Conscioux Internet Market Research

Huaiyu Hu Keer Jiang Yuning Xie

Abstract

In this project, our objective is to locate potential markets for Conscioux by taking insight of people's attitude towards veganism in Twitter. To be more precise, we are interested in the supply and demand of vegetarian. The two main datasets we use are the data containing terms "vegan" and "weight loss" scraped from Twitter and public Yelp dataset. To have a general view of how these tweets distribute across the United States, we plot the coordinates of these tweets on Google Map. At the same time, we also visualize the distribution of restaurants that provide vegetarian food by clustering Yelp data. With the preprocessed tweets data, we build several models that classify the sentiment based on one's tweet and choose the optimal one with highest accuracy. To explore other factors that may affect people's attitude towards vegan, significant factors are achieved by logistic regression.

1 Background

Veganism was created to describe the group that hold the belief of abstaining from the use of animal products, particularly in diet. The follower of this kind of diet or philosophy is known as a vegan. In these years, the concept of vegan lifestyle is favored across the country and the number of vegan and vegetarian restaurants have ballooned. Meanwhile, healthy balance attracts attention of more Americans, especially in the aspects of carbohydrates, fats, and proteins considering body status. Our corporation, Conscioux, is aim to provide a one-stop shop for plant-based health solutions for weight loss, muscle gain, and nutritional guidance. Customers are easily connected to programs and professionals that share their healthy living values. Furthermore, Conscioux serves everyone who commits themselves to the world by reducing their carbon footprint while working on improving their health.

2 Data

2.1 Data Collection

There are three datasets analyzed in this project and the first one is scraped from Twitter. Choosing "vegan" and "weight loss" as our keyword, we collect tweets as well as other

4 features that tags, coordinates (longitude & latitude), favorite count and country code. Original observations are more than 100 thousand but most of tweets are lack of geographical records considering privacy. It is necessary for us to analyze coordinates, so the number of observations after filtering out is 3574. In terms of consumers' attitude towards vegan, we use natural language processing and sentiment analysis to quantify their reviews. One new categorical feature is created to describe their attitude with "1", "0" and "-1" which means positive, neutral and negative attitude respectively.

The second dataset is from Yelp including business set and review set and in business set, there are 188593 original observations including 26 states. For the same reason as before, we need to find out where the vegan regions are located in and the characteristics of each cluster of these vegan regions. After we pick "vegan" and "vegetarian" as key words in the categories of these restaurants, the new data only involves 8 states which are Arizona, Illinois, North Carolina, Nevada, Ohio, Pennsylvania, South Carolina and Wisconsin. The number of observations after filtering out is 829. Due to the limitation of data, the results can only partly demonstrate the distribution of vegan regions in United States.

The last dataset that used to perform logistic regression is collected and summarized from three different websites. Firstly, the features in population, age, education and salary of each state are from Statistical Abstract of the United States published by *United States Census Bureau*, which is the authoritative and comprehensive summary of statistics on the social, political, and economic organization of the United States. Then, to describe features in aspects of fit, we use adult obesity rate and physical inactivity rate of each state from *The state of Obesity*. This organization creates annual report that provides the latest data on obesity as well as related health conditions every year. There is another column describing number of vegan restaurants in each state collected from *Happy Cow*, which is an open service for people to find plant-based or vegan options all over the world.

2.2 Data Visualization

Given coordinates of tweets, we can initially draw the left map, which shows the location distribution of all the tweets that contain the terms "vegan" and "weight loss". It is clear to see that the targeted tweets are concentrated in Great New York Area, Los Angeles, San Francisco, Houston and Orlando. It is reasonable that metropolis with larger population density focus more on plant-based diet and healthier lifestyle. Thus we recommend Conscioux to advertise more and help potential customer in these cities to switch to vegan lifestyle.

Before visualizing the attitudes distribution across the United States, three percentage of

tweets related to different perspectives are calculated. The percentage of positive, neutral and negative towards vegan are 54.34%, 34.61% and 11.05%. In total, the reaction is satisfied and the space for Conscioux to develop their consumers is considerable. From the right map, there are some districts that discuss about vegan a lot with many positive/ neutral attitudes and they could be the targets of Conscioux. For example, Chicago and Florida are good examples of under-service areas.

