# Social Media Analytics: Homework 4

## Xinyan Cai

## Introduction

In this report, the potential trend of food was found by Facebook posts incorporating the frequency of ingredients of food. Therefore, in order to obtain the potential food trend, the frequency of different ingredients words over 60 months is investigated to predict the pattern of certain words.

## Data

There is a big dataset includes a txt file of ingredient words and Facebook post in every year from 2011 to 2015, 5 years in total.

## Method

Firstly, I used ingredient.txt as the target list to extract various words inside the data (Facebook posts) and integrated them into a data frame.

```
setwd("C:/Users/XINYAN CAI/Desktop/data")
mydic <- tolower(scan('ingredients.txt', character(), quote = "", sep = "\n"))

docs <- Corpus(DirSource(c("C:/Users/XINYAN CAI/Desktop/data/fb2011",
                           "C:/Users/XINYAN CAI/Desktop/data/fb2012",
                           "C:/Users/XINYAN CAI/Desktop/data/fb2013",
                           "C:/Users/XINYAN CAI/Desktop/data/fb2014",
                           "C:/Users/XINYAN CAI/Desktop/data/fb2015")))
dtm <- DocumentTermMatrix(docs, control=list(dictionary = mydic, tolower=T, removePunctuation=T, removeNumbers=T,
                                             stripWhitespace=T, stopwords=c(stopwords("english"), stopwords("spanish")
dtm = removeSparseTerms(dtm,0.996)
m <- as.matrix(dtm)
df = data.frame(m)
```

Secondly, the names of months and years are obtained from the row names of the data frame. Thus, I could utilize them to make the aggregation function, which was used to find the frequency of different words.

```
#Frequency of different words
data <- data.frame(year=as.numeric(substr(rownames(df),7,10)),
                   month=as.numeric(substr(rownames(df),12,13)),
                   stringsAsFactors=FALSE);
da <- cbind(df, data)
da <- aggregate(x = da, by = list(da$month, da$year),FUN = sort)

# Construct plots of different words
Word.search <- function(i) {
  i<- tolower(i)
  plot(da[ ,i], type='l',xlab = 'Month',ylab = 'Frequency',
       main = c('Number of', i, 'in 60 months'))
}
```

Finally, a function was constructed to plot time-series graphs of different words.

```
# Construct plots of different words
Word.search <- function(i) {
  i<- tolower(i)
  plot(da[ ,i], type='l',xlab = 'Month',ylab = 'Frequency',
       main = c('Number of', i, 'in 60 months'))
}

# Cauliflower Rice
Word.search('Cauliflower')
Word.search('Rice')

# Vegetables Noodles
Word.search('Vegetables')
Word.search('Noodles')

# Pumpkin Pie
Word.search('Pumpkin')
Word.search('Pie')
```

# Validation

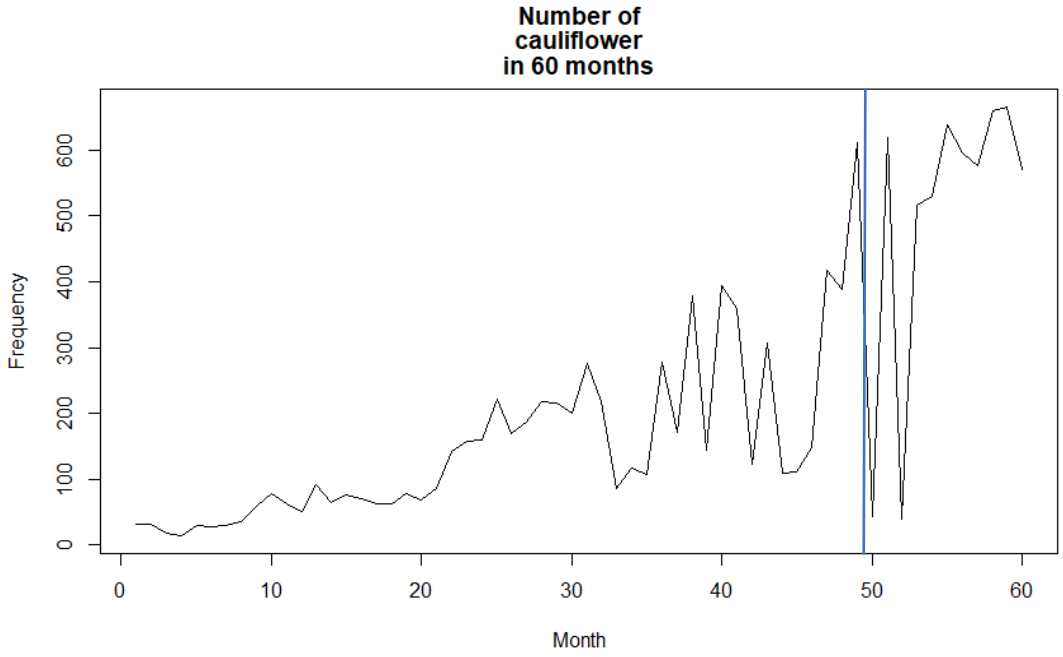
The trends of Cauliflower Rice, Vegetables Noodles and Pumpkin Pie were used to test the effectiveness of a method. It is worth mentioning that here I plotted individual words such as 'Cauliflower' and 'Rice' firstly, and compare plots of them to find a common trend. Hence, I could discover the trends of two words or more together.

