**TRAVEL COMPANION**

**Project Report**

Submitted by

**Arpit Khurana (15BCE0353)**

**Arpit Puri (15BCE0732)**

In fulfillment for the project of

**Web Mining**

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING





**School of Computer Science and Engineering**

**DECLARATION**

We hereby declare that the project entitled TRAVEL COMPANION submitted by us to the School of Computer Science and Engineering, VIT University, Vellore-14 in fulfillment of the requirements for the award of the project of web mining in Computer Science and Engineering is a record of bonafide work carried out by us under the supervision of SHASHANK MOULI SATAPATHY, Associate Professor. We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other project of this institute or of any other institute or university.

**Arpit Khurana (15BCE0353)**

**Arpit Puri (15BCE0732)**



**School of Computer Science and Engineering**

**CERTIFICATE**

The project report entitled Distributed Systems Banking Application is prepared and submitted by Arpit Khurana (Register No: 15BCE0353) and Arpit Puri (15BCE0732). It has been found satisfactory in terms of scope, quality and presentation as partial fulfillment of the requirements for the award of the project of Web Mining in Computer Science and Engineering in VIT University, India.

SHASHANK MOULI SATAPATHY

Associate Professor

**Guide**

**1. Introduction**:

In this project we want to develop an application that can be used for planning user trip. With the help of this application user can search for different location and get information about the places and their features.

Also it will provide the information about the hotels and restaurant in the places. We can also search for a hotels and restaurants and get all information about it. Finally it will analyse the user reviews and ratings and help people choose what is more suitable for them.

**2. Background Information**

To source data for data science projects, we often rely on SQL and No SQL databases, APIs, or ready-made CSV data sets. The problem is that we can’t always find a data set on our topic, databases are not kept current and APIs are either expensive or have usage limits. If the data you’re looking for is on a web page, however, then the solution to all these problems is web scraping.

We want to analyze the various websites like google places, tripadvise.com and zomato.in etc. data to solve our problem.

Essential steps for web scraping:-

i) Working out which pages to scrape

ii) Identifying the URL structure

iii) Understanding the HTML structure of a single page

iv) Use library to parse the HTML content

v) Extracting the particular data

vi) Examining the scraped data

**3. Objective of the project:**

People love traveling. So they have a large no of options and they don’t know what will be better location for them. So what should they do to decide? For this purpose, they go to various websites and manually search for location and analyze various reviews before deciding.

People open google and search for a specific location and read reviews about the place. Then they search for the hotels in the location in various websites and again search for rating and reviews. After that they have to search restaurants near the location. This is a lot of work. So we want to develop an application which gives us all the information related to places like hotels and tourist spots in a single location. Through this application we can get information like ratings, distances, features and availability of various places all confined at one place so that user does not have to search at different locations for small things. Also it will show analysed data and provide comparison He/she will find it at one place i.e.in a single application.

**4. Proposed Steps**

Main steps:-

i) Search online for trending places and show user suggested places.

ii) Ask user to enter data like location, hotel or restaurant. Extract the name of the location from the entered data.

iii) Search related Website for corresponding place.

iv) Download the content of the website for scraping.

v) Identifying the URL structure

vi) Understanding the HTML structure of a page from which data need to be scraped.

vii) Use library to parse the HTML content

viii) Extracting the particular data like review, rating, geographic location, transportation required to reach that destination etc;

ix) Examining the scraped data

xi) Do sentiment analysis to analyze user review data.

x) Store it in a csv file or show as an html page to user.

**5. Innovation component in the project:**

The innovative part of our project is that contrary to the popular sites such as TRIVAGO and others which searches and compares the different hotels as per the user need, it plans the whole sum journey for the user. From transportation of the trip to the hotels they stay in, the restaurants and the locations where they have to visit.

Our projects takes the destination and the source, the dates as input for planning the journey. Then it compares the flights and the hotels simultaneously with the help of reviews of the people and the offers given by the websites.

**6. Work done and implementation**

**Methodology adapted:**

There are 3 main steps to scrape a tourist place page:

i) Request to visit that particular webpage, just like what your browser does, and download the HTML contents into your environment. I used requests module to accomplish this. urllib2, urllib3 are also available choices.

ii) Beautiful Soup module can be used to parse HTML contents. In addition to that, it provides very powerful and useful functions, searching through the soup object to match for text and HTML tags within the page.

iii) Extract what you want from the webpage and download/store them for further analysis.

