Environmental Consciousness in the Restaurant Business

PREPARED FOR

Professor Anasse Bari

NYU, Department of Computer Science

PREPARED BY

Nellie Spektor (ns104@nyu.edu)

Andrea Waxman (aw2860@nyu.edu)

Valerie Angulo (vaa238@nyu.edu)

MAY 20, 2019

Abstract

Can being more environmentally friendly also be good for a restaurant's bottom line? This study examines the correspondence between using environmental practices in restaurants and their success. As data scientists, we intend to provide data analysis and statistics that will prove that adopting green or environmental friendly practices will drive business for restaurants and improve their ratings. To accomplish this, we must find a way to measure how environmentally conscious a restaurant is, in addition to assigning a benchmark to determine a restaurant's success. Using a collection of data from the Green Restaurant Association[1], Seafood Watch[2], and Yelp's published dataset[3] for business and user reviews, we aim to create a dataset to refer to and perform data analysis for our intended goal. Using customer ratings on Yelp, we cross examined our restaurant dataset to determine the restaurants' success to test our hypothesis: being green is *also* good business.

[1] Green Restaurant Association
[2] Seafood Watch
[3] Yelp

I. Literature Review

Before undertaking this project, we conducted extensive research about the topics we would be exploring. As we looked into what it means to be green for a restaurant, the best definition we found for our case was, compared to a traditional restaurant, a green restaurant devotes effort to the three R's: reduce, reuse, and recycle and the two E's: energy and efficiency (Gilg et al., 2005).[4] We found no works directly related to what we planned to do, but we did find some works that indirectly guided our understanding and solutions. There has been research done exploring consumer perceptions of green restaurants in the US as well as predicting restaurant success through Yelp data. Some notable papers found among our research are the following: "The Effects of Restaurant Environment on Consumer Behavior"[5], "Using Yelp Data to Predict Restaurant Closure"[6], "Predicting International Restaurant Success with Yelp"[7], and "Effects of restaurant green practices on brand equity formation: Do green practices really matter?"[8]. These papers were helpful in formulating our perception of the industry and what has already been explored through data. We derived that many consumers do care about how eco-friendly a restaurant is[9] and that Yelp contains valuable information about restaurant success because of its extensive database. Yelp is able to provide a multitude of data points to help analyze businesses listed on its site, including but not limited to, restaurants. This was good news for us and for any other works that have dealt with restaurants. The different attributes like business ID and name that Yelp data provides has worked for research in the past, as found in the paper "Predicting International Restaurant Success with Yelp". Furthermore, although previous research has paid attention to consumer attitudes and behavioral intentions toward restaurants that offer environmentally friendly foods or practices, the effects of green practices in association with brand equity formation are still under explored[10] (Young Namkung et SooCheong Jang, 2013). In an attempt to further explore this topic, Young Namkung and SooCheong Jang looked into finding information about the questions 'Does implementing green practices in restaurants influence customer-based brand equity formation, such as perceived quality, green image of a restaurant, and behavioral intentions?' and 'Are the effects of green practices different across different restaurant segments?' Their findings were derived from survey based data and just 512 valid

[4] Effects of Restaurant Green Practices on Brand Equity Formation: Do Green Practices Really Matter?
[5] Effects of Restaurant Environment on Consumer Behavior
[6] Restaurant Closure Predictions
[7] Predicting International Restaurant Success with Yelp
[8] Effects of Restaurant Green Practices on Brand Equity Formation: Do Green Practices Really Matter?
[9] Effects of Restaurant Environment on Consumer Behavior
[10] Effects of Restaurant Green Practices on Brand Equity Formation: Do Green Practices Really Matter?

responses from consumers and only focused on one type of restaurant: American diners. While their conclusions were valuable in extending the research that we are also tackling, there is still much room for improvement and differing perspectives. Our intention is to fill the research gap that exists in this particular matter. Finally, we discovered the existence of The Green Restaurant Association (2011) (GRA) which provides a nationally recognized green restaurant certification to any restaurant who meets the following requirements: water efficiency, waste reduction & recycling, sustainable furnishings & building materials, sustainable food, energy, disposables, and chemical & pollution reduction. Due to the careful and precise selection of restaurants which may be certified by their association, we chose to use the GRA as one of our main data sources for green restaurants. In conclusion, we have not been able to find any research which looks at green restaurants on a large scale. Therefore this is a great opportunity to add more data analysis about the industry to the world.