## Cauliflower Rice

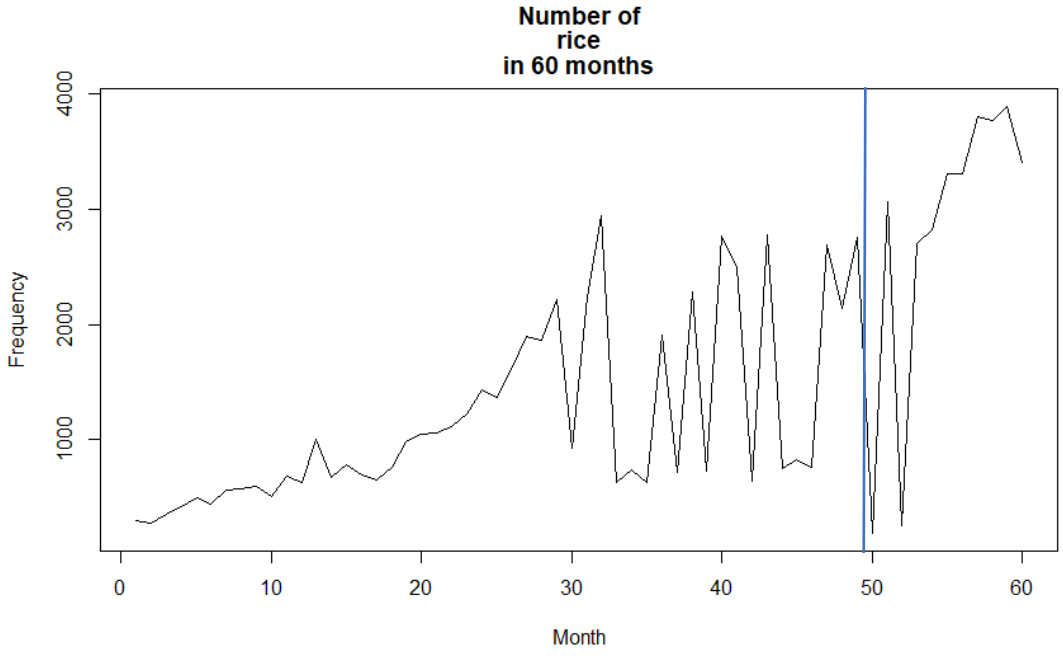
In general, the demand for cauliflower rice has been growing steadily over the past few years largely due to consumers' desire for a healthy alternative to white rice and gluten-filled grains. Therefore, both words showed upward trends. They also revealed

a similar trend after 2015 because of the expanded the cauliflower line. Consumers are more likely to get cauliflower rice so that they will mention more times about cauliflower rice on Facebook.

