Xinyu Huang

December 16, 2019

Slow Trend Report

Overview

I want to detect the emerging trends of food consumption from social media data.

I will work on over 4 million Facebook posts from 2011 to 2015. The data was stored in multiples files. Each file contains 12 monthly textual content of all posts for that year.

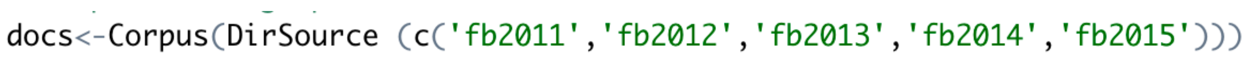
I will first construct a timeseries of potential food trends. Then I will use visual inspection to detect the changes in those timeseries.

Potential Approaches

There are several potential approaches to detect the emerging food trends such as topic modeling, word or term frequency. I will use word frequency in this project.

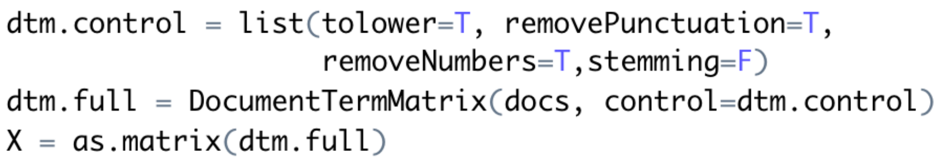
Method Used

I build a document term matrix to solve the problem.

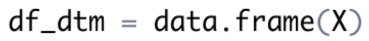


```
docs<-Corpus(DirSource (c('fb2011','fb2012','fb2013','fb2014','fb2015')))
```

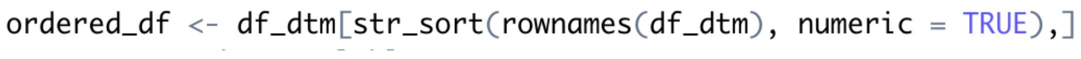
First, I set up a corpus to read in all the data.



```
dtm.control = list(tolower=T, removePunctuation=T,
                   removeNumbers=T,stemming=F)
dtm.full = DocumentTermMatrix(docs, control=dtm.control)
X = as.matrix(dtm.full)
```



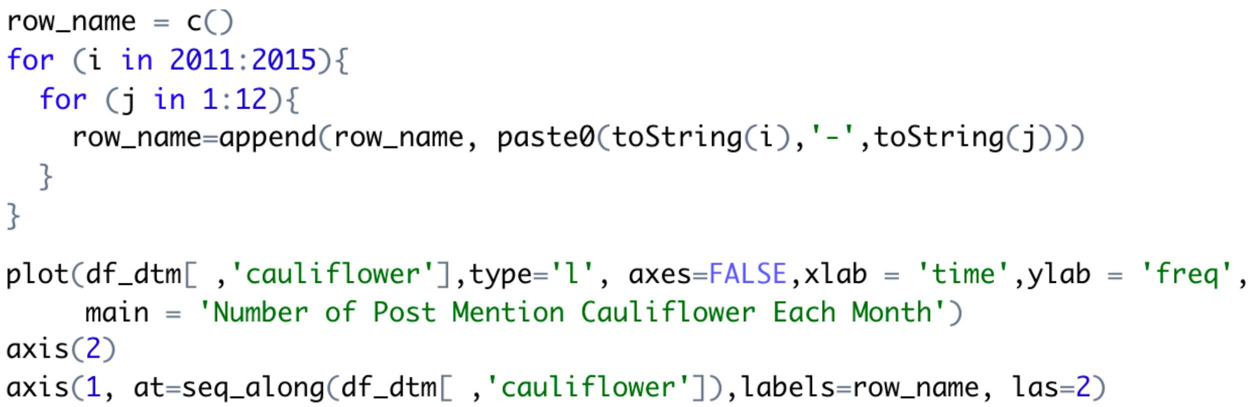
```
df_dtm = data.frame(X)
```



```
ordered_df <- df_dtm[str_sort(rownames(df_dtm), numeric = TRUE),]
```

Then I build a document term matrix that count the frequency of terms occur in the corpus. The rows represent the documents which is the timestamp (month and year) and the columns represent terms. I transfer the document term matrix into a data frame and order the data frame in the correct time sequence for future use.

The timeseries of potential foods trends are now completed. Next, I am going to use visual inspection to detect the changes in those timeseries. I want to plot the tends of number of post mention certain food each month. I am going to use Cauliflower for illustration.



