Evaluation Of Food Trends

Tan Lik Sin

7 November 2015

INTRODUCTION

Food trends come and go, it is often important for restaurants to be keenly aware of what customers want most when they dine. Mining text from Yelp dataset provided as part of Yelp Dataset Challenge, we looked at frequencies of keywords in restaurant reviews to identify what diners pay attention to when they patronise a restaurant. Is it service that is often mentioned in reviews? Or is steak more frequently mentioned?
In this study, we assumed that the higher the frequency of a keyword, the more important the word is to reviewers. We ranked the relative frequencies of keywords for each year and performed linear modelling on the words to determine their trends. The METHODS section described the process used to arrive at the final dataset for our analysis.
We looked at how these keywords increase or decrease in frequency over time to identify food trends. We also looked at keywords that remained consistent in their importance to reviewers over time. The RESULTS section illustrated our analysis and in the DISCUSSION section, we interpreted the results of our analysis and concluded our findings.

METHODS

This section describes the methods used to obtain our results. We first loaded the necessary data and required library files. We then preprocessed the data by flattening the list of lists, extracting only restaurant businesses relevant for our analysis. The reviews were sorted according to dates and the distribution by time was identified.
Text mining on the reviews were performed to extract keywords with high frequencies for each year. The key words were ranked according to their frequencies and statistical modelling was applied. A p-value was obtained from the linear model employed on each word with time as predictor. The p-value was used to determine the significance in the trend of the keyword ranking, which in turn provided us with insights on the food trends over time as described in the RESULTS section.

Load necessary data and library files

1. Load library files that will be used (Code not shown)
2. Load Yelp data into RStudio (Code not shown)

Preprocessing data and exploratory analysis

When opened in R, the Json data appeared as list of lists in R data frame. We needed to unlist and flatten the data as much as possible to enable easy processing. We then selected only reviews that were addressed to restaurant businesses. Reviews were reordered by the year that they were composed.

3. Flatten data, especially the categories column in business data file (Code not shown)
4. Extract reviews targeted at restaurants (Code not shown)
5. Sort restaurant reviews by date (Code not shown)
6. We found that year 2004 contained only 10 reviews on restaurant. We remove year 2004 due to lack of information. (Code not shown)

```
##
##    2004    2005    2006    2007    2008    2009    2010    2011    2012    2013
##      10     422    2454   10352   27437   48062   89084  135411  155988  209502
##    2014    2015
## 302500    9405
```

Text mining in reviews

We want to identify frequency of words used in restaurant reviews over time so as to pick up food trends. We divided the reviews by the year it was made and determined the frequency of keywords used in the reviews. We created a function to convert the text into a format suitable for processing, cleaned up the text in the reviews, and output 200 words with the highest frequencies for each year.
We randomly selected up to 100000 reviews for each year to determine high frequency words. The cap of 100000 reviews was selected to reduce the excessive amount of processing time needed for text mining.
We merged the top 200 highest frequency words of each year into a single dataframe. By doing that, we implicitly only considered high frequency words that made it to the top 200 every year. There were 93 words common to the top 200 of each year.
We ranked the words in the merged list to determine their relative frequencies (higher rank = higher frequency) for each year.

7. Create function to clean up text and output words with the highest frequency (Code not shown)
8. Create a data frame merging words with the highest frequencies for each year (Code not shown)
9. Rank words according to their relative frequencies (Higher rank for higher frequency, highest=93, lowest=1) (Code not shown)
10. Transpose rank dataframe (Code not shown)

Statistical Modelling

We now have words with the highest frequencies for each year. We assumed that words deemed more important to diners would appear more often and thus ranked higher. Having populated a data frame with high frequency words from reviews over the years, we looked at trends of these words over time. We generated linear models for rank of each word with time as predictor and obtained the p-value for each word to determine if the change in rank over the years was significant. The null hypothese was that there is no change over time. A low p-value suggests strong evidence against the null hypothesis.

11. Generate linear model (predictor=year) and rank each word according to the respective p-value. We also obtain the slope, and the rank of the word in 2015. (Code not shown)

RESULTS

We looked at words with the lowest p-values and words with the highest. A word with low p-value has low probability that there is no change in rank over time. This meant that rank change over time was significant. This would be of interest to us as it showed significant increase or decrease in the importance of the word to reviewers of restaurants. Conversely, a word with high p-value has high probability that there is no change in rank over time. This meant that there is insufficient evidence to suggest that the rank changed over time. This would also be of interest to us as it showed the consistent importance of the word to reviewers of restaurants.