Figure 1: *Vegan Tweets Distribution*

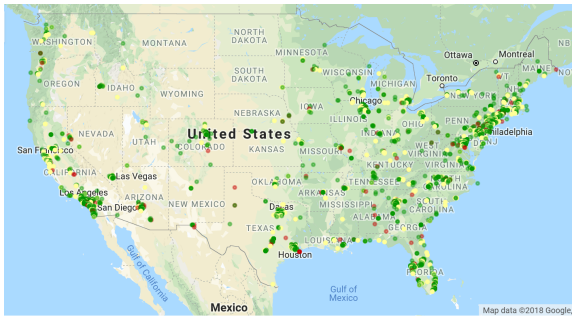

Figure 2: *Distribution with Attitudes*

2.3 Data Cleaning

2.3.1 Tweets Preprocessing

Since we are interested in people's attitude towards vegan lifestyle in Twitter, it is necessary for us to extract useful information from their tweets, tags etc. Therefore, the preprocessing of the tweets data is an essential step as it makes the raw text ready for following analysis, i.e. it becomes easier to apply machine learning algorithms to it. The main goal of this step is to clean up the noise that are less relevant to what we are trying to looking for in millions of tweets such as punctuation, special characters, numbers and terms that do not carry much weights in context to the tweets. If we preprocess the dataset well, then we would be able to get a better quality feature space.

What we do in the text cleaning are listed as below:

- Decode HTML to general text and get rid of URL links.
- The tweets handles are masked as "@user" due to user privacy concerns. Therefore, the tweets handles barely provide any relevant information about the nature of tweets for us.
- It is also necessary to get rid of the punctuations, numbers and even special characters since they wouldn't be helpful in differentiating different types of tweets.

- The text used with hashtags can provide useful information about the tweets. It might be risky to get rid of all the text together with the hashtags. So we leave the text intact and just remove the "#".
- Once we have executed the above steps, we split every tweet into single words or tokens which is an essential step in NLP step.
- During the cleaning process, negation words are split into two parts, and the 't' after the apostrophe vanishes in filtering tokens with more than one syllable. This makes words like "can't" end us as same as "can". We fix this by manually converting these negation words to two separated words in the cleaning function.

2.3.2 Text Visualization

The first text visualization is controversial word cloud. A word cloud represents word usage in a document by resizing individual words proportionally to its frequency, and then presenting them in random arrangement. The word clouds of negative and positive sentiment are shown below:

Figure 3: *Negative Sentiment Word Cloud*

Figure 4: *Positive Sentiment Word Cloud*

It is reasonable that words such as "gluten free", "fitness", juice" appear a lot in positive word cloud, and words such as "chicken", "butter" are frequent words in negative word cloud.

Another interesting part in Twitter that reflects sentiment is hashtag. Hashtags in Twitters are synonymous with the ongoing trends on Twitter at any particular point in time and they also contain much valuable information. Therefore, for our dataset, we try to check whether they help us distinguish tweets into different sentiments. As expected, most of the terms are relevant to healthy lifestyle.

To implement other data visualization in the following step, we use the count vectorizer to calculate the term frequencies. In order to get a full picture of the dataset, we implement it with stop words first and do not limit the maximum number of terms. Based on the term frequency data frame we know that the high-frequency stop words (such as "to", "the", "for", "of" etc.) are roughly same frequent in positive, negative and neutral classes, so we remove these stop words to get a more meaningful dataset. Now we have the frequency