Sentiment Analysis:

i) Remove stop words from the text

ii) Give machine a lot of data so that it can learn how to analyze sentiment

iii) Create an REST API so that we can analyze future data

**Tools used:**

**Language:** Python

**Version:** Python 2.7

**i)** **Beautiful Soup:** Beautiful Soup provides a few simple methods and Pythonic idioms

for navigating, searching, and modifying a parse tree using Python parsers like lxml

and html5lib. It automatically converts incoming documents to Unicode and

outgoing documents to UTF-8.

**ii)** **Selenium** The web driver kit emulates a web-browser (I chose chrome driver) and

executes the JS scripts to load the dynamic content.

**iii)** **Text Blob:** Text Blob is a Python (2 and 3) library for processing textual data. It

provides a simple API for diving into common natural language processing (NLP)

tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis,

classification, translation, and more.

**iv)** **Pandas :** pandas is a Python package providing fast, flexible, and expressive data

structures designed to make working with “relational” or “labeled” data both easy

and intuitive. It aims to be the fundamental high-level building block for doing

practical, real world data analysis in Python.

**v)** **NumPy :** NumPy is a library for the Python programming language, adding support

for large, multi-dimensional arrays and matrices, along with a large collection of

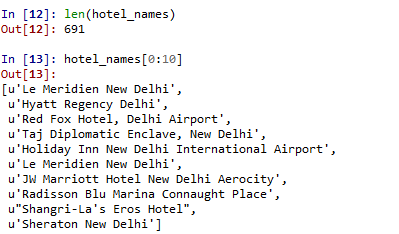
high-level mathematical functions to operate on these arrays.

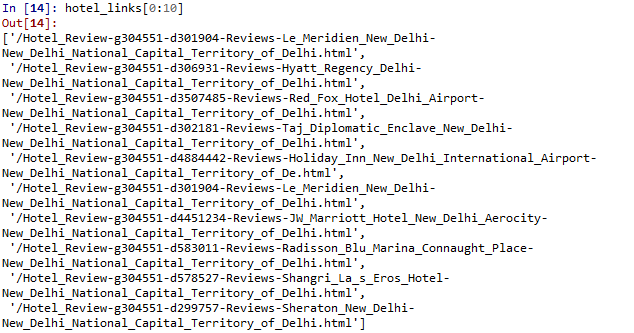
**vi) Jupyter:** The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. The Notebook has support for over 40 programming languages, including Python, R, Julia, and Scala. Leverage big data tools, such as Apache Spark, from Python, R and Scala. Explore that same data with pandas, scikit-learn, ggplot2, TensorFlow.

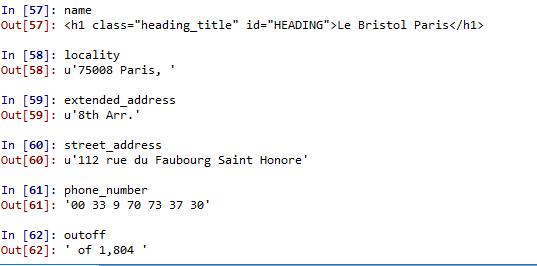
**Results and Discussion :**

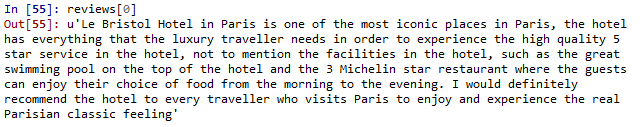
* ***Hotels***: we start by training the selenium driver to extract data from a particular site by filling details according to the input given by the users. Once the details have been filled up then all the hotels in that particular city will be found by running a loop in the script upto the last page. Also a link to the particular hotel detail page would be found which will give us the price details, address, rating and comments about that particular hotel that will help the user to select the appropriate hotel. All this data scraping is done by the help of a tool named beautiful soup.