II. Business Understanding

With the global warming issue becoming a trending topic in the past decade and environmental concerns coming to light, many of us have become aware that businesses tend to be the number one source of energy consumption and waste disposal in the world, hurting our environment greatly. Taking this knowledge a step further, we looked into the restaurant business and the statistics in relation to the environment. These were some of our findings: the average restaurant uses 500,000 kilowatt hours (five times the energy of other commercial buildings), consumes 800,000 gallons of water, and discards about 100,000 pounds of waste every year. We are in the midst of an environmental revolution in which many restaurants have adopted environmentally friendly practices. Using recyclable or compostable packing materials, designating receptacles for trash, recycling, composting, and using organic or local ingredients, are just a couple of ways restaurants are playing their part in this revolution. A revolution that has sparked environmental conversations, a rise in the vegan and vegetarian lifestyles, and an increasing interest in using organic or local produce, over the last decade. That being said, there are still numerous restaurants and businesses that do not heed the consequences of their immense energy and water consumption and hefty waste disposals. We know that there are restaurants that *do* make an effort to reduce their carbon footprint and make the environment a priority, but do we know if it is profitable? Would it just be easier for owners to ignore the environment and focus on the business and ROI of the restaurant? Or is there a correlation between the two that restaurant owners might find interesting? By learning more about the industry and their relationship with our ecosystem, we formulated the problem we aim to help solve with data. We hypothesized that restaurants can be more successful if they are environmentally friendly.

III. Data Understanding

To test our hypothesis, the data sources that we chose to work with were the Yelp dataset, Green Restaurant Association data, Seafood Watch data, and alternative data sources such as Restaurant Inspection data and environmental food blogs. The semi-structured data that we worked with was the Yelp dataset and the restaurant blogs, which were in a text list format. The Yelp dataset contains over 50,000 restaurants and their reviews in json format, with attributes such as business id, address, city, state, zip code, latitude/longitude, business categories and text reviews. The Green Restaurant Association sent us a csv that contains the names of almost 600 restaurants, their address, city, state, zip code and a green rating, a 1-5 point value determined by the organization that represents the level of a restaurant's environmental practices. These practices include composting, using recyclable materials, using energy-saving methods, etc. The Seafood Watch data was used to supplement the Green Restaurant Association data since there wasn't much data available from the GRA. The Seafood Watch sent us a csv that contains the names of 1,825 businesses that were verified to have good seafood practices, such as making sure the food was fresh and local, that the fish were caught in an environmentally responsible manner, and that food preparation was safe. The alternative sources that we looked at were blog posts from Plant-Based Restaurant Guide NYC and NYC Restaurant Inspection Grades data. Plant-Based Restaurant Guide NYC was used to manually supplement our green dataset with more environmentally responsible restaurants; this data was in the form of text blog posts listing green restaurants. We wanted to use the NYC Restaurant Inspection Grades data to see if other features besides a restaurants environmental consciousness led to a restaurants success, but ultimately did not use this set for reasons that will be discussed in the next section. This dataset included restaurant name, location and grade. With these datasets, we wished to be able to predict which restaurants were green based on the yelp text reviews and from these predictions, create our own green classifier.

