Trend Detection Report

Tianyao Liu

Simon Business School

Background and Research Aim

This report presents the natural language processing (NLP) on Facebook posts from 2011 to 2015. By doing this analysis, I aim extract information from the large corpus to discover early food trend by probing into the data. This task could be divided into two parts: constructing time series of potential food trends and visualize the changes over time.

Potential Methods

Firstly, Distributional semantic modeling could be used to collect distributional information in high dimensional vectors, and to define semantic similarity in terms of vector similarity. By using this, we could find that the more semantically similar two words are, the more distributionally similar they will be in turn, and thus the more that they will tend to occur in similar linguistic contexts. In this case, food related topic could be constructed, and we could discover similar vector distributions in those Facebook posts.

Secondly, if we want to build a classification model, Neuron network is another suitable method. With inputs on features, we could determine whether people are more inclined to health diet over time or not. We could choose to build a 'non-deep' feedforward neural network with one hidden layer between input layer and output layer, or a deep neural network with several hidden layers. One thing to be careful about is over-fitting problem. Validations here is necessary to increase accuracy. Same logic could also be applied to is K-nearest Neighbors (KNN), which is also a useful model to do text-classification

If we want to get the importance of a word to a document in a collection or corpus, we could apply term frequency and inverse document frequency (Tf-Idf). This value increases proportionally to the frequency of a word in the document (tf) but decreases along with the increase of word frequency in the whole corpus (idf), which helps to adjust for the fact that some words appear more frequently in general.

Data Analysis Process on Chosen Method

In this project, I chose linear semantic analysis to examine the trend changes on two words: '*cauliflower*' and '*Zoodle*', representing social media users' interest on cauliflower rice and veggie noodle. To be more specific, firstly, I used *Corpus()* to import the huge corpus containing 60 elements. Secondly, I made some transformations on the raw data and constructed document-term matrix (DTM) with sparse terms with sparsity above 0.97 being removed. After that, I got frequency of all terms and frequency of *'cauliflower'*. I constructed a new column containing the file name and frequency using *cbind()*, and used a for loop to extract time in a 'year-month' pattern corresponding to every frequency. In order to have a clear look of the frequency ranking, I ordered the data frame based on frequency. In the visualization part, I re-ordered the dataframe based on time sequence and used *ggplot()* to plot a line chart regarding the trend of 'cauliflower' over five years.

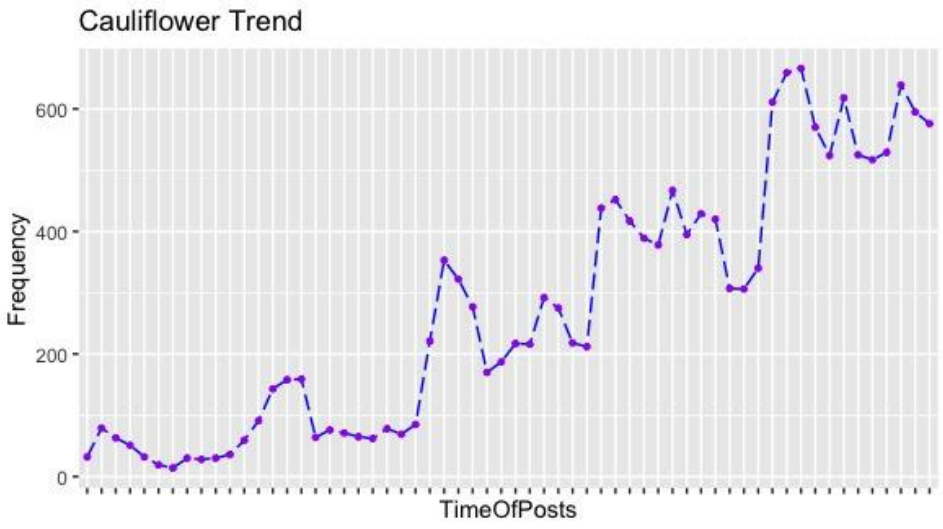

Figure 1: Cauliflower Trend

I also did the same on ‘Zoodle’ and observed its trend.

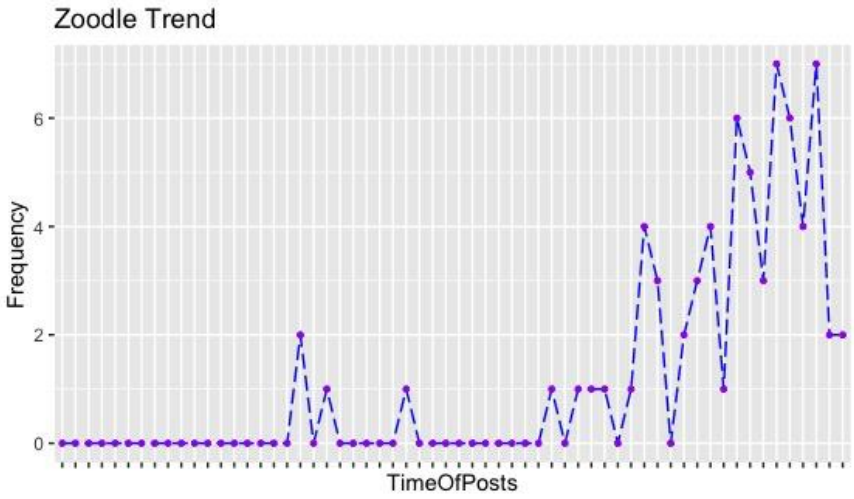

Figure 2: Zoodle Trend

Main Findings

- Both ‘cauliflower’ and ‘zoodle’ has an upward trend from 2011 to 2015. This means that they are more frequently being mentioned on social media.
- Frequency of the term ‘cauliflower’ is the most on 2015-11 as 666 times, and the least on 2011-12 as 14 times.
- The term ‘zoodle’ is relatively new, because it did not appear before 2015-03. After that month, frequency increased steadily. It is the most on 2015-07 and 2015-10 as 7 times.
- Based on those observation, we could conclude that people are paying more attention on healthy food concept and are likely to share them on social media.
- Deeper analysis may be needed for further explanations.

Insights and Suggestions for Large Supermarket Chains

- **Supply chain**: Prepared more stock and more frequently re-stocking on healthy food ingredients such as cauliflower, zucchini, squash and cucumber.

- **Operation**: Put those ingredients on front shelfs which are easy to be reached by customers.
- **Promotion**: Launch promotions (such as campaigns) on social media where people discuss food-related topics frequently.