16.2     29.8
0229       19       31
0301     17.4     29.6

[61 rows x 2 columns]
```

Our program will display the weather changes in the selected city. Detailed and past year's temperature will be displayed.

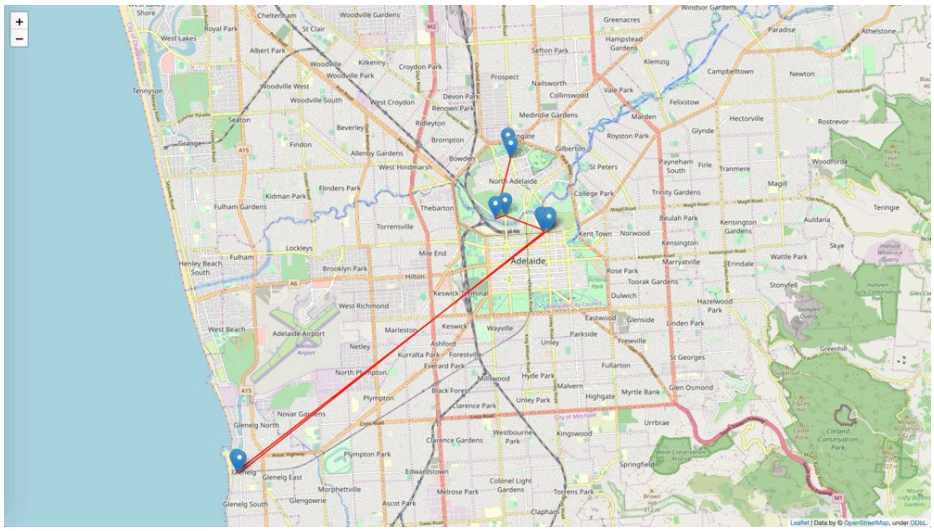
d). Airline Price Pie Chart

```
Airline Information in Line Chart
```



Airline information will be displayed as pie chart.

e). Map Visualization



In the `map/map.html` , a route map will be displayed to guide the customers where to go and how to play in selected city.

f). Plan Steps

```
Route plans detail:
Step 1: O'Connell Inn
Step 2: Stop 6 O'Connell St - West side
Step 3: 多伦斯河带状公园小径
Step 4: Stop 2 Montefiore Rd - West side
Step 5: 南澳大利亚艺术馆
Step 6: Stop G1 North Tce - North side
Step 7: 格雷尔海滩
Step 8: Stop 17 Moseley Square
Step 9: 阿德莱德大学
Step 10: Stop Tram University
```

A more specific route plan will be displayed below.

g). Cost and Distance Steps

```
Costs and distance
+----------+-----------+
|   Item   | Money($)  |
+----------+-----------+
|  Hotel   |   105     |
| Airline  |   511     |
|   Bus    |   5.0     |
|  Total   |  621.0    |
+----------+-----------+
Walking Distance in city: 847.46m
```

65% Costs and diatance information

about the route are displayed as table.

h). Word Clouds