IV. Technical Obstacles

We initially wanted to look at green restaurants in New York City, but upon exploring the yelp business dataset, we found only 13 entries for New York City from using a python dictionary containing variations of New York and searching for rows that contained these values in the city column and the value "New York" in the state column. From these 13 business entries, only 3 were restaurants. From this discovery, we realized that the yelp data we were using was a subset of a much larger dataset. We then decided to broaden our data to all restaurants in the US contained in the yelp dataset. When looking to integrate the Restaurant Inspection Rating data to our set, we found that almost all of the restaurants had a grade of "A" and that there weren't many matches on restaurants in our set, mostly because this datasource is a NYC Public Dataset. Because we only had 3

restaurant entries for NYC we decided not to incorporate the restaurant inspection grade into our project.

After joining the GRA and Seafood Watch data with our cleaned and filtered yelp data, we found that there were only 28 matches. Not having enough data on environmentally friendly restaurants was the biggest challenge for our project. We also had trouble using RapidMiner on the data because RapidMiner requires at least 100 rows of data to run analyses on and we wanted to run analyses on our rows that matched with GRE data and Seafood Watch, which totaled less than 100 entries. To combat these issues, we decided to mark restaurants as green if their text reviews contained certain keywords, which suggested that reviewers were applauding the restaurant for its eco-friendliness. We also decided at this point to manually supplement our available green restaurant data with information from environmentalist blog posts.

V. Data Preparation

To clean and prepare our data, we used Python. Specifically, we used the pandas, NLTK, and NumPy python packages. The first step in combining our data was converting the yelp json files into csvs. To do this, we found a script online that formats json files to csv using pandas. Our next step was to remove businesses that were not restaurants from the yelp business data. We used a python dictionary that contained restaurant related words such as “food”, ”drink”, ”dessert”, ”bakery”, ”diner”, etc. which we found to be representative of restaurant businesses in the categories column of the yelp businesses. We were then able to keep only rows that contained one or more of the restaurant dictionary words in the categories column. Because yelp provided the reviews in a separate file from the business data itself, our next step was to merge the yelp reviews data onto our yelp business data by left-joining the csvs on business id. Although the business id column contained all unique ids, there were many duplicate restaurants. For example, every McDonalds restaurant had a unique business id. We wanted to look at chains as a whole, so we grouped the data by name and updated the column containing user’s reviews to have a list of all reviews from all locations of a chain. This assumes that every restaurant which is not a chain has a unique name; although this is most likely not the case, this was the easiest way to deal with restaurant duplicates and we figured that there would be very few individual restaurant that shared names and the number of errors caused by this merge would be small. The columns that we kept from our combined yelp dataset are business name, address, city, state, zip code, latitude/longitude, review, and yelp star rating. After getting our merged yelp dataset, we needed to integrate our other data sources. To do this, we merged the restaurants from the Green Restaurant Association with the yelp dataset by name, also creating a new column that contained the GRA’s green rating, which was 0 if the restaurant did not exist in that dataset. We added 1 to a restaurant’s GRA rating if it could be found in the Seafood Watch dataset. This might not have provided a the ideal set of results due to the exact matching of the restaurant names, so we tried to account for this

by writing a python script that utilizes Spark's DataFrames for fuzzy matching using Levenshtein distance. However, from trying this out with a Levenshtein distance of 1, we obtained many matches that were incorrect, so ultimately did not end up using this method. To add in data collected from blogs, we manually inserted a score for restaurant names mentioned in the green restaurant blogs. Because our yelp restaurant data was so limited, we decided to give our own green rating through finding restaurants whose yelp reviews contained environment-related words. These were words such as "compost", "recycle", "vegan", "local", and more. If these words comprised over 1% of the total words in a restaurant's reviews, we labeled the restaurant "green". From this method, we were able to increase our data size to about 2,600 green restaurants and added an equal number non-green restaurants to our set in order to have a balanced dataset on which to train a model.

