Emerging Trends of Food Consumption Detection

1. Potential Approaches

There are many approaches in social media trend analysis. In this case, the detect the trend, we first need to decompose a Document-Term matrix.

I. Unigram-models

We can use the tm package in r to either construct the top words based on tm or TF-IDF (Term Frequency-Inverse Document Frequency). This approach is easy to interpret, and faster to run, especially on large datasets. But the problem with this approach is the high-dimensional vector space representation, and its learning and estimation has to deal with curse of dimensionality.

II. Latent Semantic Analysis (LSA)

LSA is a method frequently used to analyze relationships between a set of documents and the terms they contain by producing concepts related to the documents and the terms. This can be done with truncated SVD decomposition. LSA can be easily extended and embedded in other more complicated models. But it cannot be generalized to new documents and is prone to overfitting.

III. Latent Dirichlet Allocation (LDA)

LDA is a probabilistic model that decomposes document-term matrix into two low-rank matrices: document-topic distribution and topic-word distribution. The topic proportions is a K-dimensional Dirichlet distribution and each topic is a V-dimensional Dirichlet distribution. Number of topics of LDA needs to be specified by user and this method is time-consuming. But as a generative model, LDA is able to generalize the model it uses to separate documents into topics to documents outside the corpus.

2. Method Used

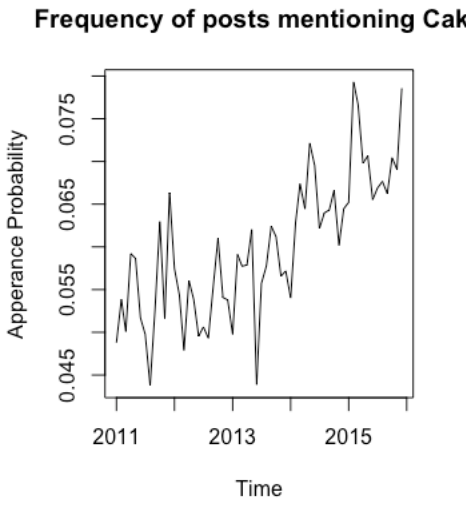
LDA is chosen to conduct the topic modeling task. Considering the long running time of LDA method, I only checked the top words appeared in Facebook posts that relates to food in 2015 and have a grasp of popular terms, which can be considered to test the term frequency in later processes.

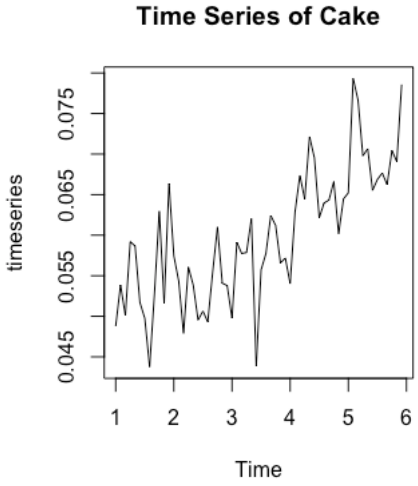
Secondly, I defined a function to return the monthly percentages of posts containing the term we want to explore in all the posts (60 in total, correspond to the time range (i.e. 60 months in 5 years)). I calculated the percentage here instead of absolute counts because percentage is a more reasonable metric regardless of the size difference in different months and years.

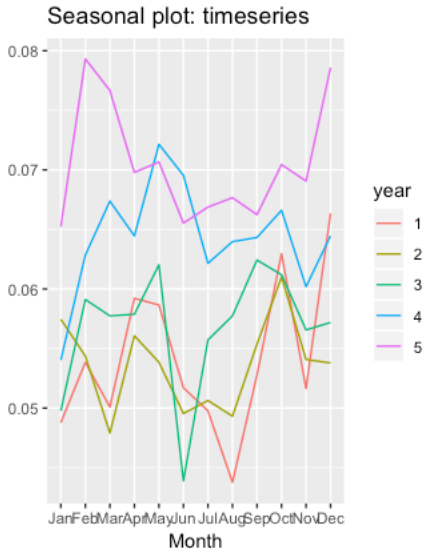
3. Result

Finally, I checked time series trend of different terms. Based on the result of LDA model, I first checked “cake”. There is an increasing probability in Facebook posts of term “cake”, so I checked “cheese cake” and “cup cake” for further exploration. I also used ts() function in R to convert numeric vector into R time series object to see the trend, the plots are the same and the original plot has a clearer x axis, so I did not use the time series object to plot. However,

