CIS 434 Final Project Report: Food Trend Detection

MSBA Miao Xi

Introduction

Social media analytics is very useful for all consumer-facing industries to measure their performance and help to craft future improvement. One typical implementation is in food industry. Using social media analytics, industry professionals are able to detect trend by identifying food related buzz words and the change of frequency at almost real-time level. This greatly accelerates the process of product development. Companies with such technological advantage are able to quickly take-up unfulfilled needs of consumers. In this project, 4 million food-related Facebook post are leveraged to discover food develop a methodology that is able to discover and detect possible food trend among time.

Methods

There are several approaches toward this task. They differ from each in terms of underlying assumptions and how to define topics.

Bag of words Assumptions

The bag of words assumption regards language as a collection of independent terms.

The most commonly used bag of words model is uni-gram model, which assume each word are drawn from the same distribution of vocabulary regardless of words happen before or after it. By using uni-gram model, the simplest way to measure the trend of a food word is to count the frequency it appears in Facebook posts for a given time period and use the frequency as an indicator for trend. But this approach has some limitations. First it only measures a single word, it is hard to detect any combination of words happening together, hence it could miss some interesting trends which are expressed in a combination of words. One way to solve this issue is to define term as a combination of 2 or more words. But this approach would further increase the feature space to a document term matrix (DTM), and add more burden for computing. Second, a simple count could provide miss-leading result as the increase of frequency could be caused by the increase of sheer amount of all Facebook posts, not necessarily because certain trend is happening. More carefully constructed statistics shall be used as a measure of trend.

Another way to implement bag of word assumptions is to find busty keywords [1]. This method is the very similar to measuring frequency but it dramatically shortened the measure periods to minutes. This method is more capable to be used in monitoring live stream social media data to discover heated discussion. Not very applicable in the scope of this project.

Bag of words assumption could also be applied to this task in form of co-occurrence analysis. This is the method used in this project. It relies on the construction of a co-occurrence matrix, and identify frequently co-occur word pairs. This is an improvement compared to uni-gram model because it could detect the trend that consisting of two words, which is more suitable for the task as many food names are combinations of 2 words. Further detail of the method would be explained in later sections.

Latent Dirichlet Allocation (LDA)

The LDA method are based on the assumption that each document is a liner combination of topics and each topic is a linear combination of words. By using this generative model, underlying topics in a corpus could be discovered. Each month of Facebook food post could be seen as a corpus, topics extracted from each corpus, and the weight change over corpus for similar topics could be used as measurement of trend over time. This method also has its limitation as it is very computational intensive. The process of topic discovering is a very random process. It is hard to guarantee to find similar topics across corpus.

Co-occurrence Analysis for Food Trend Detection

Dictionary

The nature of this project is food trend detection. All words in Facebook posts that are not related to food are all unnecessary. Here a food word dictionary is kindly provided by the professor and TAs. It would be used to filter out all non-food related words to reduce the dimension all of term feature space.

Data Processing

Data preprocessing is a very important step for this project. It involves many memory and CPU intensive steps. The 4 million Facebook food posts across 5 years are provided in separate monthly csv files. It is still too big to construct single DTM. After each file was read into memory it is further broken down into smaller blocks, 2000 posts per block. And preprocessing code are written in iterations. Each iteration would process one block.

Workflow if each iteration can be summarized as:

1. Construct a full DTM on all words.
2. Subset DTM columns with only food related words form food dictionary.
3. Transfer each row of DTM into a logical vector.
4. Use this logical vector to subset all words of food DTM to acquire a smaller Facebook post data, each row corresponding to each post, and the column contains only food words extracted from original posts.
5. Use the new Facebook data construct a co-occurrence matrix.
6. Extract word pairs that co-occur more than 4 times in the 2000 post blocks.
7. Construct a dataframe that contains the word pairs and their co-occurrence frequency and the year month combination of the posts.

Trend Index

As discussed above, pure frequency count should not be used as a measurement of trend as it subjected to the increase influence of total Facebook posts growth. An index need to be constructed to compensate for such limitation. The Trend Score (S) are built as follows:

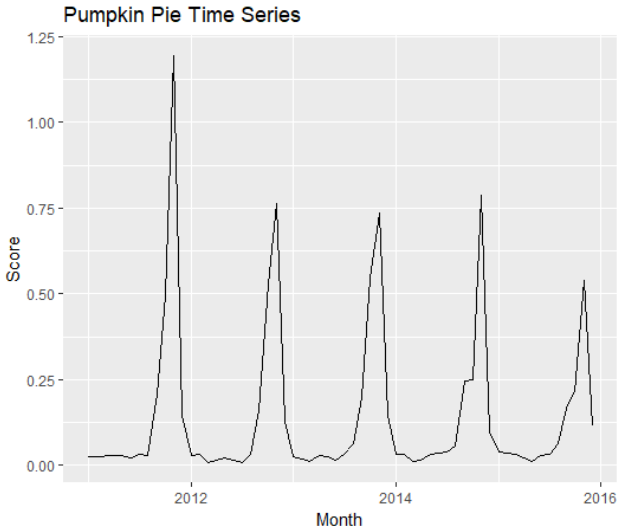
$$S = \frac{F_m}{N_m} \times 10$$

F_m is the sum of all co-occurrence frequency for a single word pair in a given month m. N_m is the total number of Facebook post in the same given month. This score measures the frequency of word pair co-occurrence relative the growth of the total posts, which compensate for the limitation for the raw count of frequency