12. We look at top 20 words with the lowest p-values, the slope and the rank in 2015

```
head(textfit, 20)
```

```
##       word       pvalue        slope 2015rank
## 12     can 9.736799e-08 -0.004455400     68.0
## 15    come 7.123623e-07  0.007106336     74.0
## 34     get 1.611609e-06 -0.001468542     83.0
## 48    make 3.378913e-05 -0.003235868     55.5
## 31   fresh 4.608340e-05  0.006658176     49.0
## 27  experi 8.165293e-05  0.007379910     43.0
## 88    vega 2.071697e-04 -0.013664785     29.5
## 57     one 2.107986e-04 -0.002016218     82.0
## 76   steak 2.879929e-04 -0.006072909     10.0
## 72  servic 2.898972e-04  0.002327383     88.0
## 14 chicken 3.031536e-04  0.005600178     71.0
## 89    wait 3.344518e-04  0.011536297     72.0
## 54   never 3.649191e-04  0.006060590     39.0
## 6      bar 3.762620e-04 -0.011449472     42.0
## 77   still 8.138125e-04 -0.003534645     15.0
## 25   everi 9.105304e-04  0.001294421      9.0
## 52    menu 1.767611e-03 -0.002177840     69.0
## 45     lot 1.872188e-03 -0.003683743     19.0
## 93    will 1.960810e-03  0.006011399     77.0
## 56   night 2.078508e-03 -0.008487093     35.0
```

13. We look at top 20 words with the highest p-values, the slope and the rank in 2015

```
tail(textfit, 20)
```

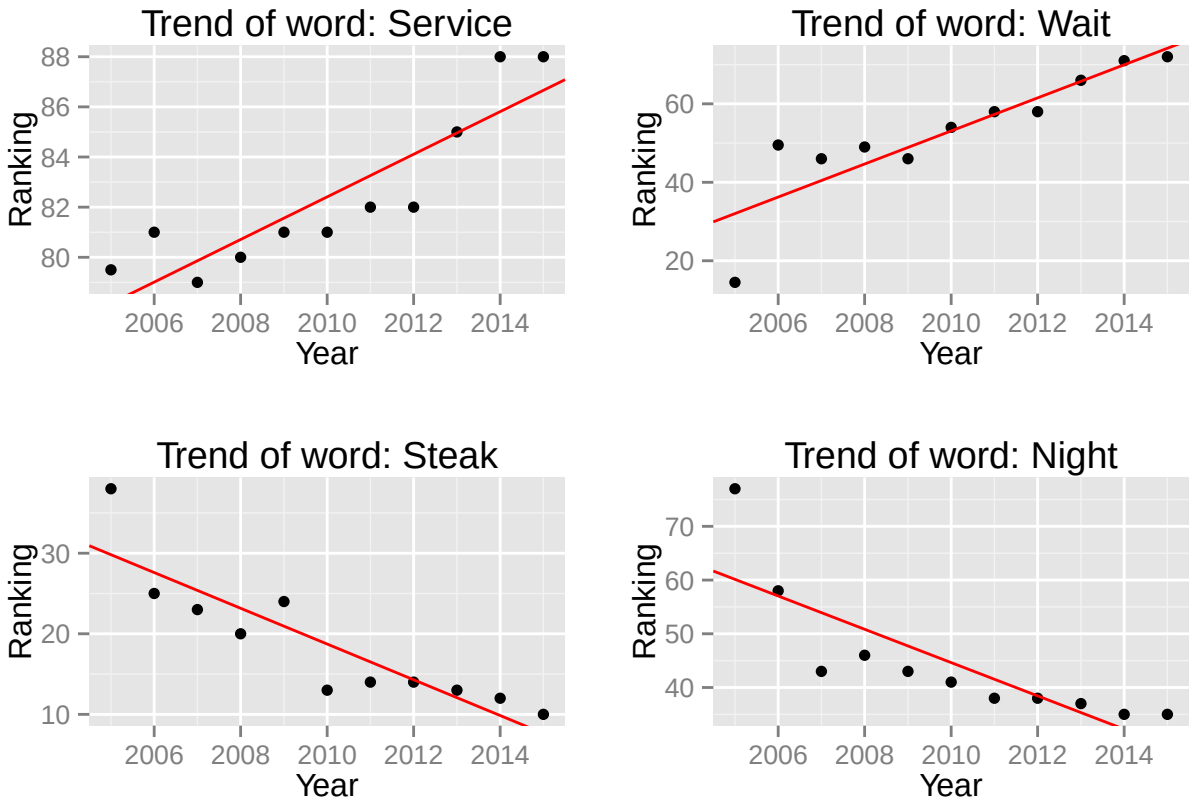
```
##     word    pvalue         slope 2015rank
## 59 peopl 0.5584213  6.717444e-04       32
## 60 pizza 0.5588910  1.904965e-03       60
## 37 great 0.5615272  3.237633e-04       89
## 63 price 0.5843431  9.588405e-04       64
## 3   area 0.6381481  2.240366e-04        4
## 91   way 0.6645597 -5.604478e-04       22
## 23  even 0.6647534  7.088763e-04       67
## 22   eat 0.7240995 -5.603859e-04       63
## 36  good 0.7243179  7.463551e-05       92
## 53  much 0.7272009  1.081905e-03       47
## 43 locat 0.7752994  6.975171e-04       34
## 16   day 0.7909152 -2.862226e-04       28
## 39  just 0.8223433 -9.973911e-05       85
## 20  dish 0.8458262  3.728492e-04       46
## 92  well 0.8746425  1.616772e-04       61
## 41  like 0.8756795  1.242418e-04       86
## 26 excel 0.8961893 -4.471006e-04       25
## 67 salad 0.9081698 -1.743374e-04       45
## 78  tabl 0.9169715 -1.367533e-04       62
## 38  high 0.9696412  6.260722e-05       12
```