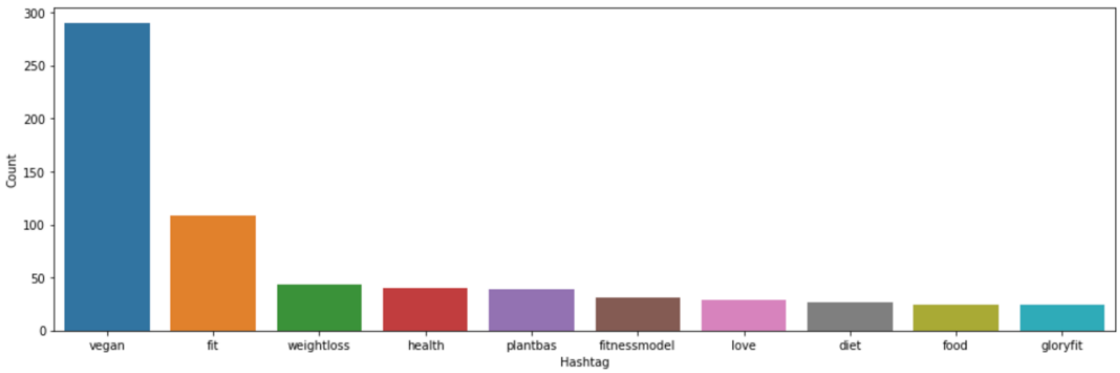

Figure 5: *Histogram Density of Hashtags in Twitters*

data without stop words and it is shown in Figure 6. Moreover, here we also draw bar plot for top 30 words in positive tweets that is shown in Figure 7, this may provide appropriate keywords for Conscioux to mention in their advertisement:

	negative	positive	neutral	total
vegan	70	360	312	742
thanksgiving	20	17	165	202
love	16	62	57	135
fitness	19	55	58	132
free	14	78	40	132
happy	11	28	85	124
diet	12	54	51	117
new	8	55	51	114
day	9	26	51	86
healthy	10	46	19	75

Figure 6: *Word Frequency Table*

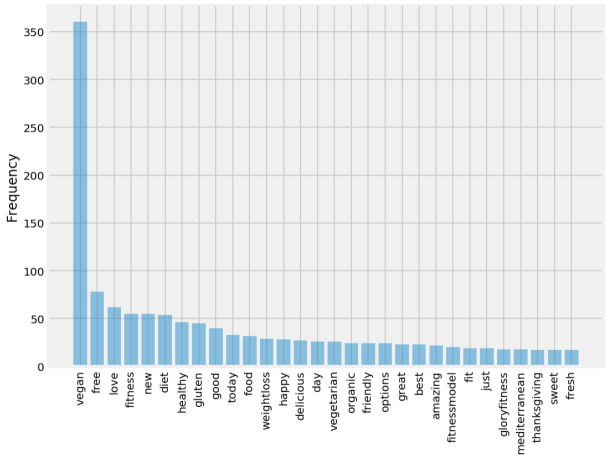

Figure 7: *Top 30 Tokens in Positive Tweets*

Unfortunately, words frequency could only provide partial information about the importance of what people say in tweets. To explore them deeper, we come up with a way that is able to quantify how important a word is among all the tweets in that class. Intuitively, if a word appears more often in one class compared to another, then this can be a measure of how much the word is meaningful to characterize the class. The positive/negative/neutral rate is defined in the way below (in the following parts we just take positive class as an example):

$$positive\ rate = \frac{positive\ frequency}{positive\ frequency + neutral\ frequency + negative\ frequency}$$

The problem of this measure is that words with highest positive rate have zero frequency in the negative and neutral tweets, but overall frequency of these words are too low to be convincible. Therefore, we consider another metric that is the frequency a word occurs in

the class, which is defined as below:

$$positive\ frequency\ precentage = \frac{positive\ frequency}{\sum positive\ frequency}$$

Furthermore, we want to come up with a metric that reflects both information in these two metrics. Firstly, the CDF values of both positive rate or positive frequency percentage are calculated. By calculating the CDF values, we can see where the value of either positive rate or positive frequency percentage lies in the distribution in terms of cumulative manner. For example, taking "protein" as an example, there are around 86.94% of the tokens taking a positive rate value less than or equal to 0.8667, and 73.27% taking a positive frequency percentage value less than or equal to 0.0029. Then we calculate a harmonic mean of these two CDF values and from where we notice the last column provides a more meaningful measure of how important a word is within the class (negative and neutral classes are not shown here).