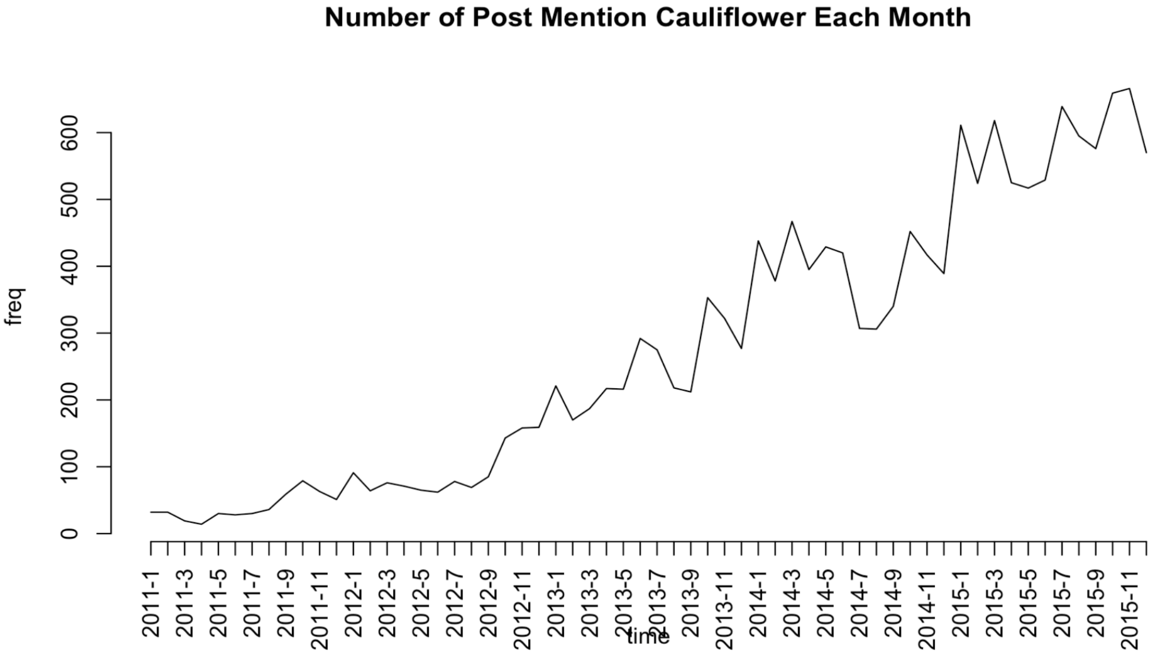
```
row_name = c()
for (i in 2011:2015){
  for (j in 1:12){
    row_name=append(row_name, paste0(toString(i),'-',toString(j)))
  }
}

plot(df_dtm[ ,'cauliflower'],type='l', axes=FALSE,xlab = 'time',ylab = 'freq',
     main = 'Number of Post Mention Cauliflower Each Month')
axis(2)
axis(1, at=seq_along(df_dtm[ ,'cauliflower']),labels=row_name, las=2)
```

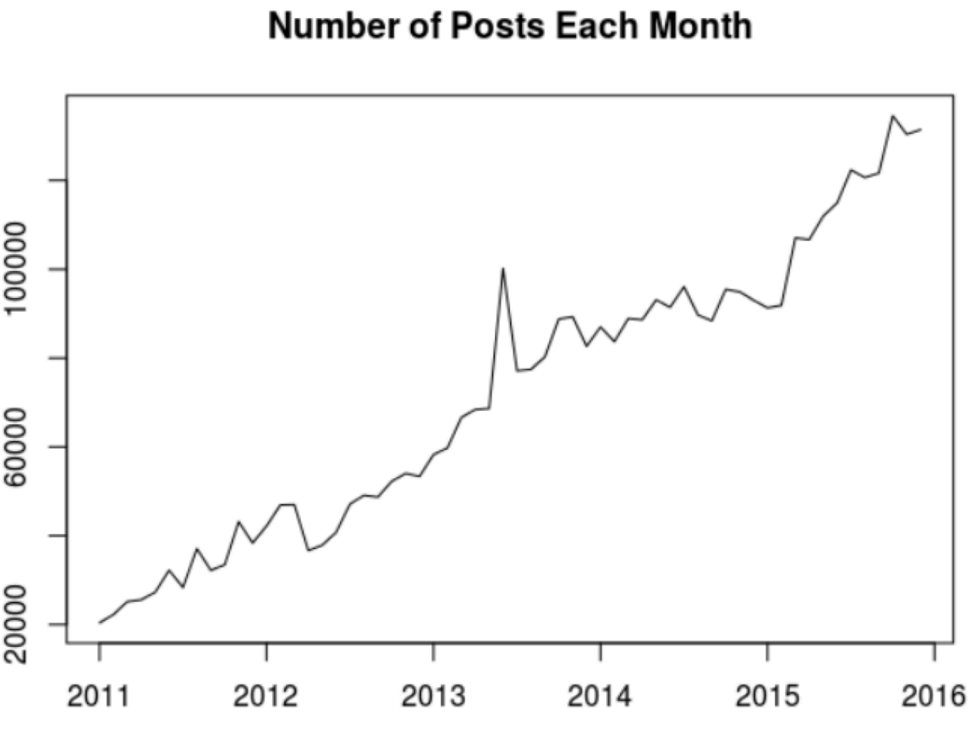
At start, I build a time vector for future use as time label from January 2011 to December 2015. Then I plot the number of post mention Cauliflower each month with respect to the time label I just created. The x axis is the time label. The y axis is the frequency. Column ‘cauliflower’ in the document term matrix data frame represents the frequency in order of time. If I want to plot the trends for any other food, I could just replace the ‘cauliflower’ with that food’s name.