DISCUSSION

The RESULTS section displayed the top 20 words with the lowest and highest P-values and their corresponding slopes and 2015 rank. A low P-value indicates that there is significant change in rank while a high P-value suggests

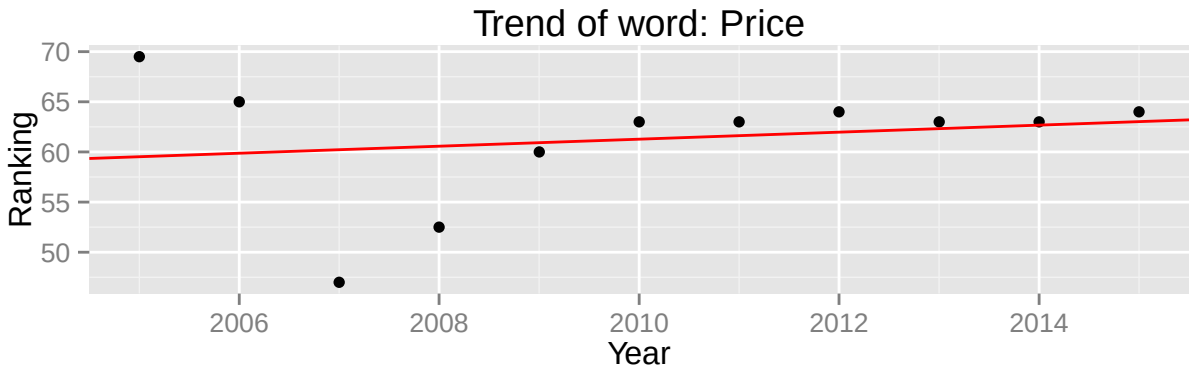
little probability of change in rank. A word with positive slope has increasing rank and hence importance over time while the converse is true. A high 2015 rank demonstrates a high frequency of the word in 2015 reviews and thus are important to reviewers while a low rank suggests that the words are lower in importance.
For top 20 words with the lowest P-values, it is important to pay attention to words with positive slope and a high 2015 rank. These words are increasing in importance and are relatively important to reviewers in recent times. From the table generated in (12), words like "service", "chicken" and "wait" have positive slope and high 2015 rank. It suggests that reviewers are paying increasing attention and emphasized in recent times the service a restaurant provides, chicken as an ingredient to the food they eat, and the amount of time they have to wait for seats to the restaurant. The ranks of "service" and "wait" over the years as well as the regression lines are illustrated in the plot of (14).