```
Place comments cloud:
多伦斯河带状公园小径
Loading model cost 0.569 seconds.
Prefix dict has been built succesfully.
```

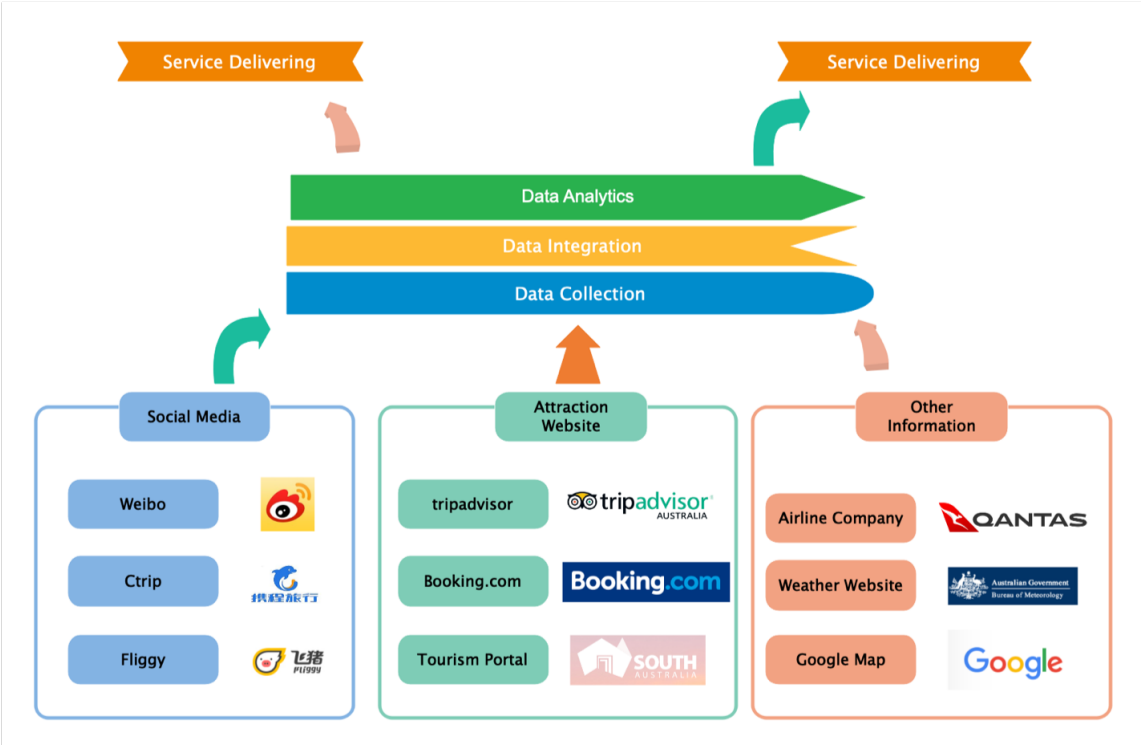
Word clouds corresponding with every spots are listed below. Each cloud demonstates the word that has the highest frequency in the comments.

3. Scrape

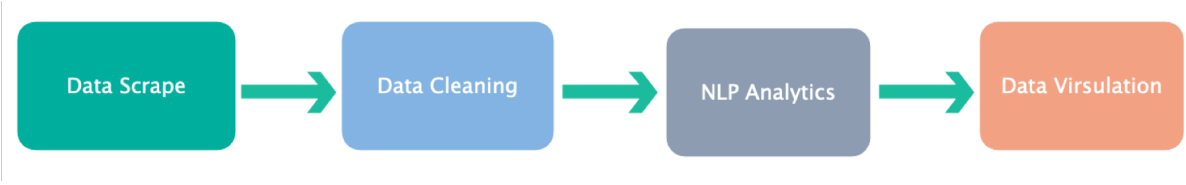
3.1. Abstract

Data sources we use include the follows:

- Ctrip comment data
- Weibo data
- Tripadvisor hotel data
- Weather data
- Airline data
- Adelaide Metro Bus data



3.2. Analytics Process



3.3. Ctrip Comments Scrape

Before running the flight data scraping program, please follow these steps:

- Download and install Chrome browser
- Download chromedriver from https://sites.google.com/a/chromium.org/chromedriver/downloads. You must choose the chromedriver version based on your chrome browser version.
- Uncompress the file and paste it to Python installation directory.

Ctrip Scraping Library

Library	Version
pandas	0.25.1
beautifulsoup4	4.8.0
selenium	4.0.03

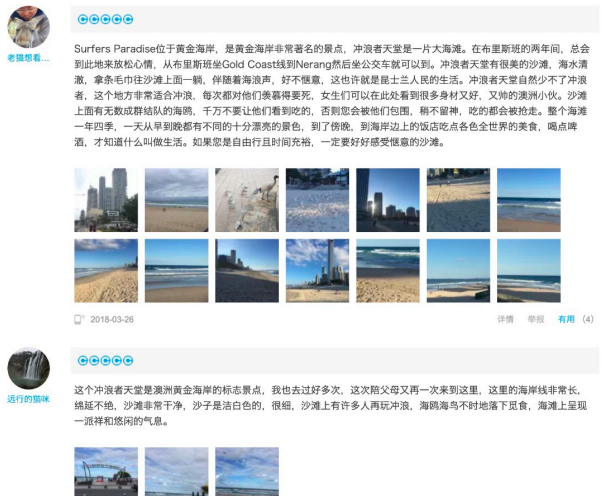
Running

Run `ctrip_comment.py` to scrap the hotel information in Ctrip websites. After the sunning of this file, you will get the final raw comment data.

Web Pages

Ctrip.com International, Ltd. (doing business as Ctrip) is a Chinese provider of travel services including accommodation reservation, transportation ticketing, packaged tours and corporate travel management.

In the codes, we discover two different page format in the comments page. So there are two parse logic, displaying following pages:



Basic logic

- Get all tourism urls from comment website `https://you.ctrip.com/countrysightlist/australia100048.html`
- Simulate Browser: get all the data hidden in html
- Simulate click: trun page