Result Presenting

Below is the plot of change of posts frequency mention cauliflower in time series from 2011 to 2015.



We can observe that the frequency increase from around 30 in January 2011 to more than 600 in December 2015 (almost 20 times). But we need to take the change of number of food-related Facebook posts into account.

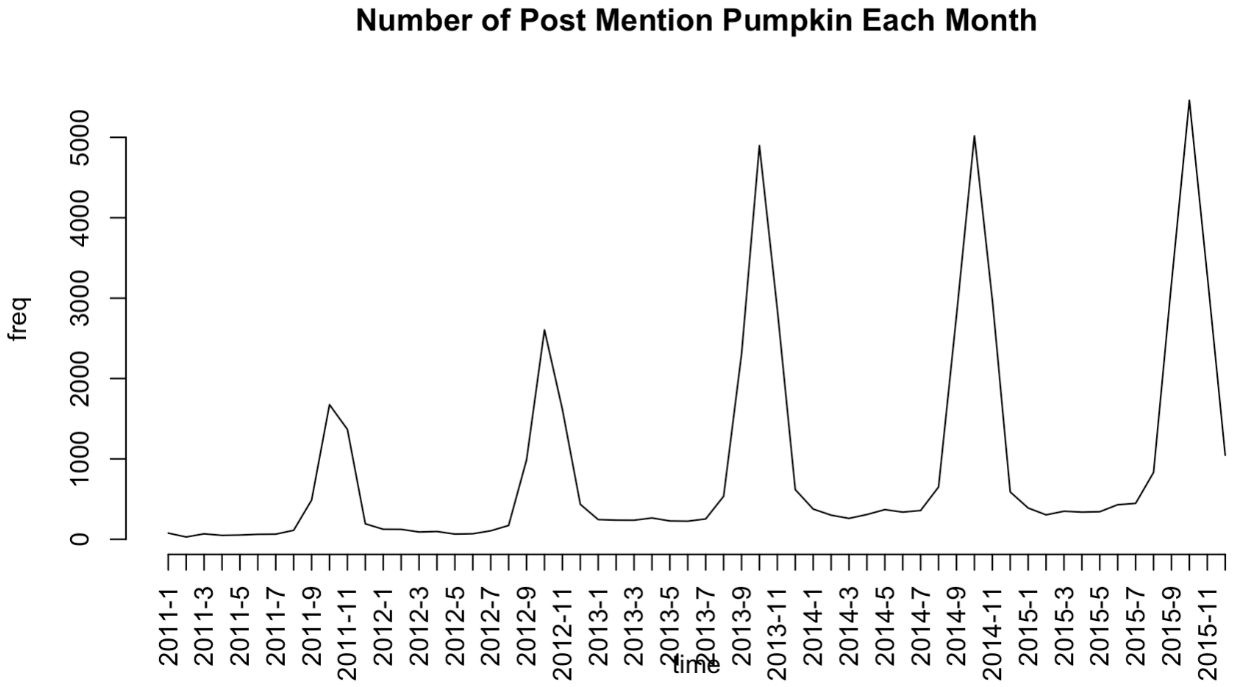


From the plot above, we can observe an increasing trend of food-related Facebook posts. The food-related Facebook posts in the end of 2015 is nearly six times of posts in the beginning of 2011 which is less than 20 times. Therefore, there exist a huge increase of cauliflower mention rates among all food-related Facebook posts.