	negative	positive	neutral	total	pos_rate	pos_freq_pct	pos_rate_normcdf	pos_freq_pct_normcdf	pos_normcdf_hmean
available	1	16	2	19	0.842105	0.003595	0.851794	0.801265	0.825757
friendly	3	24	7	34	0.705882	0.005392	0.727013	0.925948	0.814509
special	1	17	4	22	0.772727	0.003819	0.793860	0.821517	0.807452
protein	1	13	1	15	0.866667	0.002921	0.869399	0.732736	0.795239
check	0	13	2	15	0.866667	0.002921	0.869399	0.732736	0.795239
sunday	1	14	2	17	0.823529	0.003145	0.837469	0.756824	0.795107
mediterranean	2	18	5	25	0.720000	0.004044	0.741973	0.840418	0.788133
organic	2	24	10	36	0.666667	0.005392	0.683331	0.925948	0.786351
options	3	24	9	36	0.666667	0.005392	0.683331	0.925948	0.786351
good	7	40	16	63	0.634921	0.008987	0.645951	0.995933	0.783641

Figure 8: *Table of Positive Frequency*

2.4 Feature Extraction

2.4.1 Count Vectorizer

Count Vectorizer is one approach to extract features from words by tokenizing, counting and normalizing. Firstly, all the reviews are collected and transformed into the bag of words model, which ignores grammar and order of words. After that, the frequency of each vocabulary will be summarized and normalized, so this converts them to a numerical presentation.

In this section, we want to explore the necessity of removing stop words in this step. It is also assumed that removing stop words is a necessary step, and will improve the model performance. Moreover, we also define our custom stop words list. So here we run the same test with, without stop words and without custom stop words then compared the result. The model we choose to evaluate different count vectors is logistic regression. In this project, we also extend the bag-of-word to trigrams. N-grams are simply all the combinations of

adjacent words or letters of length n that can be found in the text. In the plot as below, we compare the performance of unigram, bigram and trigram. Based on figure 10, the bigram and trigram perform better in smaller vocabulary and the unigram performs comparably better in larger size of vocabulary.

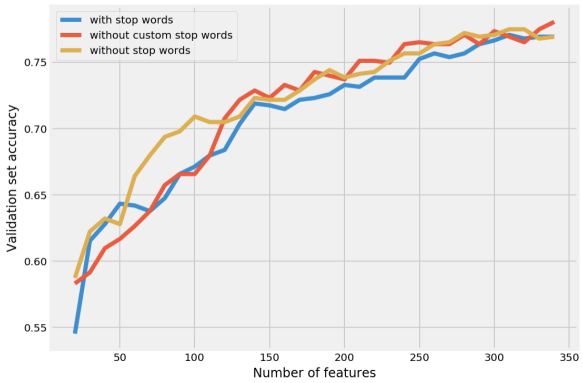

Figure 9: *Accuracy with/without Stopwords*

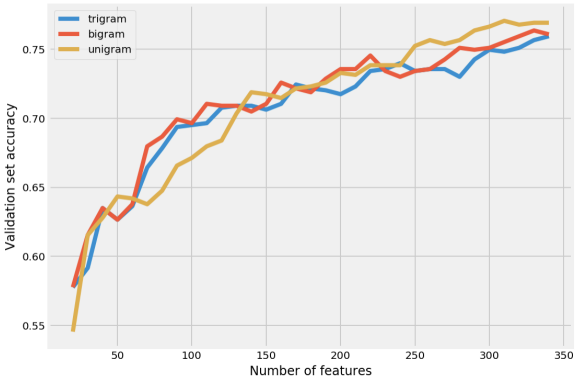

Figure 10: *Accuracy of N-gram Test Result*

In order to evaluate the classification performance of a model, there are many different metrics that can be used. Here we create the confusion matrix and the accuracy is defined as:

$$accuracy = \frac{TruePositve + TrueNegative}{TruePositve + FalsePositve + TrueNegative + FalseNegative}$$

When the distribution of classes is well balanced, accuracy is able to give you a good picture of how the model is performing. However, if one of the class is dominant in the dataset, then accuracy might not be good enough to evaluate the model. To explore further into the confusion matrix, we adopt the precision and recall where precision tells us the proportion of True Positive in the set of all positive predicted data, and recall is the proportion of data that actually is positive are predicted positive:

$$Precision = \frac{TruePositive}{TruePosive + FalsePositive}$$

$$Recall = \frac{TruePositive}{TruePostive + FalseNegative}$$

Finally, F1 score us the harmonic mean of precision and recall. So by calculating the harmonic mean of the two metrics, we can get how the model is performing both in terms of precision and recall:

