Data Mining Capstone Task 1

September 14, 2015

1 Abstract

2 Implementations

In this task, I used *python* toolkits to perform the analysis. In particular, I used packages such as *gensim*, *sklearn* for the topic extraction.

For data visualisation, I used *D3* to do all the drawings.

2.1 Topic mining of all restaurant reviews (Task 1.1)

To understand what people are talking about among the review data, I used topic model LDA to extract 10 topics from all the reviews that are for *restaurants*.

I applied *TfidfVectorizer* to vectorize the raw review data, after filtering all reviews and only collect those for restaurants. The TF transformation was made to be linear, and I enabled IDF reweighting. I specified the n-gram range to be 1 to 2, i.e. terms with 1 or two words were collected.

By using *D3*, I applied *word cloud* to visualise the data. Essentially, each cloud represents a topic. The font size shows the significance of the corresponding term in this topic. Ten colours are used to distinguish the topics.

Figure 1 shows the result.

The result is also shown in this link.

Here are several observations:

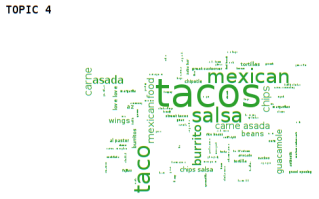
1. Cuisines / foods are very popular topics – topics 4, 5, 7, 9 more or less emphesize a particular cuisine, e.g. Mexican tacos, pizza, Japanese sushi, and Thai / Chinese food, while topics 0, 6, 8 emphesize certain foods or drinkings.

(a) Topic 0

(b) Topic 1

(c) Topic 2

(d) Topic 3



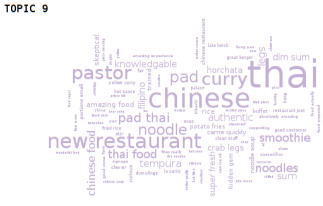
(e) Topic 4

(f) Topic 5

(g) Topic 6

(h) Topic 7

(i) Topic 8



(j) Topic 9

Figure 1: Ten topics mined from all restaurant reviews

2. General comments (usually good) to the restaurants – topics 1 and 3 clearly show a good impression, while topic 2 possibly suggests an inferior impression in terms of waiting (and then asking for managers). In particular, topic 2 also suggests that *time* is a very important topic, or factor, when customers are reviewing a restaurant.

3. Locations are also mentioned, especially "Las Vagas Valley", spreaded amongst topics 0, 1, 6.

2.2 Topic mining of positive and negative reviews (Task 1.2)

In this task, I managed to explore the topic distribution for subsets of all the reviews. In particular, this section introduces the observations made upon subsets of negative reviews (reviews with star number <= 2) and positive reviews (reviews with star number >= 4).

The topic model used is still LDA, with identical configurations as in task 1.1.

Figures 2 and 3 show the result.

The result is also shown in this link (positive views) and this link (negative views).

From these results, it is shown that:

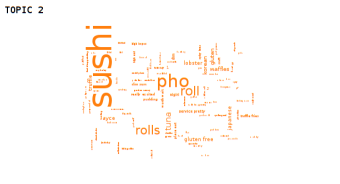
1. Both positive and negative reviews still talk much about food or cuisines themselves frequently. For instance, in negative review topics, tacos, crab legs, sushi, carne asada, pizza, hot dogs are the top topics. While for positive reviews, still pizza, sushi, Indian food, chicken, breakfast, oysters are mentioned.

2. Compared with all reviews, these subsets now shows something different in different subsets.

    - In the positive review topics, there is no negative impression when looking at the top words in each word cloud of each topic. And even for the positive topics, the phrasing are quite general ("great place", "amazing", "definitely coming") and not touching specific factors that influence customers' rating.
    - However in the negative review topics, many specific terms indicate what the customers are looking for – "portion size", "short ribs", "limited menu", "service", "time", i.e. amount of food, menues, service, and time. At the same time, general comments (e.g. "just okay") can still be seen.