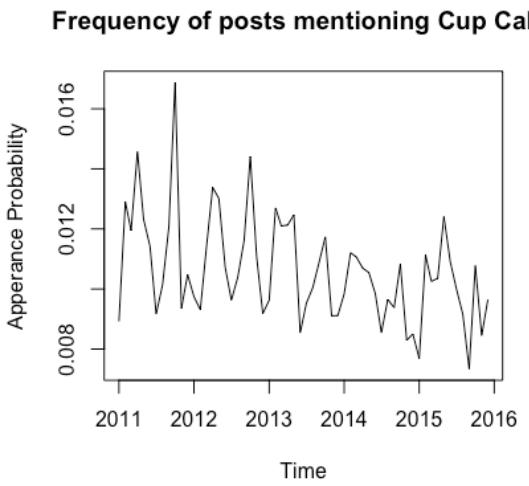
time series object is useful to plot seasonal trend. As the result of ggseasonplot shows, detailed trend analysis in terms of monthly change can be conducted.

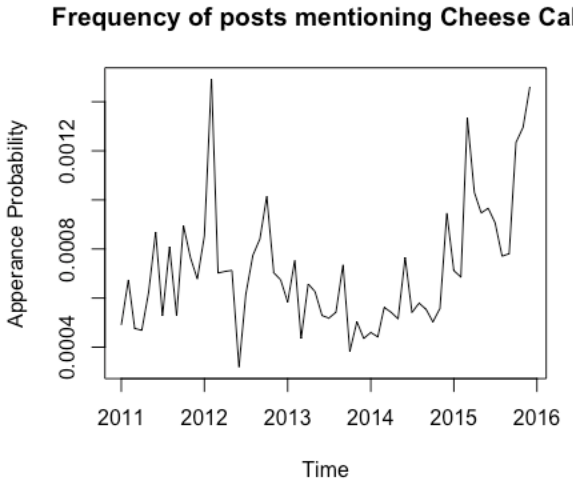




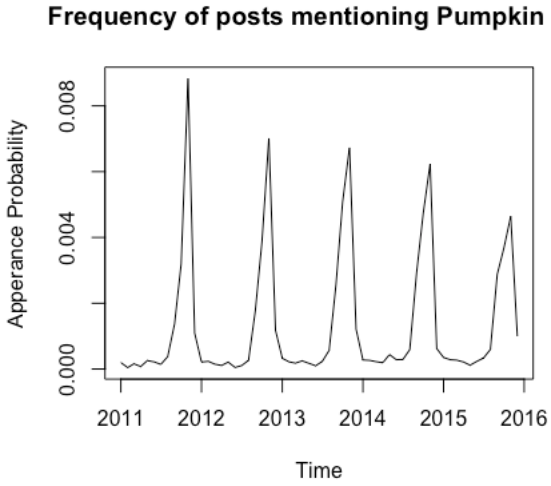


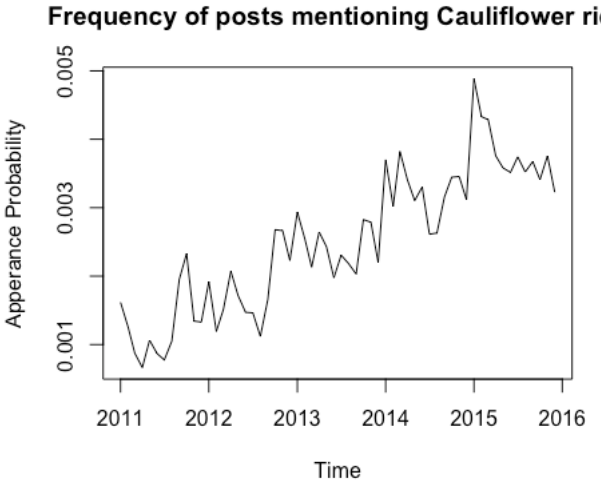
The frequency plots of “cup cake” and “cheese cake” indicate that more posts were mentioning cheese cakes from 2011 to 2015. And it is likely that the increase in “cake” frequency is partly due to the increase in “cheese cake”.