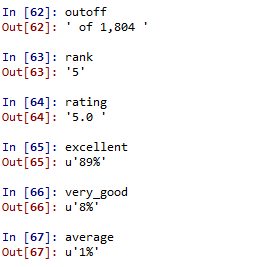


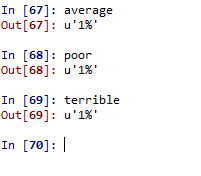






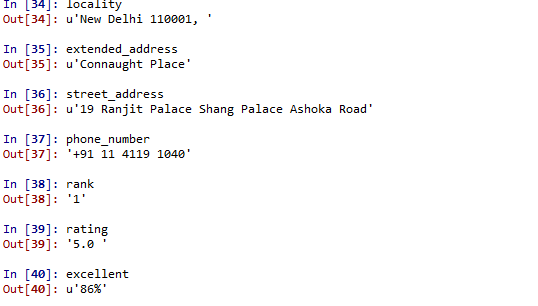


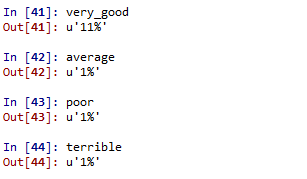




* ***Restaurants***: we start by training the selenium driver to extract data from a particular site by filling details according to the input given by the users. Once the details have been filled up then all the restaurants in that particular city will be found by running a loop in the script upto the last page. Also a link to the particular restaurants detail page would be found which will give us the price details, address, rating and comments about that particular restaurants that will help the user to select the appropriate hotel. All this data scraping is done by the help of a tool named beautiful soup.

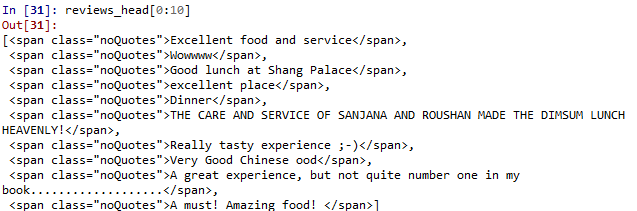


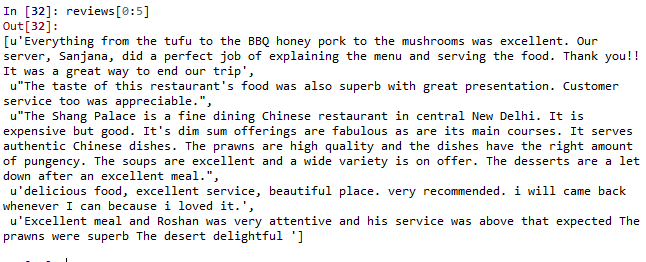






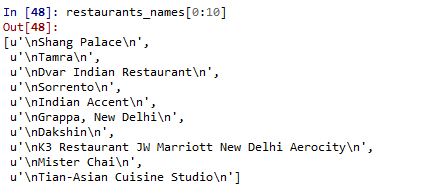




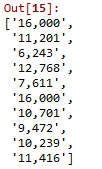




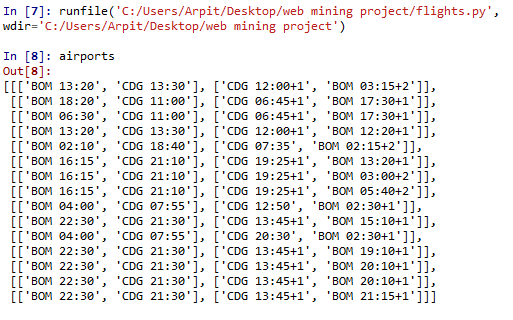


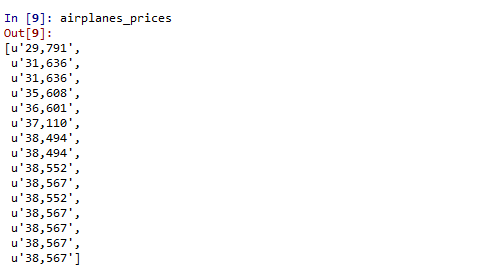


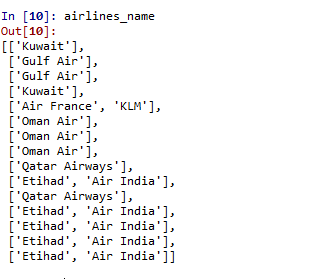




* ***Flight Details***: we start by training the selenium driver to extract data from a particular site by filling the details of flight i.e. whether we are planning to book the tickets for one-way trip or multi-way trip, dates of arrival and departure, etc. Once the details have been filled up then all the flights corresponding to the details will be found out and the user can select the appropriate flights.