(a) Topic 0

(b) Topic 1



(c) Topic 2

(d) Topic 3

(e) Topic 4

(f) Topic 5

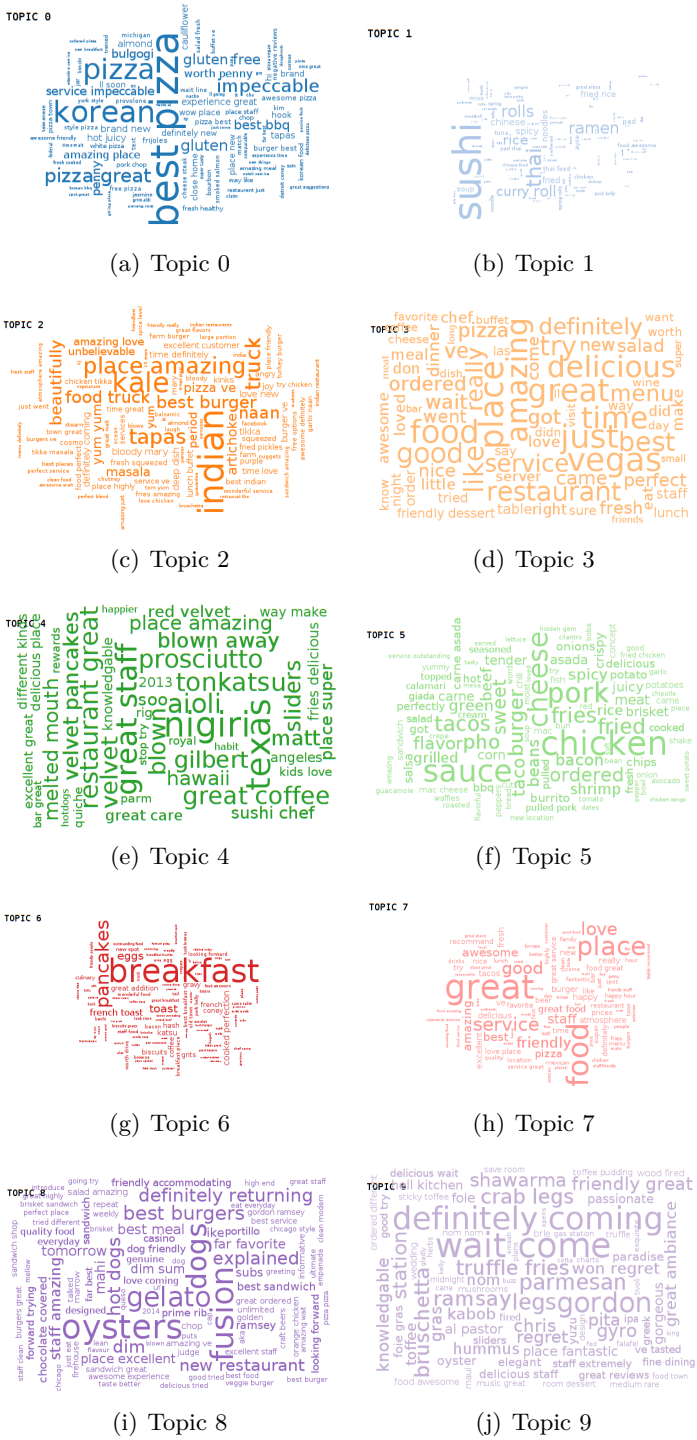
(g) Topic 6

(h) Topic 7

(i) Topic 8

(j) Topic 9

Figure 2: Ten topics mined from negative restaurant reviews



(a) Topic 0 (b) Topic 1

(c) Topic 2 (d) Topic 3

(e) Topic 4 (f) Topic 5

(g) Topic 6 (h) Topic 7

(i) Topic 8 (j) Topic 9

Figure 3: Ten topics mined from positive restaurant reviews

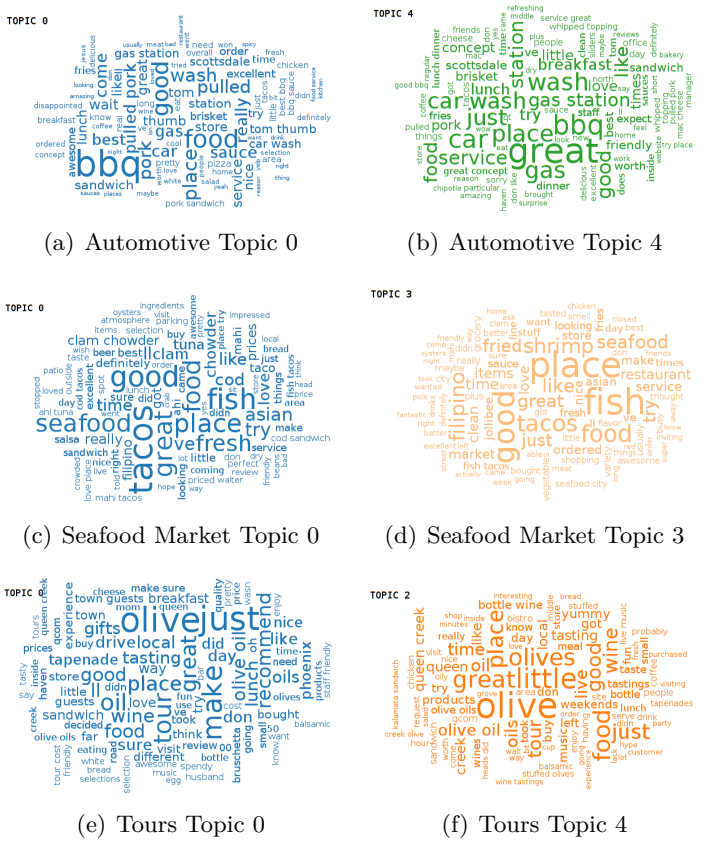
2.3 Topic mining of reviews from different cuisines

Since reviews are made to different restaurants, and different restaurants do have different categories, or cuisines. Intuitively, one could imagine the topics mentioned in the reviews could vary.

Here I did some topic extraction upon reviews from different categories, and hereby present the comparison between category "Automotive", "Tours", and "Seafood Market".

Figure 4 shows a collection of all three categories, 2 topics sampled for each.

The result is also shown in Automotive, Seafood Markets, and Tours.



(a) Automotive Topic 0 (b) Automotive Topic 4

(c) Seafood Market Topic 0 (d) Seafood Market Topic 3

(e) Tours Topic 0 (f) Tours Topic 4

Figure 4: Ten topics mined from reviews of categories "Automotive", "Seafood Market", and "Tours"