name	review_text	green_boolean	overall_green_rating	GRA_rating	green_words_rating	alt_rating	stars
GREEN NEW AMERICAN VEGETARIAN	absolutely love this place i ve been here man...	green	3.0	0.0	2	0.0	4.500000
SKINNYFATS	i was excited to try this place when my hubby...	green	3.0	0.0	2	0.0	4.500000
STARBUCKS	this is a cool starbucks great atmosphere ton...	not_green	0.0	0.0	0	0.0	3.257036
WALGREENS	stars for pleasant employees i give it more b...	green	3.0	0.0	2	0.0	2.754266
WHOLE FOODS MARKET	came to this whole foods in search of hot sau...	green	3.0	0.0	2	0.0	3.576923
TRUE FOOD KITCHEN	truly over rated it wasn t bad but it wasn t ...	green	3.0	0.0	2	0.0	4.000000
GREENS AND PROTEINS	i had the green shake and the turkey club wra...	green	3.0	0.0	2	0.0	4.000000
PITA JUNGLE	this pita jungle is located in the food court...	not_green	2.0	0.0	2	0.0	3.843750
FLOWER CHILD	this place is so good little pricey but the f...	green	3.0	0.0	2	0.0	4.400000
SPROUTS FARMERS MARKET	i haven t been in too many sprouts but this h...	green	3.0	0.0	2	0.0	3.894737

VI. Modeling

After we had our dataset finalized, we wanted to group the restaurants into two clusters: green and not green. To do this, we wanted to convert each restaurant row into a tf-idf matrix using the reviews to find our top words and then implement the k-means algorithm on our matrix. We wrote a python script and used pandas and NLTK to create a dataframe out of the business name and text columns from our dataset and performed preprocessing on the reviews such as removing stopwords, removing punctuation and lemmatizing the words. We then transformed the data into a tf-idf matrix and were able to see the top words generated by our models. What we noticed was that a lot of words were similar between the two groups so we decided to look at the actual topics being generated. We visualized the topics by making word clouds and by looking at those we realized that we needed to add more stopwords to our list to try to get more information that would separate the two groups. Below are the words from each group that we found to be the best separators. Although many words are shared between the groups, we noticed

“vegan”, ”green”, ”vegetarian”, ”local” and “veggies” to be included in the green restaurants. We decided to use a different technique in classifying the groups so that the overlap wouldn’t affect our model.

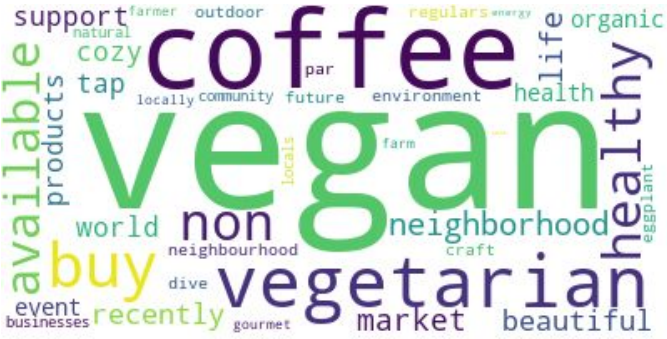
Green

Non-Green

Our next step was to take the aforementioned 5,196-row dataset with 2,598 Green restaurants and 2,598 Not Green restaurant and imported it into RapidMiner. We used the Auto Model feature of the application to run Feature Engineering and Machine Learning Algorithms. The algorithms’ goal was to predict whether a restaurant was green or not green primarily based on the contents of its reviews. Because we had so little overlap between our “green” datasets and our yelp data, many of the restaurants classified as green in this final dataset were chosen because of “green” keywords in their reviews. Given this, we knew that many of the features that the models would use to classify restaurants in the test data would be the appearance of these words, but it was interesting to see which words were most predictive. Displayed in this table are the features of the review text that were most correlated to the “green” or “not green” column.

Attribute	Coefficient ↓	Std. Coefficient
review_text:green	454.237	0.551
review_text:healthy	409.705	0.739
review_text:vegetarian	320.269	0.400
review_text:life	301.105	0.318
review_text:future	238.145	0.220
review_text:locals	229.414	0.230
review_text:greens	227.180	0.293
review_text:environment	213.216	0.211
review_text:eggplant	194.076	0.259
review_text:living	191.378	0.160
review_text:outdoor	188.040	0.223
review_text:community	181.055	0.266
review_text:locally	137.669	0.176
review_text:organic	119.305	0.292
review_text:vegan	111.110	0.305
review_text:natural	105.625	0.161

Here we represent the words above which have the highest Pearson Correlations with the “green” label column using a word cloud. Their size is representative of how often they appear in reviews of the green restaurants.



