Text Mining Project Report

Project Introduction

- Analyze over 200,000 Facebook posts in 5 years from 2011 – 2015 and detect the trends of two types of health foods, “cauliflower rice” and “zoodle” (zucchini noodles), from their frequencies of being mentioned.

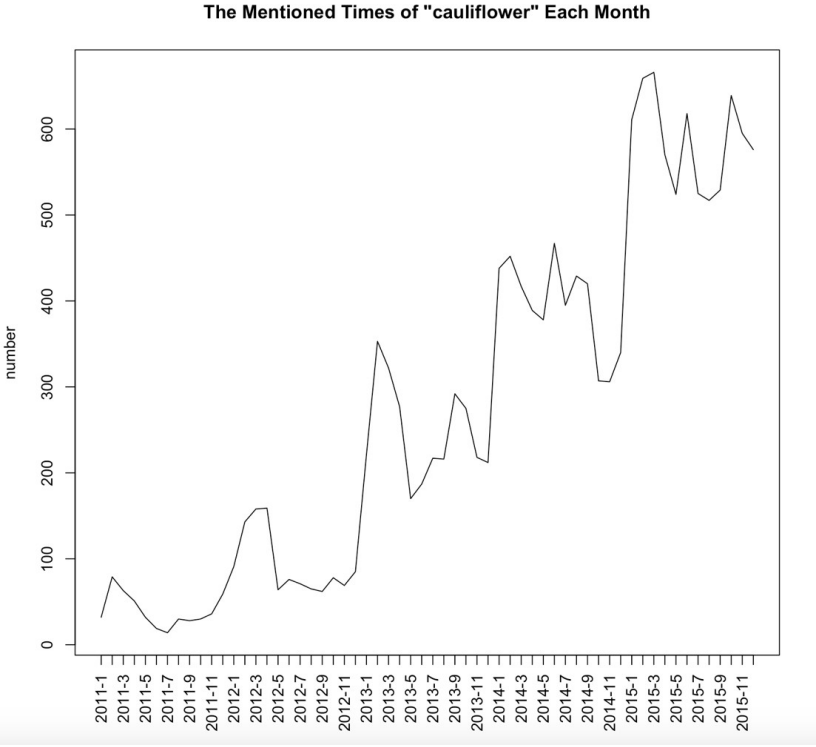
Potential Approaches

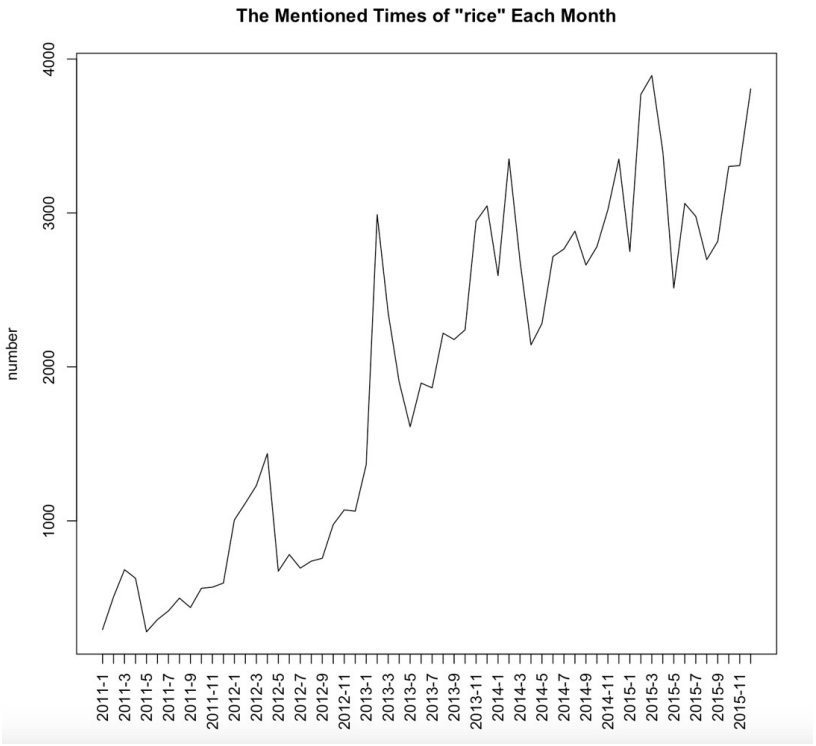
1. Use “topics” modeling and LDA. However, choosing the words related to certain topics and deciding the weight of those words are kind of subjective.
2. Output the food related words with highest frequency as word-cloud and detect the trends of those in the time series. However, since the words show most frequently are more likely to be those most popular foods and would be mentioned all the time, this method is not efficient to help us detect the trends of some newly thriving foods, just like “cauliflower rice” and “zoodle”.
3. Build the DTM (Document-Term Matrix) of the data files and join it with the ingredients’ d.
4. Use LSA to see the frequency that the word “rice” appears in the window of the word “cauliflower”. (Whether close to the frequency of “cauliflower” itself.)

Here I choose the third one and plot the line charts showing the times of the food words being mentioned by month.

Output & Interpretation:

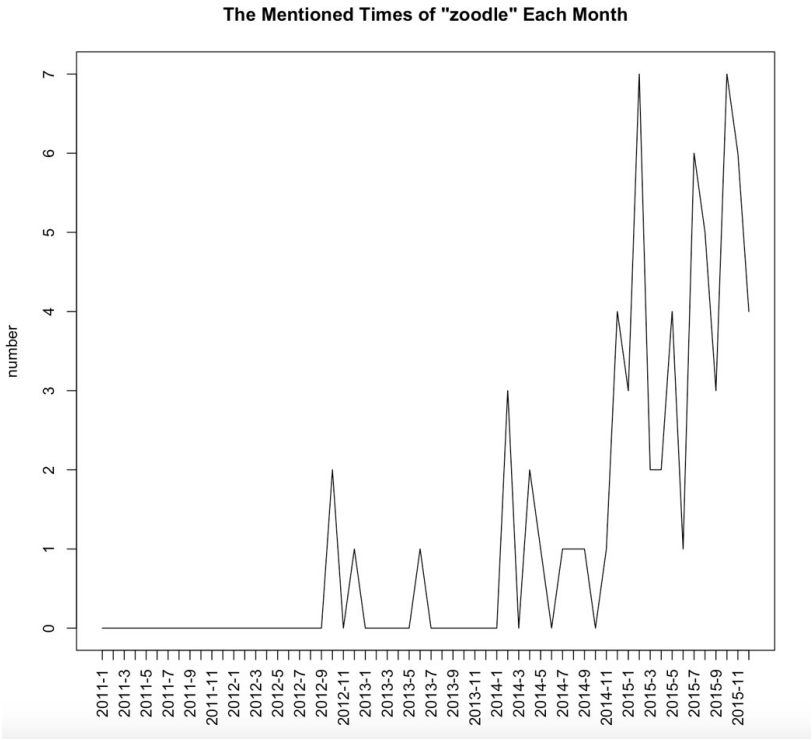
The followings are the plots show the frequency of occurrence of “cauliflower” and “rice”.





According to the two graphs, there's a positively correlated trends between the frequency of "cauliflower" and "rice". (The number of times that "rice" being mentioned is much larger since it's a kind of staple foods. But we still can detect a trend from the plot of "cauliflower" which can somehow represent the trend of "cauliflower rice".)

Comparatively, the frequency of "zoodle" shows seasonality: increasing trends in around December every year (the end of the year or the beginning of the next year).



To conclude, "zoodle" draws increasing attention year by year from 2011-2015.