CKME 136 – Capstone Project

# Introduction

It can be very difficult to find Vegan restaurant options in any city. For my Capstone project I would like to do a twitter sentiment analysis on tweets and their respective geotags to find the most vegan-friendly cities in the world.

# Literature Review

**https://github.com/asmitapoddar/Twitter-Sentiment-Analysis-R**

The literature that I have reviewed above outlines procedures involved with doing a twitter sentiment analysis. Although the topic was very different (mood of top trending tweets), I believe the procedure would be very similar. The authors of this experiment extracted tweets from twitter, cleaned the data, created a database, applied the naïve bayes algorithm, then calculated the percentage of positive vs negative tweets.

**http://analyzecore.com/2014/04/28/twitter-sentiment-analysis/**

The source above outlines the procedure to analyze Twitter posts in R by calculating the number of positive, negative and neutral words and plotting the data into a chart. The source code in the above article will be helpful in guiding me through this procedure.

# Dataset

I will be using the TwittR package in R to extract tweets on positive or negative experiences that vegans had in large cities using geotags. Once extracting the tweets, I will clean the data by removing any emoticons, extra punctuation, stop words, and URLs, then create a database.

# Approach

## Step 1: Extraction of Tweets

1. Create twitter application
2. twitteR - Provides an interface to the Twitter web API
3. ROAuth - R Interface For OAuth
4. Create twitter authenticated credential object

## Step 2: Clean Tweets & Create and load database

The tweets are cleaned in R by removing extra punctuation, stop words, redundant blank spaces, emoticons, and URLs. Once the tweets are cleaned, I can create a database by using term frequency to determine the Features or Labels I will use to classify if the tweet is positive (there are vegan options), negative (there are no vegan options), or neutral (there are some vegan options).

**Step 3. Algorithms**

**Lexical Analysis:​** ​By comparing uni-grams to the word database, the tweet is assigned sentiment score - positive, negative or neutral and overall score is calculated.

**Naive Bayes Machine Learning Algorithm:** ​Training data sets are used to teach the machine what kind of sentences are categorized as positive and what kind are categorized as negative

**Calculate Percentages**: Determine which city is the most vegan friendly by calculating and ranking the cities that have the most positive tweets.