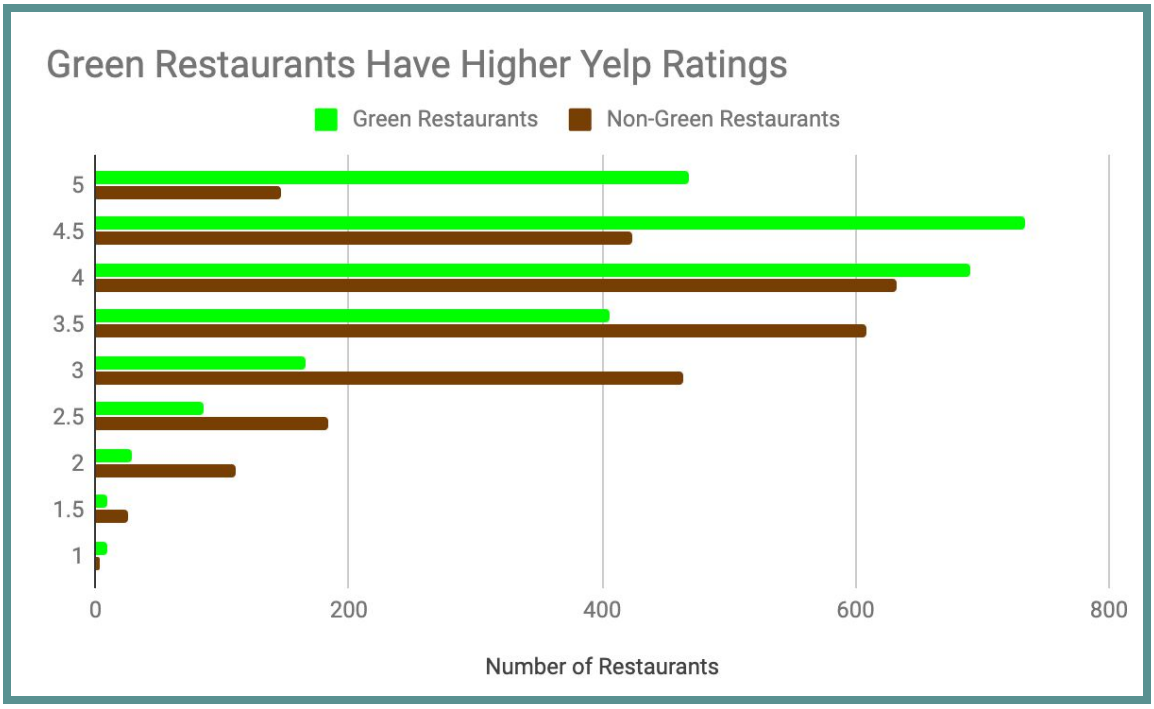
Using these highly-correlated features, a Generalized Linear Model proved to be most accurate from the Auto-Model process on RapidMiner. This can be used to refine our list of green keywords as well as classify new restaurants as green or not green. The following confusion matrix displays its results. Upon improvements of our dataset and classification algorithms, we can use this as a baseline for future evaluation.

accuracy: 78.17% +/- 0.58% (micro average: 78.17%)

	true not_green	true green	class precision
pred. not_green	2479	871	74.00%
pred. green	375	1983	84.10%
class recall	86.86%	69.48%	