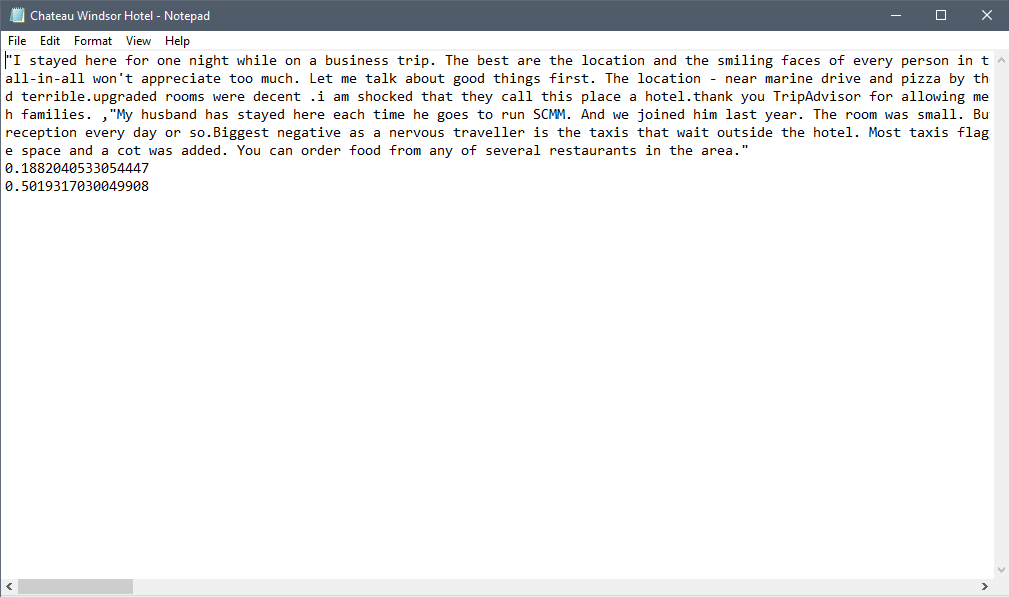




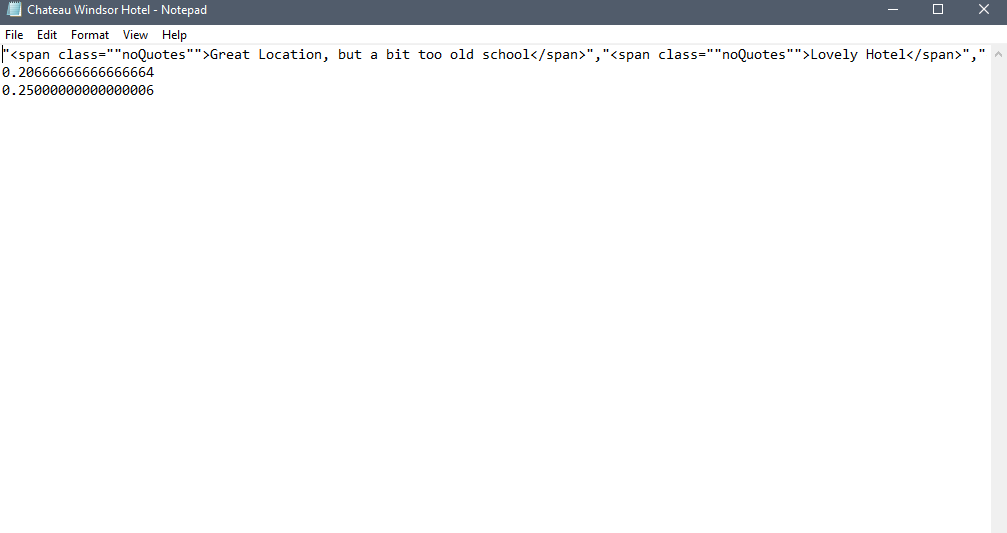


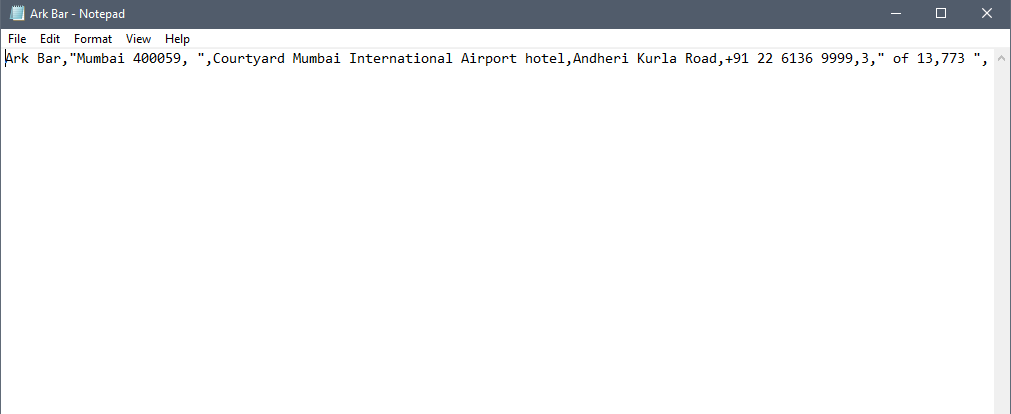
**CSV File Output**

**Hotels Review**

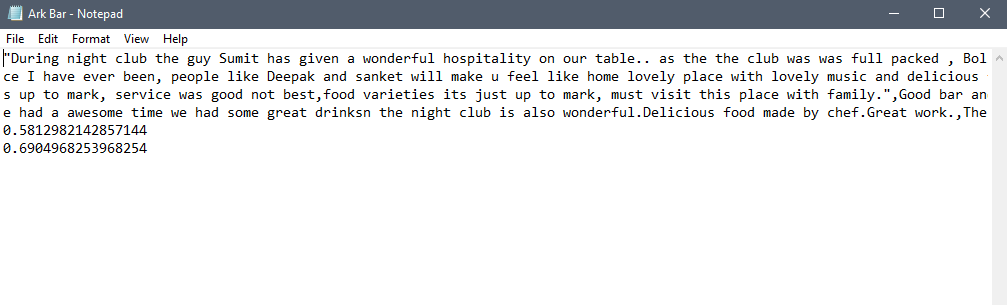


**Hotels Review Head**

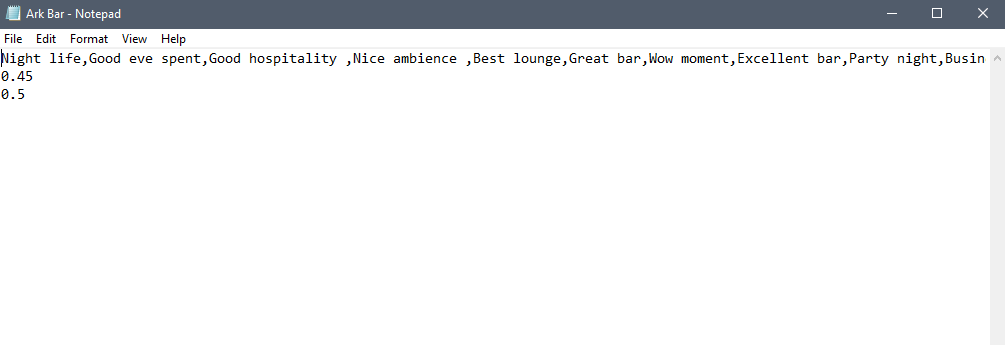
**Restaurants Detals**



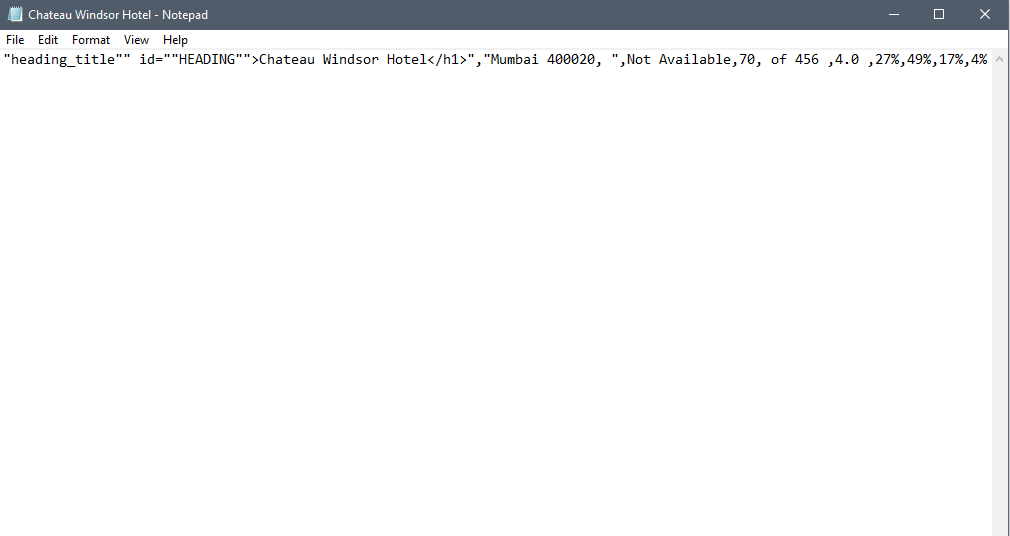
**Restaurants Review**



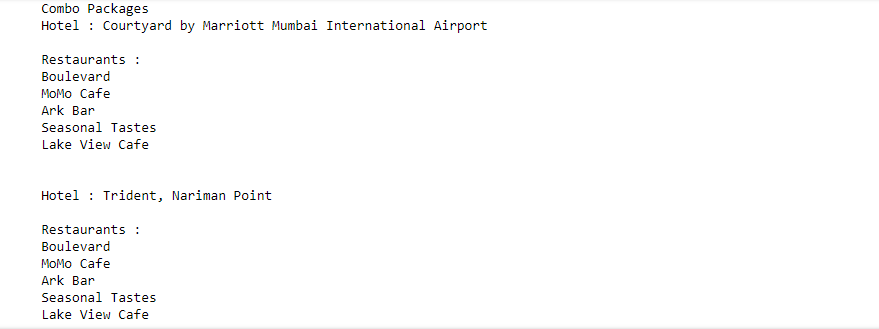
**Restaurants head review**



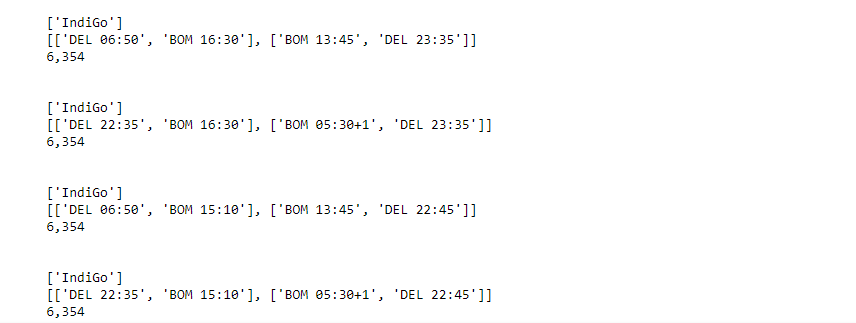
**Hotels Details**



**Combo Packages**



**Recommended Flights**



**7. Results and discussion**

Our project basically deals with the idea of providing a cheap and a memorable trip to all the users who has subscribed to our site. On our site we have different exciting combos and packages for the users which they can choose according to their convenience. Our combos sets up a complete package which includes, flight tickets, restaurants reservations and also hotel reservations. There is a huge list of exotic restaurants and hotels to choose from and the list contains only the best places in the entire city. For the case of restaurants, you can select from your favorite cuisines including Indian, Chinese, Mughlai, etc. and select