As is shown, although food is still the popular topic, there indeed are

some divergences among the three categories:

1. The food mentioned are different, of course. (BBQ for automotive, fish and shrimps for seafood, and olive oil for tours.)

2. In particular, speciality of the restaurants are reflected in the reviews. Especially, in the automotive category, "car wash" and "gas station" are mentioned frequently. And it is understandable – a restaurant that is relevant to automotives, is very likely to be a restaurant near a gas station, or some automotive service stations.

2.4 Exploration of the star distribution

As is mentioned in the observation of Task 1.1, the topics extracted shows that at least two of them indicate a general positive impression, while only one mentioned "time", which might potentially reflect some neutral or negative impression.

So I then become curious about how the positions of all the reviews distribute – do people tend to comment positively or evenly?

I choose to observe star distribution to examine this topic.

Figure 5 shows how reviews of 1-5 stars distribute.

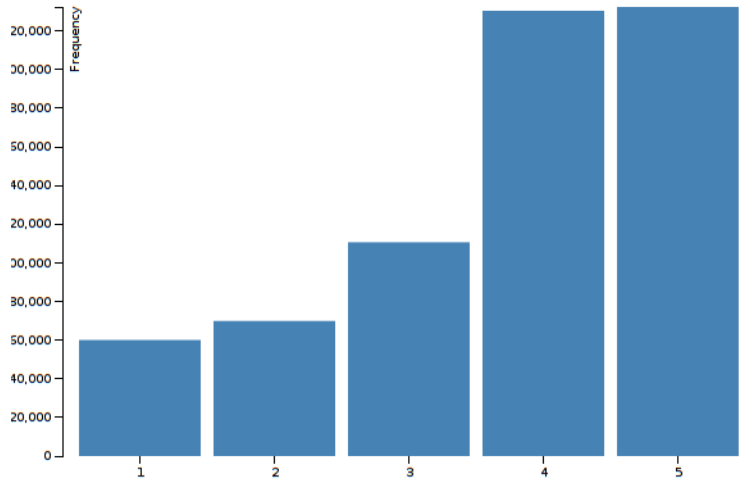


Figure 5: Ten topics mined from positive restaurant reviews

As is shown, the reviews are highly skewed, where positive reviews takes majority of all reviews.