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall}$$

Because 52.66% of tweets in our dataset are positive, 36.26% are neutral and only 11.08% are negative, then in our case neutral takes the place of "negative" in the definitions above. By analyzing in this way, we build the model which targeting the group of people who loves

vegan lifestyle and the potential customers who might be interested in vegan lifestyle in the future. In the classification below that the model has slightly higher precision in neutral class and higher recall in positive class, but this averages out by calculating the F1 score.

```
null accuracy: 63.92%
accuracy score: 74.97%
model is 11.05% more accurate than null accuracy
--------------------------------------------------------
Confusion Matrix

          predicted_neutral  predicted_positve
neutral                 224                 32
positive                 80                292
--------------------------------------------------------
Classification Report

             precision    recall  f1-score   support

    neutral       0.71      0.25      0.37        79
   positive       0.67      0.87      0.76       258

avg / total       0.76      0.75      0.74       715
```

Figure 11: *Classification Results*

2.4.2 Tf-idf Vectorizer

When analyzing text data, it is necessary to transform them into floating point values, but only counting amount of words is not enough because we need to explain how important a word is to a document in a corpus. Therefore, term frequency–inverse document frequency (tf-idf), as one of the most popular term-weighting schemes in text mining, is introduced. The main idea of tf-idf is that word with high-frequency in all the corpus is less important than that occurs frequently in only several documents. Term frequency tf(t,d) is count of each term t occurs in document d and inverse document frequency idf(t,D) measures how much information the word involves. The formula of tf-idf is:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D).$$

Once we instantiate Tfidf vecorizer and fit the transformed data to logistic regression, and then check the validation accuracy for a different number of features. Based on Figure 10, we can see that Tfidf yields better result than count vectorizer in unigram, bigram and trigram.

3 Clustering

3.1 Methodology of Clustering

Clustering algorithm is one method to partition groups with similar characteristics and assign them into clusters. After fixing clusters, they are interpreted by assigning a label or

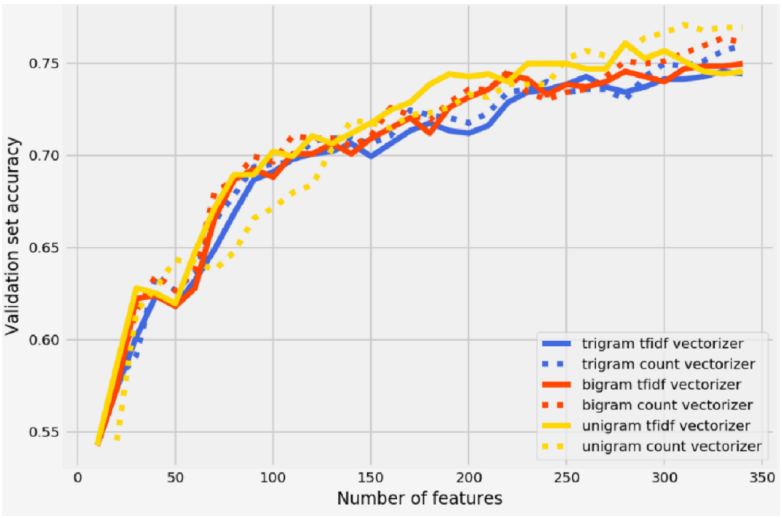

Figure 12: *Comparison between Tf-idf & Count*

category to each cluster. There are a large number of rules to perform clustering, and the most popular algorithms are connectivity models, centroid models, distribution models and density models. In our project, we will use one of centroid models, k-means ++, to employ clustering.

K-means++ is improvement of k-means approach, while k-means clustering is an iterative refinement technique to segregate n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Given initial random k means, all the sample points will be assigned to nearest mean to minimize summation of Euclidean distance. The second step is to update the cluster centers to the average of its assigned points. Repeat the steps as above until the assignments no longer change, then we will achieve final approximation of clusters. However, the main shortcomings are that the result is not guaranteed to be optimal and various choices of initial means will lead to totally different clusters.