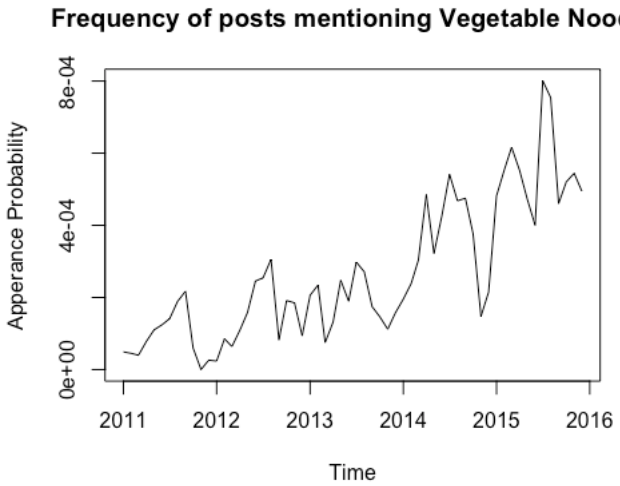




I used “pumpkin pie”, “cauliflower rice” and “vegetable noodle” to validate my function.

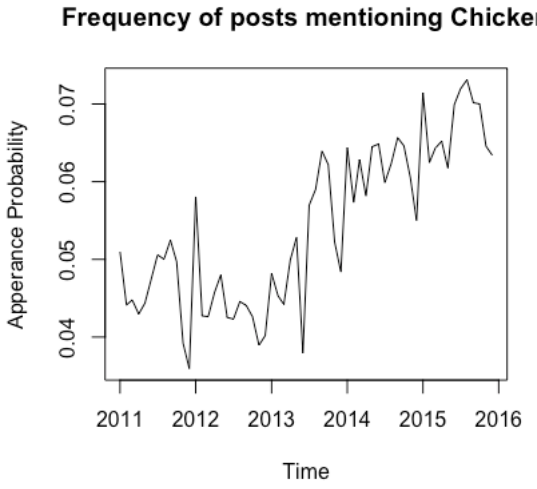


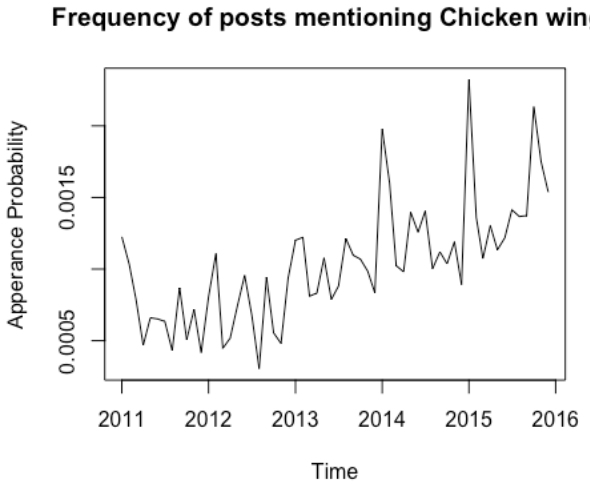




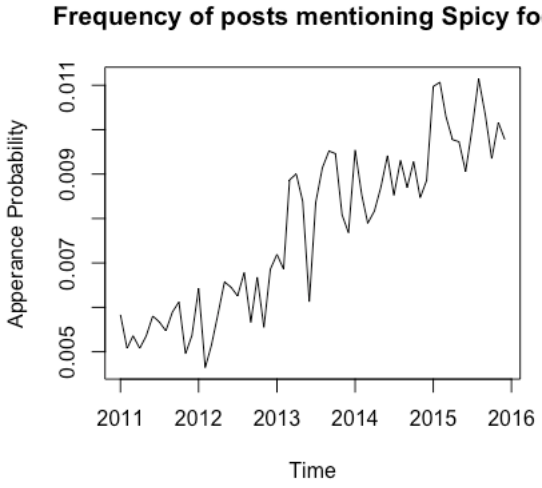
“Pumpkin pie” are more frequently mentioned in the end of each year, which may be due to Thanks-Giving (validated by the seasonally consumed food). And it shows increase in both cauliflower rice and vegetable noodles, which corresponds to the recent food trends.

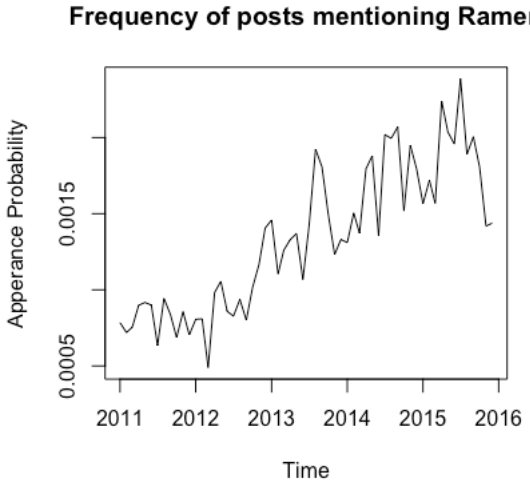