dates on your trip and make reservations. For the case of hotels, one can select hotels according to the fare or rating like hotels with 5 star rating are expected to be more expensive than those which are rated low and finally for flight ticket one can enter the details of their arrival and departure place and dates and can select from a list of flights available on that day. Our site is different from other sites in such a manner, that our site provides an entire combo and the user have to select from the list of combos of restaurants, hotels and flights whereas in other sites people individually first select hotels, restaurants and then hotels. Also our ratings is basically based on the reviews posted by the people for the particular hotel or a restaurant. By using sentimental analysis on the comments posted by the people on different sites we were are able to extract the best of the hotels and the restaurants out of all options available in a particular city and make suitable combos including the selected places. We basically aim to provide our users with the best service.

The combos are based on sentimental analysis scores of the restaurants and hotels. The combo with higher restaurant to hotel ratio will be showed on the top of the combos list. We aim to calculate the ratio for all the top hotels and restaurants like for e.g. top 10 hotels and restaurants(100 combos), and make and show only 20 combos(approx.) out of total list of 100 which has the highest restaurant to hotel ratio.

**8. Conclusions:**

The project mainly deals with the topic of web scraping, where we extract information about different tourist places and hotels thereby giving us the rating, the geographic location and availability of a particular place. This further helps the user to decide which place is the best to go considering their convenience. Also the user can go through the reviews posted by different users who have already visited that place. The users will also be able to see the list of the top trending places in a particular week including the all-time tourist places.

**9. References**

i) Mitchell, Ryan. Web scraping with Python: collecting data from the modern web.Reilly Media, Inc, 2015.

ii) Nikovski, Daniel N., and Alan W. Esenther.Method for Extracting Data from Web Pages. U.S. Patent Application No. 12/239,859.

iii) McKinney, Wes. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython.Reilly Media, Inc, 2015.