Thus, here we use k-means++ that addresses obstacles mentioned above by specifying seeding method to find better initial cluster centroids. Firstly, randomly choose one data point as the first cluster center and then calculate distance between each observation and this center point. After that, each subsequent new center is chosen from the remaining data points with probability proportional to its squared shortest distance. When k means are chosen, the next procedure is the same as the standard k-means optimization iterations. This modification is much more competitive than k-means by achieving a lower potential value as well as shorting running time.

3.2 Yelp Dataset Clustering

In order to find the location of vegan regions in United States and characteristics of each region, we apply clustering on Yelp dataset. It is obvious that these regions are characterized by the physical proximity and similarity of restaurants. As a result, we use longitude and latitude to cluster for proximity, and categories to cluster for similarity. The subset including top 15 most popular categories of restaurants except "food" and "restaurants" is created to establish the similarity matrix, which is a location matrix with two columns and a categories matrix with fifteen columns. The spatial coordinates and restaurant categories have different units of scale, so it is essential to normalize the data separately first and then give them a proper weight to combine these two matrix together. With the scaled dataset, we apply k -means++ method to find clusters. In order to determine the number of clusters, error and silhouette coefficient against the number of clusters are calculated and plotted as Figure 13.

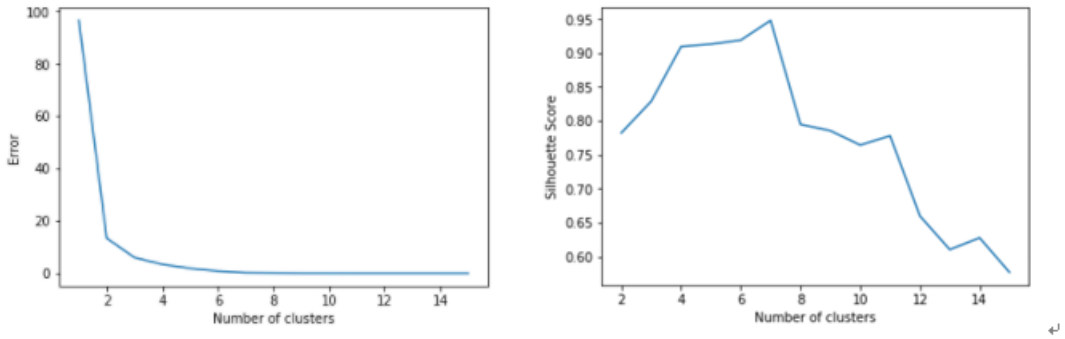

Figure 13: *Choosing Number of Cluster by Error & Silhouette Coefficient*

Lower value of error and a higher value of silhouette score indicate a model with "better defined" clusters, so the value of 7 is chosen as the number of best clusters. To find top three labels of each cluster, we count the categories in each cluster and then pick top three labels. We also plot a heat map with Google map to see coordinates more directly as Figure 14.

From Figure 14, we can find eight vegan regions whose centers are Las Vegas, Phoenix, Cleveland, Pittsburgh, Charlotte, Madison, Champaign and Philadelphia separately. However, using k-means++ method, we only have seven clusters. Due to a small amount of vegan restaurants in Philadelphia, the results of our clustering are consistent with the reality. Combining these results together, the conclusion table is shown as below. Besides, we also cluster the restaurants using the text of the restaurant reviews in an unsupervised fashion.Therefore, we use the document-term matrix approach and tf-idf method to achieve this goal. The result of having 7 clusters is exactly the same as the result got by business dataset.

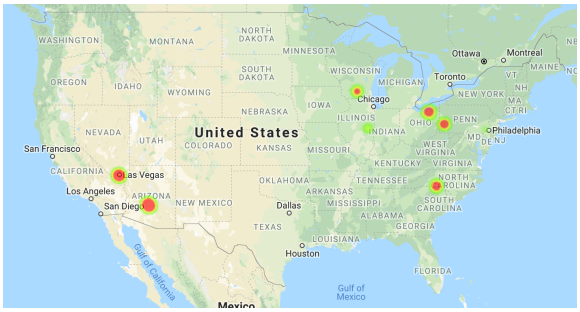

Figure 14: *Heat Map of Yelp*

City	Cluster Characteristics		
Las Vegas	Mediterranean	Mexican	Sandwiches
Phoenix	Mediterranean	Mexican	Indian
Cleveland	Mediterranean	Bars	Mexican
Pittsburgh	Sandwiches	Salad	Gluten-Free
Charlotte	Indian	Sandwiches	Bars
Madison	Sandwiches	Indian	Gluten-Free
Champaign	Mexican	Bars	Vegetarian

