CS 839 Spring 2018, Project Stage 2

Team Members:
Arpith Neelavara (neelavara@wisc.edu)
Bhargav Tangirala (btangirala@wisc.edu)
Aribhit Mishra (amishra28@wisc.edu)

Web Sources

We decided to extract Restaurant data from the following 2 sites:
https://www.yelp.com/
https://www.tripadvisor.com/
We decided to use these sites as they showed up frequently on web searches when we searched for Restaurant reviews.
The data in Yelp was very well-structured and had tags with easily-distinguishable class attributes.
The data from TripAdvisor was slightly trickier since some of the attributes we wanted (like opening and closing times), did not exist for all restaurants.

Tools Used

We used an open-source Python package called BeautifulSoup (https://pypi.python.org/pypi/beautifulsoup4) to extract the data. It parses any HTML page as a list of tags and its attributes. We can then search for specific tags based on some attribute, and then extract the text field (which corresponds to the text being displayed on the web page).

Data Extracted

We decided to extract data relating to Restaurants (name, address, type, etc.). We first checked restaurant data in Madison, but there did not appear to be enough data to fill 3000 tuples. Thus, to ensure that we had sufficient number of tuples, we picked Restaurants in New York. Any other major city might have worked, but we decided on New York arbitrarily.

Finally, we extracted 2 CSV files corresponding to data extracted from yelp.com and tripadvisor.com.
Yelp_Restaurants.csv has 3160 tuples.
TripAdvisor_Restaurants.csv has 3891 tuples.

Yelp_Restaurants.csv has the following attributes:

- Name: Name of the restaurant
- Address: Address of the restaurant
- Phone: Phone number of the restaurant
- Cuisines: Cuisine the restaurant serves (separated by ';')
- Take Out: Yes/No depending on whether restaurant offers Takeout service

- Saturday Opening Time: Time in 12 hour clock with AM/PM or 'CLOSED' or 'OPEN 24 HOURS'
- Saturday Closing Time: Time in 12 hour clock with AM/PM or 'CLOSED' or 'OPEN 24 HOURS'
- Sunday Opening Time: Time in 12 hour clock with AM/PM or 'CLOSED' or 'OPEN 24 HOURS'
- Sunday Closing Time: Time in 12 hour clock with AM/PM or 'CLOSED' or 'OPEN 24 HOURS'

TripAdvisor_Restaurants.csv has the following attributes:

- Name: Name of the restaurant
- Address: Address of the restaurant
- Phone: Phone number of the restaurant
- Cuisines: Cuisine the restaurant serves (separated by ';')
- Saturday Opening Time: Time in 12 hour clock with AM/PM or 'NaN' if data was missing
- Saturday Closing Time: Time in 12 hour clock with AM/PM or 'NaN' if data was missing
- Sunday Opening Time: Time in 12 hour clock with AM/PM or 'NaN' if data was missing
- Sunday Closing Time: Time in 12 hour clock with AM/PM or 'NaN' if data was missing
- Take Out: Yes/No depending on whether restaurant offers Takeout service