Stage 1 - Data Retrieval:

We have two datasets to begin with, reviews.json (a list of reviews) and businesses.json (A list of restaurants and identifying data). Both were provided by Yelp, as part of academic purposes agreement.

1. Use businesses.json to extract restaurant names, address, cuisine and identifying ID for reviews.json (which uses an ID rather than restaurant name to identify a review's target)
2. We assign each restaurant a unique ID, using an increment as we process them then pass the data to two dictionaries
   Restaurant_dict: which uses the unique ID as a key, and a list of data as its values
   Food_dict: which uses cuisines as keys and lists of unique IDs as values

Eg: Restaurant 1 is a McDonalds, assign it a UID of 1
Restaurant_dict = {1:[Name, Address, List_of_reviews, etc]}
Food_dict = {American: ['1'], Burgers: ['1']}

Restaurant 2 is a Mexican themed burger restaurant, assign it a UID of 2
Restaurant_dict = {1:[Name, Address, List_of_reviews, etc], 2:[Name, Address, List_of_reviews, etc]}
Food_dict = {American: ['1'], Burgers: ['1', '2'], Mexican ['2'}

We plan to use this form of data structure, because our input on the consumer-facing website will take a cuisine or cuisines as an input and return relevant restaurants. Taking the example above, entering Burgers will return restaurant UIDs 1 and 2, which we can then use to retrieve data from the restaurant_dict.

Stage 2 - Ranking and Twitter introduction

1) We can use the review score of each restaurant to determine a list of the top ten restaurants, based on their Yelp review score and number of reviews.
2) Using this curated list of restaurant names and information, we can use the twitter API to identify tweets regarding the restaurant. We have a python module that allows us to access the official twitter search API, and authentication tokens from Twitter.

Stage 3 - Data Analysis:

1) Using a python module, VADER, we can perform sentiment analysis on tweets to assign them a score that is representative of whether they are negative or positive.
2) We weight these tweets according to their likes and the poster's popularity amongst their followers
3) We can then compare the overall disposition of tweets against the general disposition of yelp reviews
4) Using this information, we can adjust the ranking of the ten curated restaurants

Stage 4 - Data Visualization and Representation:
Checklist:

1. Assemble a website
2. Input box, user can input a cuisine or cuisines of their choice
3. Create or access visual tools to display our data, eg word clouds or meters to represent sentiment. Particularly if, for example, tweet sentiment does not match yelp reviews