Validation

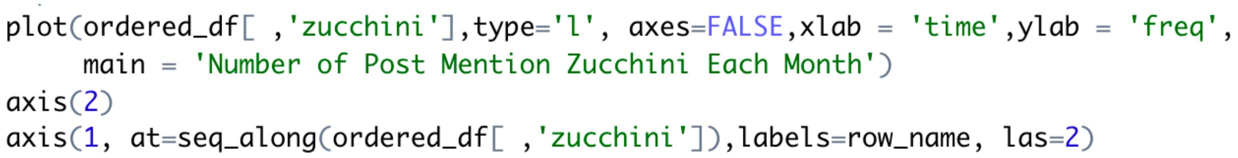
First, I will use seasonally consumed food to validate the method. First, I would like to check for the number of post mention pumpkin each month. If my method can detect a surge in thanksgiving season each year, then it might work.

```
plot(ordered_df[ ,'pumpkin'],type='l', axes=FALSE,xlab = 'time',ylab = 'freq',
     main = 'Number of Post Mention Pumpkin Each Month')
axis(2)
axis(1, at=seq_along(ordered_df[ ,'cauliflower']),labels=row_name, las=2)
```

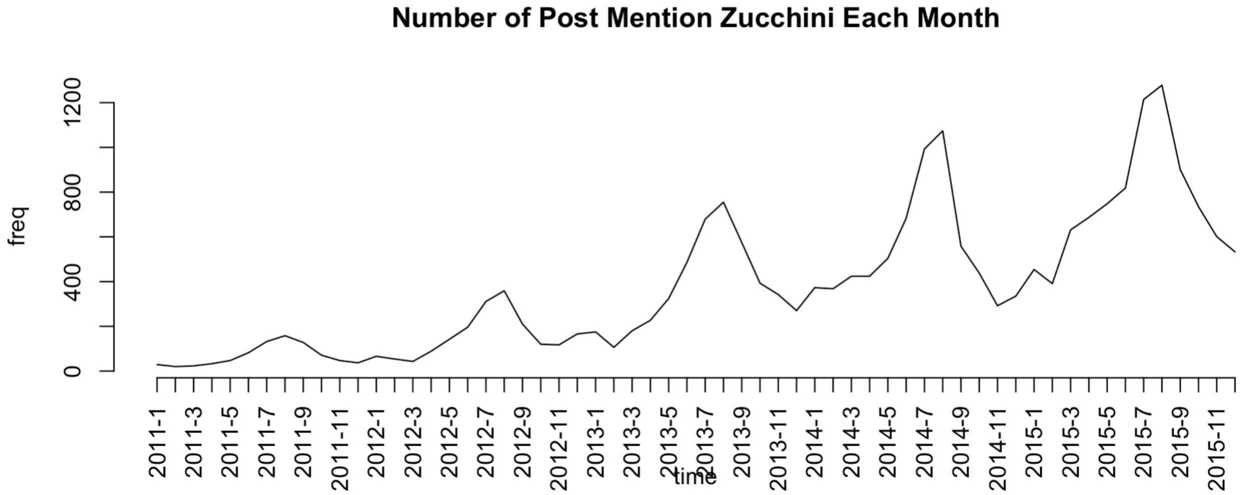


From the plot, we can clearly observe the surge of Facebook post mentioned pumpkin during thanksgiving period every year.

Given the fact that veggie noodle is a recent food trend and Zucchini might be one of the most widely used ingredients of veggie noodles. So next, I would like to check for the number of post mention Zucchini each month.



```
plot(ordered_df[ ,'zucchini'],type='l', axes=FALSE,xlab = 'time',ylab = 'freq',
     main = 'Number of Post Mention Zucchini Each Month')
axis(2)
axis(1, at=seq_along(ordered_df[ ,'zucchini']),labels=row_name, las=2)
```



From the plot, we can clearly observe the increasing trend of Facebook post mentioned Zucchini. Moreover, it shows a seasonal characteristic. There are surges of Facebook post mentioned Zucchini during the summer. One understanding might be people tend to eat more Zucchini during the warmer weather.

Conclusion

I use word frequency method to detect the emerging trends of food consumption from social media data. First, I construct a timeseries of potential food trends. Then I use visual inspection to detect the changes in those timeseries. After the validation using different ground truth, I am confident that this method could reflect some true food trends in reality.