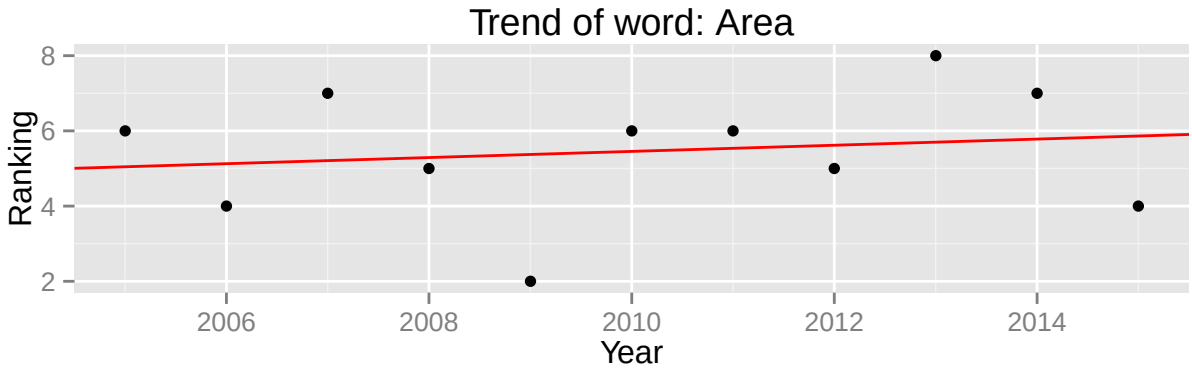
14. We look at keywords "Service", "Wait", "Steak" and "Night" to see how their importance changed over time (Code not shown)



On the other hand, it is equally important to consider words with negative slope and a low 2015 rank as these words are decreasing in importance and might not be relevant to reviewers in 2015. These words include "steak" and "night" and suggests that having a steak at night might not be the highest on a customer's mind. The ranks of "steak" and "night" as well as the corresponding regression lines are illustrated in the plot of (14).
For top 20 words with the highest P-values, we noted the relatively flat slope as compared to words with low P-values. This is consistent with the high P-values in suggesting little change in the rank over the years. From the table generated in (13), we noted that words like "price" and "pizza" have relatively high ranks. This suggests that customers are relatively conscious about the pricing of a restaurant and pizza remained consistent in its popularity over time. On the other hand, "area" and "salad" received relatively lower rank in 2015. This may suggest that customers are not sensitive to the neighbourhood of the restaurant and that salad is consistently less popular than pizza. The ranks of "Price" and "Area" are plotted in (15).

15. We look at keywords "Price" and "area" to see how consistent the keywords were over time. (Code not shown)





Returning to our objective of this study, we have identified an increasing trend in the importance of service provided by restaurant, food containing chicken, as well as length of time a customer has to wait. Restaurateurs or potential restaurant owners will want to pay attention to the service and time a customer has to wait as this will likely impact their reviews on Yelp. Equal attention should also be focused on dishes with chicken. We have also identified steak and night with a decreasing trend in importance so restaurateurs might want to focus less on marketing and publicising their steaks.
Perhaps unsurprisingly, price and pizza are often mentioned in food reviews, consistently over the years. This showed that diners remained price sensitive and restaurateurs must adopt a brilliant pricing strategy to attract customers. Pizza remained an important item that should remain in the menu for a long time to come. On the other hand, area and salad are less frequently mentioned. If the restaurant is attractive to customers, the neighbourhood should not matter much. Given the rise in "eat clean" on social media, it seemed a little baffling that salad, the dish often associated with healthy food, remained lower in ranking than pizza.

Further studies

Due to constraints of time, this study could only cover a limited scope of what might have been achieveable given the rich dataset available. We described what can further this study below:
1. By merging the keywords with high frequencies in each year, we implicitly only considered high frequency words that made it to the top 200 every year. Words that only appeard in the top 200 in later years were not considered in the study. These emerging keywords and thus food trends can be considered in further studies.
2. We arranged the keywords by the year the reviews were made. We can further analyse intra year seasonal food trends by looking at the frequency distribution in finer time resolution.
3. By analysing single words, we are missing out on the context of the reviews. The context might be important for deeper understanding of the preferences of diners.
4. As observable in the top 20 words, there remained many words that were not useful in our analysis. More detailed data cleaning can be performed such that more word trends might appear.