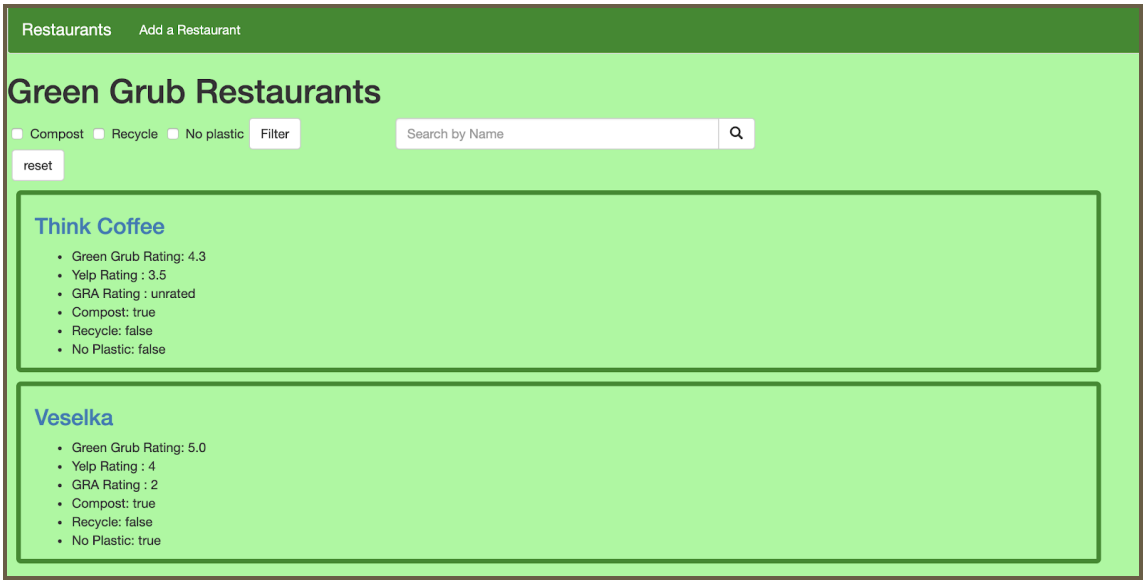
VII. Evaluation & Discussion

Our Hypothesis was that restaurants which are more environmentally friendly are more successful. To test our hypothesis, we decided to use Yelp ratings as a proxy for success. We defined environmentally friendly rather loosely given our constraints. We classified a restaurant as environmentally friendly (or green) if at least 1% of the words in its reviews were words related to being environmentally friendly or if it appeared in one of the lists acquired from the Green Restaurant Association, Seafood Watch, and Plant-Based Restaurant Guide. Using these definitions we found that green restaurants are more successful, proving our hypothesis to be correct.



This graph represents how a vast majority of the restaurants with 5 or 4 star ratings on Yelp are environmentally friendly. The lower rated restaurants also tend to be not green according to our metrics. Consumers who view a restaurant as being environmentally friendly, or at least those who mention it in their reviews, also tend to like the restaurant more. As various researchers have found via smaller scale studies, many consumer's perception of a restaurant improves when the restaurant is more environmentally friendly. We looked at this theory from a data science perspective and found the same result. There is still much to learn about the relationship between taking green initiatives, consumer perceptions and a restaurant's success.

VIII. Deployment and Future Work



Finding out which restaurants were eco-friendly and which were not proved to be difficult. This was troubling both through the perspective of our project and personally, because we want to support restaurants that support the environment. Hence, we created a website to make finding green restaurants easier and more accessible to the public. The website is structured similarly to Yelp, except the reviews are meant to focus on how eco-friendly the restaurant is. It would have been extremely helpful to have a dataset from users of this kind of website for this project. We are eager to continue working on it and launch it so we can use the data we collect on it to add to our database and improve our results. While Yelp is extremely valuable and widely used, it has no specific information about how green restaurants are. Our website acts as a crowd-sourced database in which one can rate and review restaurants that they think are doing great things for the environment, as well as those which have a lot of room for improvement. Our goal is to use this new data to expand our findings as well as to publish our results. We also want to make it easier for anyone interested in looking for green restaurants to be able to find it all in one place. The website right now is the first iteration of the concept. There is a lot of work to be done on both the contents of the site and the user interface. The website is currently deployed through a courant server here[11]. And a video demo can be found here[12].