Following the same logic and procedure of cakes, I also explored the trend of “chicken” and conclude the increasing trend in chicken may be partly due to chicken wings.

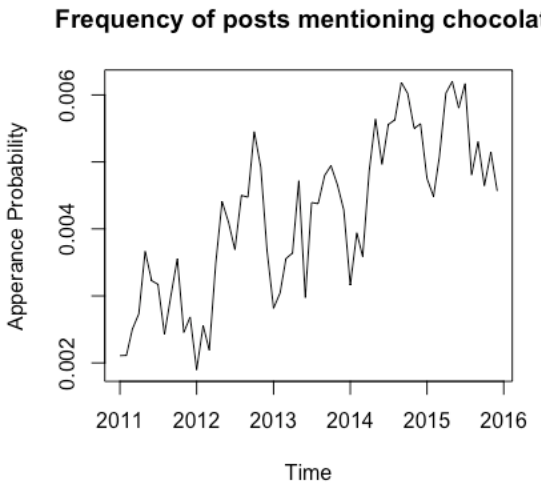


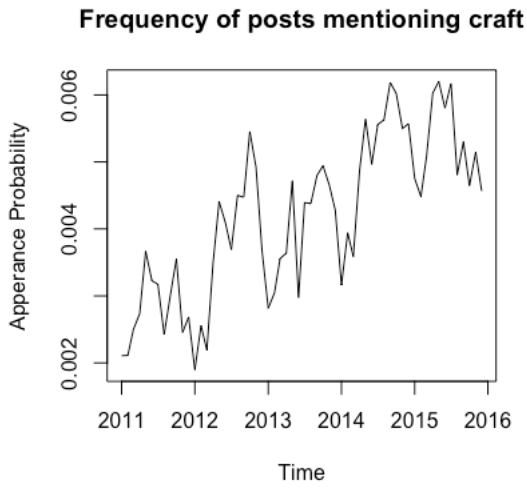


LDA model in previous steps is only used to return top words, but there are also emerging trends not mentioned much by majority posts. It shows that spicy food, ramen, chocolate and craft food are all emerging trends.






Frequency of posts mentioning chocolate
Apperance Probability
0.002
0.004
0.006
2011
2012
2013
2014
2015
2016
Time


Frequency of posts mentioning craft
Apperance Probability
0.002
0.004
0.006
2011
2012
2013
2014
2015
2016
Time