## Figure 1:



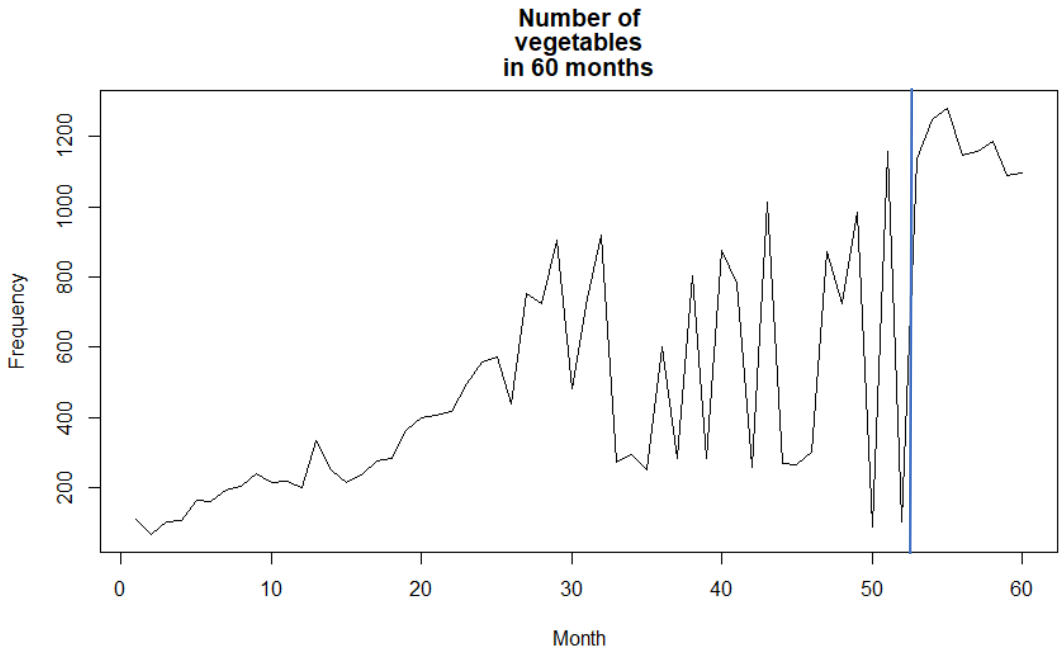

## Figure 2:



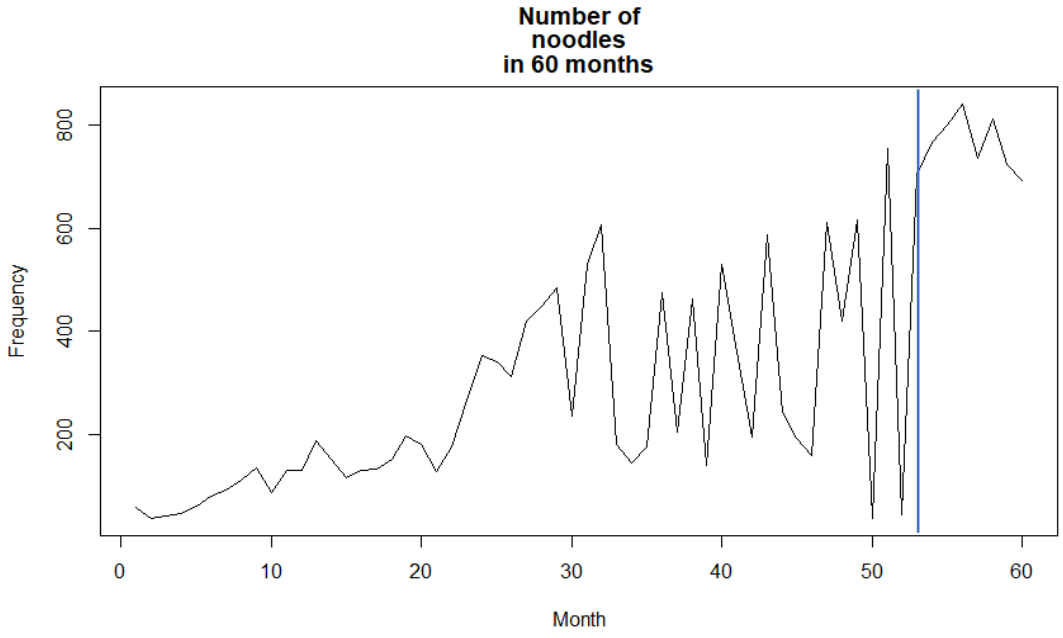
## Vegetables Noodles

In January 2015, Vogue indicated that vegetable noodle was good for people so that there were large increases of Facebook posts of both words of 'vegetables' and 'noodles'.