Because we had so few restaurants in our yelp dataset that were certified by the GRA or Seafood Watch, we were forced to base most of our findings on whether reviews contained “green” words. This approach has some drawbacks. For instance, a review which says, “The **green** wallpaper was nice and the place had a **local** charm” will be analyzed the

[11] http://linserv1.cims.nyu.edu:30053/restaurants
[12] https://drive.google.com/file/d/1zg7izxPZZceq8zSPmiEYwXfMANSDLmwK/view?usp=sharing

same as, “I love that this place is big about going **green** and getting **local** produce”. This realm of problems is often encountered in text-based machine learning and there are many advanced natural language processing techniques that can be explored to improve our classifications. Another way to improve our modeling techniques and better be able to prove our hypothesis would be to find more data about the restaurants that have been certified as good for the environment. If we could get review text from yelp or some other reviewing platform regarding these restaurants, we could learn what features truly indicate an eco-friendly restaurant. Other data about the restaurants and their neighborhoods which can improve our modeling can be found in the various datasets provided by NYC Open Data. We may also be able to extract interesting features by scraping restaurant’s websites and looking for signs of “going green” initiatives. We could also be more systematic in scraping environmentalist blogger’s websites for restaurant recommendations or even partner with the bloggers. Along with these, there are many avenues to be explored and a wealth of potential for improving our models and encouraging more restaurants to go green!

Milestones and Reporting

Milestone	Tasks	Date
1 - Analysis		
1.1	Business Understanding	3/13/19
1.2	Analysis and design stage, gather data and create system mockup & plan	04/10/19
1.2	Design work plan (distribution of tasks to each member of team)	04/23/19
1.3	Present abstract and project plan	5/08/19
2 - Development		
2.1	Create database, multiple steps	05/12/19 (finalized)
2.2	Import existing data	04/09/19
2.3	Clean data (tfidf, lemmatize, confusion matrix)	05/03/19
2.4	Analyze data	05/2019
3 - Testing & Modeling		
3.1	Put final data file through rapid miner)	05/11/19
3.2	Make word clouds, green vs non green	05/12/19
3.3	Get error analysis and analyze accuracy	05/13/19
3.4	Finalise documentation & present results	05/15/19
4 - Deployment		
4.1	Deployed website, environment version of Yelp	05/15/19

11. References

30 Ways (and Days) to a More Sustainable Restaurant. (2008, September 30). Retrieved from https://www.starchefs.com/features/trends/30_sustainability_ tips/index.shtml

Alifierakis, M. (2018, January 07). Using Yelp Data to Predict Restaurant Closure. Retrieved from https://towardsdatascience.com/using-yelp-data-to-predict-restaurant-closure-8aafa4f72ad6

Green Restaurant Association. (n.d.). Retrieved from http://www.dinegreen.com/

Jin, Q. (2015, February 18). A Research Proposal: The Effects of Restaurant Environment on Consumer Behavior. Retrieved from https://scholarsarchive.jwu.edu/cgi/viewcontent.cgi?article=1037&context=mba_student

Plant Based Restaurant Guide to NYC. (n.d.). Retrieved from http://www.model4greenliving.com/plantbased-restaurant-guide-nyc

Restaurant Inspection Information. (n.d.). Retrieved from http://a816-restaurantinspection.nyc.gov/RestaurantInspection/SearchBrowse.do

Seafood Watch - Official Site of the Monterey Bay Aquarium's Sustainable Seafood Program. (n.d.). Retrieved from https://www.seafoodwatch.org/

Yelp Open Dataset. (n.d.). Retrieved from https://www.yelp.com/dataset

Namkung, Young, and SooCheong (Shawn) Jang. *Effects of Restaurant Green Practices on Brand Equity Formation: Do Green Practices Really Matter?* Science Direct, June 2013, www.sciencedirect.com/science/article/pii/S0278431912000928?via%3Dihub.