```
page=1
while(page<=5):
    user_list = body.find('ul', {"class": "comments"})
    for user_info in user_list:
        for li in user_list:
            entry = {}
            h4 = li.find("h4")
            p = li.find("p")
            user_time = li.find("div", {'class': 'user-date'}).find('span')
            entry["city"] = city_name
            entry["place"] = place_name
            entry["rating"] = h4.text
            entry["comments"] = p.text
            entry["user_time"] = user_time.text
            entry["url"] = href
            df = df.append(entry, ignore_index=True)
    next_page = browser.find_element_by_class_name("down ")
    next_page.click()
    print(href + " page " + str(page) + " has been finished.")
    page += 1
```

- Use beautifulsoup to collect data

Cleaning
Run clean_data.py to clean the data.

3.4. Hotel Data Scrape

Hotel Scraping Library

Library	Version
Library	Version
pandas	0.24.2
numpy	1.16.1
beautifulsoup4	4.7.1
selenium	3.141.0

Running
Run Hotel*Worm.py to scrap the hotel information in different Australia cities (Sydney, Melbourne, Adelaide, Canberra, Brisbane). After the sunning of this file, you will get four csv files which named as raw*data*hotel*SD.csv, raw*data*hotel*MEL.csv, raw*data*hotel*AD.csv, raw*data*hotel*CAN.csv, raw*data*hotel*BR.csv.

Cleaning

Run Hotel*Clean.py to clean the data, which will replace the city column data from number to city name. Then it will slice the "AUD$" in price column. Finally, it will output Hotel*Clean.csv.

Description

This script file will scrape hotel information from https://www.trivago.com.au/. The data fields include city, hotel name, customers' rate, location and price. It is about different day's price for each hotel from 2019.11.01 to 2019.11.29.

parameter	type	description	example
city	string formula	The city which you want to look up	"Adelaide","Sydney","Melbourne","Canberra","Brisbane"
time	string formula	the start date and end date you want to look up	DD-DD, like "01-02", "29-30"
name	string formula	Hotel name	"InterContinental Sydney"
rate	string formula	Customer's rate	"Excellent"
type	string formula	Hotel type	"Hotel"
location	string formula	Hotel location	"Sydney, 0.7 km to Sydney Opera House"
price	string formula	Hotel price	"342"

3.5. Airline Data

Flight Scraping Library

Library	Version
pandas	0.25.1
XlsxWriter	1.2.1
beautifulsoup4	4.8.0
selenium	4.0.0a3

Before running the flight data scraping program, please follow these steps:

- Download and install Chrome browser
- Download chromedriver from https://sites.google.com/a/chromium.org/chromedriver/downloads. You must choose the chromedriver version based on your chrome browser version.
- Uncompress the file and paste it to Python installation directory.

Then you can run the program and it will automatically scrape data from Trip.com and write it to Excel. The destinations include Sydney, Melbourne, Brisbane, Canberra and Adelaide.

Data cleaning code

Run the program and the names of departure airports and arrival airports will be translated into common language. Also the 'AU$' will be eliminated in the price column.

3.6. Weather Data

Weather Scraping Code

Library	Version
pandas	0.25.1
requests	1.2.1
beautifulsoup4	4.8.0

Running:

Start running the project.

Description:

You'll scrap the weather data of the 5 major cities in Australia——Canberra, Adelaide, Sydney, Melbourne and Brisbane from 2013-11 to 2018-05. The weather information including the date, the weather description about the starting of the day, the weather description about the end of the day, the highest temperature of the day, the lowest temperature of day and the wind information and so on.