Validations

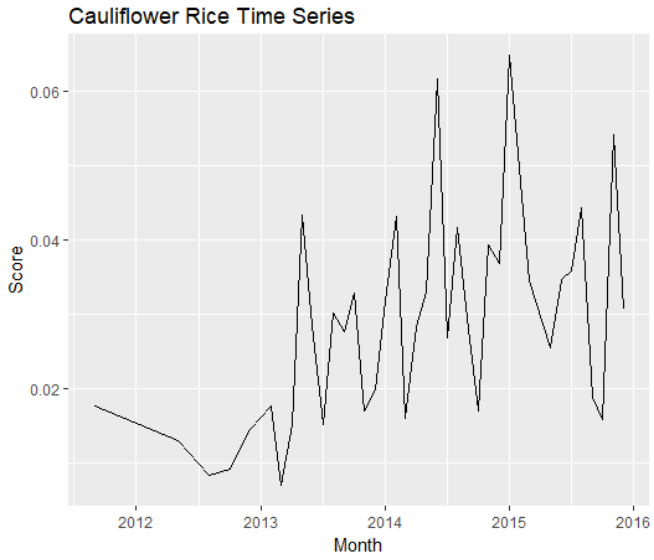
Here several ground truths are used to validate the detection method. Their trends measured by the detection method are visualized as a time series graph and compared to the background knowledge of certain fact of the food industry.

Pumpkin Pie validation



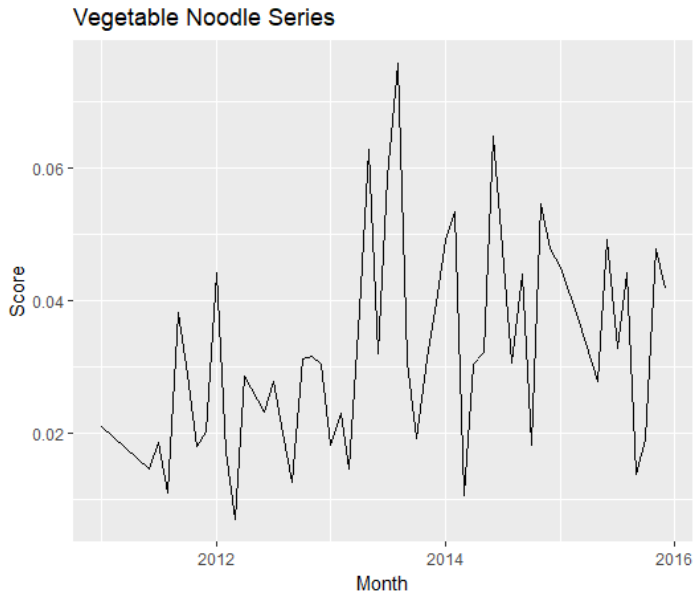
High seasonality for pumpkin pie at thanks giving could be clearly observed from the detection method.

Cauliflower Rice Validation



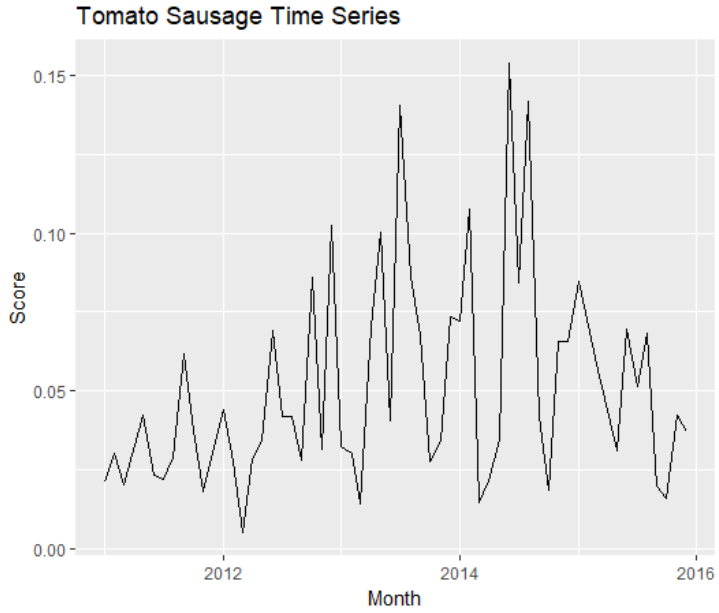
The trending of cauliflower rice is also captured in the graph, although substitute to some strong seasonal fluctuation, a general upward trend could be clearly observed.

Vegetable Noodle Validation



Vegetable Noodle is subjected to more fluctuation but its trending from early 2013 can still be observed.

Trend Detection



After the detection method has been validated, it is implemented to discover some interesting trend. Here we could see that between 2013 to mid-2014, people prefer mentioning tomato and sausage together. But lately this trend has declined.