Figure 15: *Cluster Characteristics*

4 Logistic Regression Analysis

4.1 Methodology of Logistic Regression

In regression analysis, logistic regression is used to estimate the parameters when dependent variable is binary {0 & 1}. Considering the simple model that involving only one independent variable, we assume that $y_i \sim Bernoulli(\pi_i)$ where $\pi_i = P(y_i = 1)$. The mean responses π_i are now related to a predictor via a logistic function,

$$\pi_i = P(y_i = 1) = \frac{exp\{\beta_0 + \beta_1 x_i\}}{1 + exp\{\beta_0 + \beta_1 x_i\}}.$$

To estimate parameters β_0 and β_1, maximum likelihood estimation (MLE) is used and the formula is as below.

$$D(\beta_0, \beta_1) = 2\sum_{i=1}^{n} y_i log\frac{y_i}{\pi_i(\beta_0, \beta_1)} + (1 - y_i)log\frac{1 - y_i}{1 - \pi_i(\beta_0, \beta_1)}$$

Increasing number of variables will capture underlying structure more accurately, but this will result in over-fitting problem. Therefore, variables are selected by Akaike information criterion (AIC), which adds penalty term related to number of parameters so as to avoid over-fitting.

4.2 Regression on State Dataset

To explore the factors that may affect people's sentiment towards vegan lifestyle for 50 states in the United States, we choose eight features which are total population, sex proportion, median age, the percentage of teenagers, the percentage of people with Bachelor degree or more, average monthly earning, adult obesity rate and number of restaurants in each state. The binary response variable is created by summing up sentiment values in each state and the top 30% of the states are set as 1 and the rest as 0.

After applying AIC algorithm, we build the logistic regression model that obtain five significant variables, which are total population, median age, percent of teenagers, the percentage

of people with Bachelor degree or more and average monthly earning. The coefficients in regression results provide useful information about factors leading to sentiment towards vegan. Only the coefficient of bachelor percentage is negative and all the others are positive, this means that the district with larger population, higher median age, higher percent under 18 years old, lower bachelor percentage and higher salary focuses more on vegan and fitness. Therefore, the areas that satisfy these conditions are reasonable to be targets of Conscioux.

Regression Results

	Dependent variable:
	Attitude to Vegan
Total	0.001***
	(0.0003, 0.001)
Median_age	1.879**
	(0.101, 3.656)
Percent_under_18_years_old	2.058*
	(-0.058, 4.175)
Bachelor_degree_or_more	-0.207*
	(-0.451, 0.037)
Average_Month_Earning	0.002
	(-0.0003, 0.003)
Constant	-124.459**
	(-243.125, -5.793)
Observations	51
Log Likelihood	-15.744
Akaike Inf. Crit.	43.487
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 16: *Summary Table of Logistic Regression*

5 Conclusion

Based on analysis of visualization, clustering and regression, there are several suggestions given for Conscioux when picking up potential markets.

- Potential customers are mainly located by the yellow points in the Figure 2, thus Conscioux should focus more on metropolis like New York City, Los Angeles, Houston and Orlando.
- Given the specific tweet of one user, his/her sentiment could be estimated using the model built in the section of feature extraction.
- Based on the cluster characteristics and hash tags, related keywords should be used when advertising, e.g. gluten-free for Pittsburgh.
- According to the results of logistic regression, districts with higher population, higher median age, higher teenager percentage, lower percentage of bachelor degree or more and higher average monthly earning should be aimed as targets.

6 Future Work

In our project, there are some extensions for our work as below. Firstly, the power of our laptops and twitter privacy restricts the number of sample size in analyzing. For the similar reason, the free Yelp dataset only provides part of restaurants in 26 states. Given full data, more features will be extracted and the results will be more convincing. In addition, it would be great to recognize pattern with k-nearest neighbors (KNN) algorithm and this will detect whether the location is good to develop business for Conscioux.