## Figure 3:



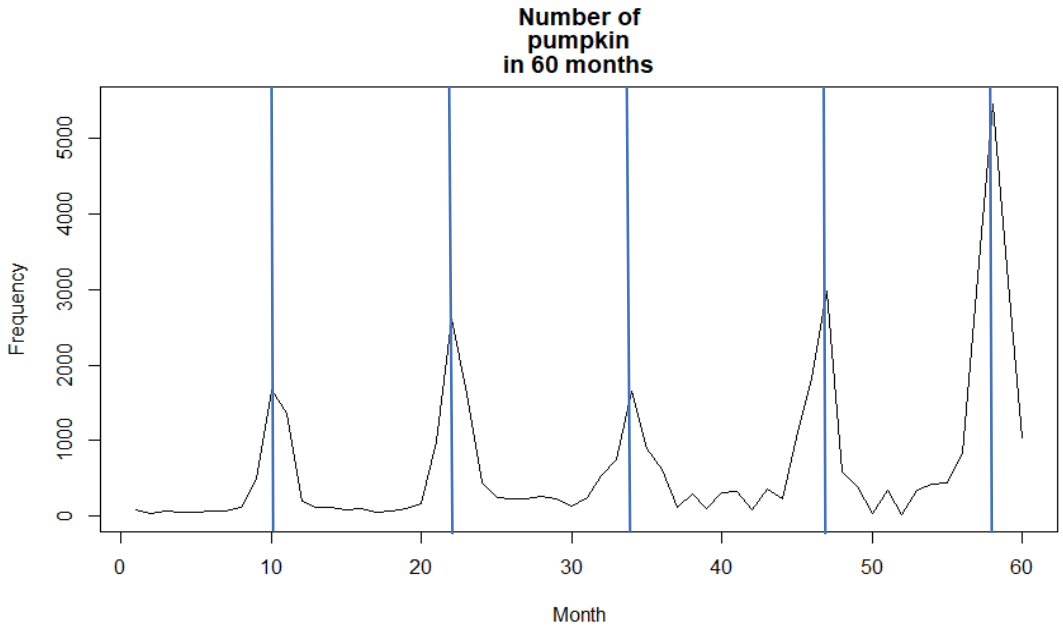

## Figure 4:



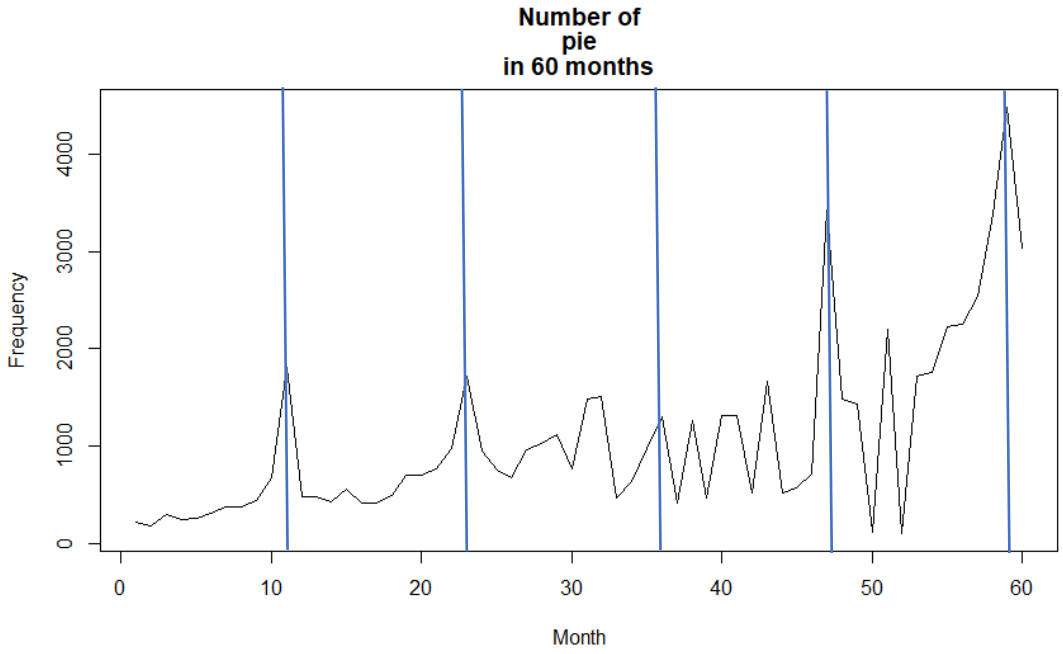
## Pumpkin Pie

During Thanksgiving, the pumpkin and pie occur together to a large extent. The trend was much more obvious for pumpkin might because it is a special word that is highly used in Thanksgiving Day.

## Figure 5:




## Figure 6:



## Further Improvement

Although the figures can accurately show the trend of different words in general. There are still a large number of drops or spikes that are hard to be explained. Thus, it means that the method was not precise enough. Next time, the co-occurrence matrix can be tried to test whether a better result can be derived through this method.