Data Mining Capstone Final Report

A Review of Project Subtasks and Exploration of Key Factors that Influential Customers Care

1 Abstract

In this final report, I reviewed all of the six sub tasks that have been completed throughout this capstone project, summarizing that useful information can be provided to customers by mining data from the review texts joint with additional information. In addition, I also do extra explorations upon the factors that are cared more by key influential customers, which, as a useful information retrieved by data mining techniques, can be provided to restaurant owners as a reference.

2 Project Activities

In this section, I am firstly presenting a review of all the six tasks that have been done over this capstone project. In addition to the contents that I also put in each task report, I am also showing some extra works and results done as an extension of each task topics.

2.1 A Review of All the Sub-tasks

2.1.1 Exploration of the dataset, and a review of Task I

It is important to get a general impression of the data before going deeper and mine more useful information from it. In task I, as was requested, I used *LDA* topic model to extract topics from all the review text, as well as two contradictory review text subsets. In addition, I also explored the data from multiple perspectives.

People talk about dishes

Following figures of word cloud visualization demonstrate what people are talking about in reviews:

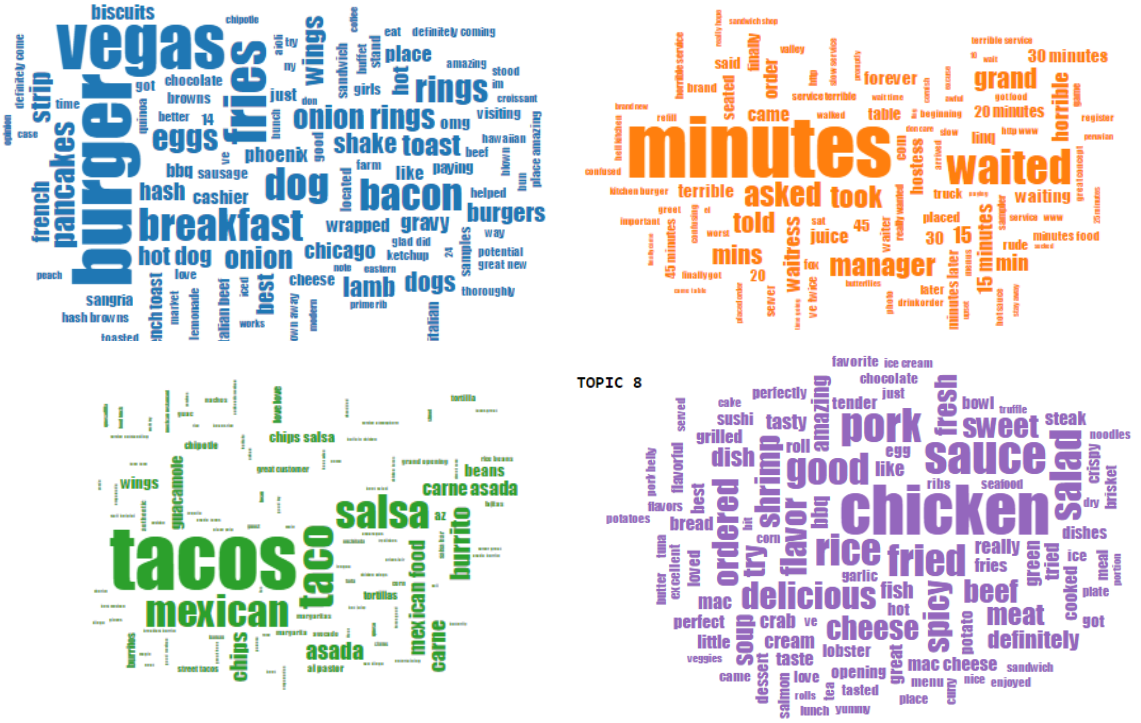


Figure: Four topics (out of ten) extracted from all reviews using *LDA*

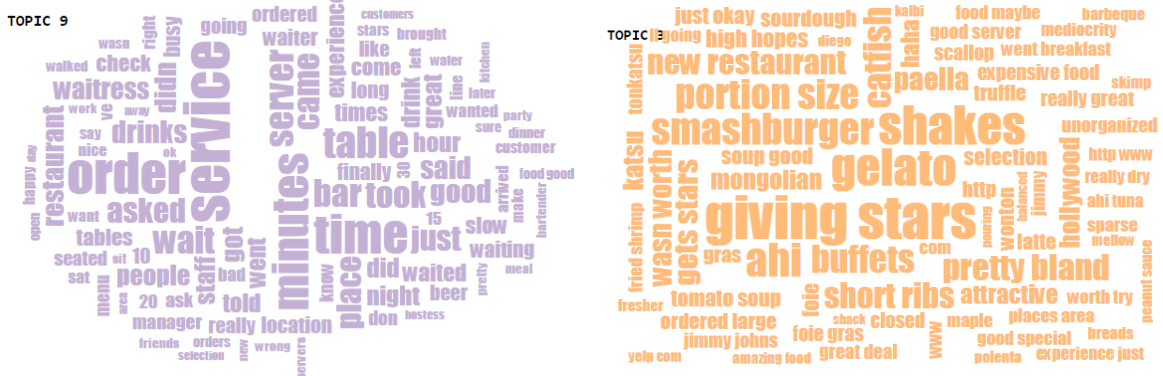


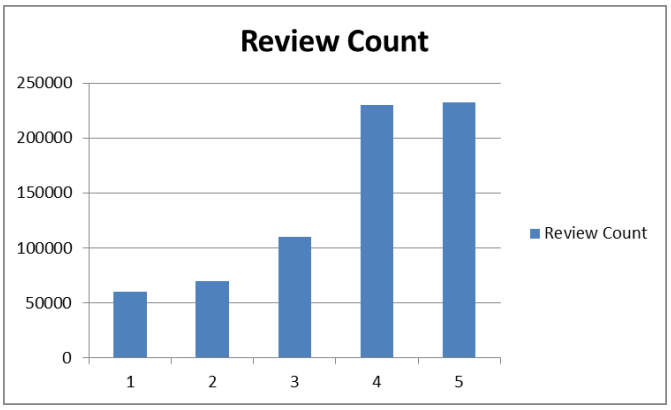
Figure: Two topics (out of ten) extracted from negative reviews (STAR < 3) using *LDA*

Figure: Two topics (out of ten) extracted from positive reviews (STAR > 3) using *LDA*

It is evident to see from these clouds, whatever it is from all reviews or from subsets of reviews, people mainly are talking about dish names. And less commonly, they talk about the services. While even though the ratings vary, the common topics do not differ much, as empirically perceived from these clouds.

People tend to rate highly

While here is another impression: there is hardly any negative terms visible in any of these word clouds, even for those generated for negative review subset. Interestingly, the overall rating distribution is skewed – people tend to rate highly, or people tend to rate when they are satisfied:



Increasing reviews with temporal patterns

After knowing the overall rating distribution, it is interesting to know how reviews are accumulated over time. Following chart illustrates how reviews are increased over the past ten years daily:

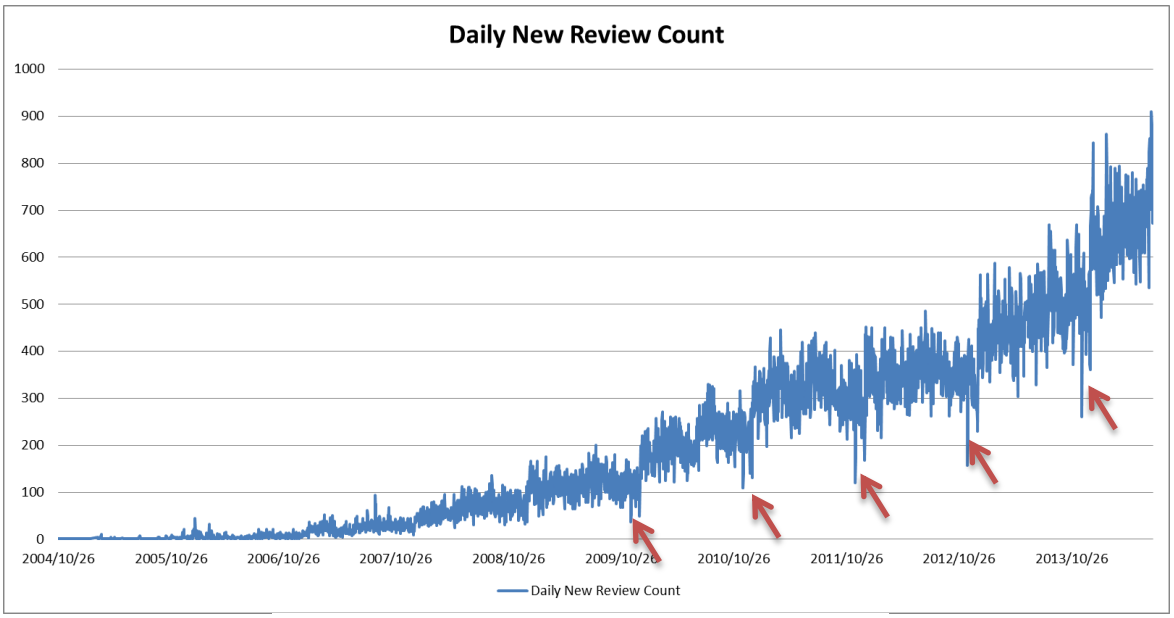


Chart: Daily review count.
Pointed parts all have a 2-spike pattern, and all of them occur around Nov. 24th, and Dec. 25th.
Spikes mean low review count increase.

So it shows that people possibly don't review much on Thanksgiving and Christmas weeks. Maybe they don't dine outside much at all on these special days?

Review length distribution

And how do the reviews look like in terms of their length? Following chart answers this question:

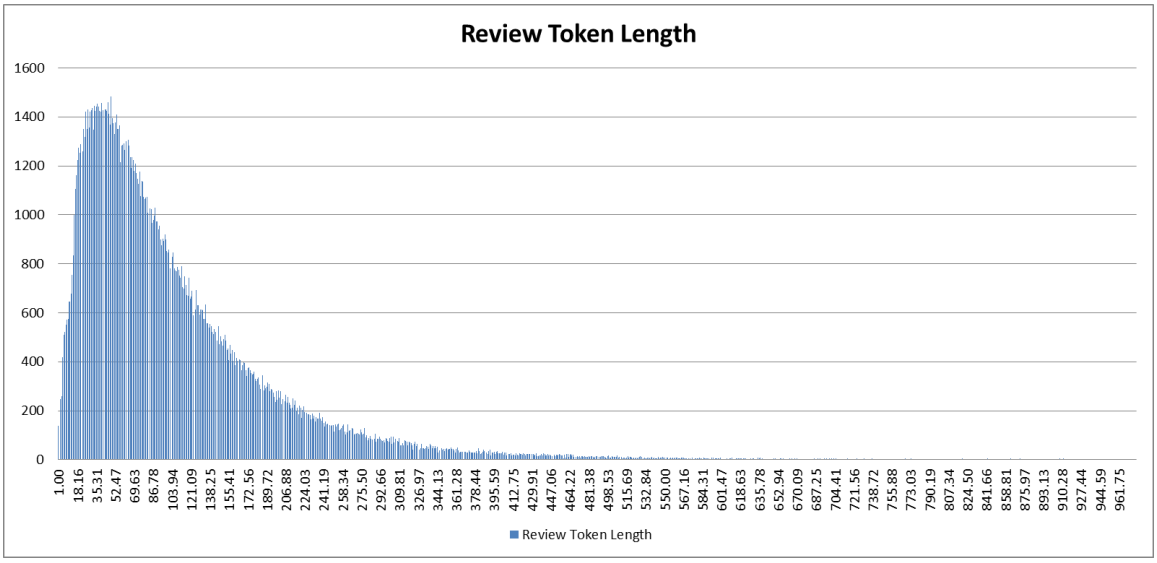


Chart: Review token length distribution
X-axis is the review length, Y-axis the frequency.

2.1.2 Cuisine clustering and review of Task II

Restaurants are mostly intuitively categorized by cuisines. And in this task, with knowing that people are mostly talking about dishes in task I, a cuisine similarity map can be mined by comparing their topics, since intuitively cuisines differ in dishes mostly from each other.

I firstly used TF-IDF vectorizer and LDA model to extract the topics on the raw review texts of each category, and then computed cosine similarity between categories, using the vectors of topic weights as returned by LDA model. In particular, I used sublinear TF modifier, and at most 10000 features as LDA model's input, and 100 topics were specified to be extracted.

Following is the visualization of the similarity heatmap between every two cuisines. After comparing with the results computed by approaches without TFIDF or topic extraction, it is evident that differences between certain cuisines got magnified, since the effects of common words are reduced by TFIDF modifiers.

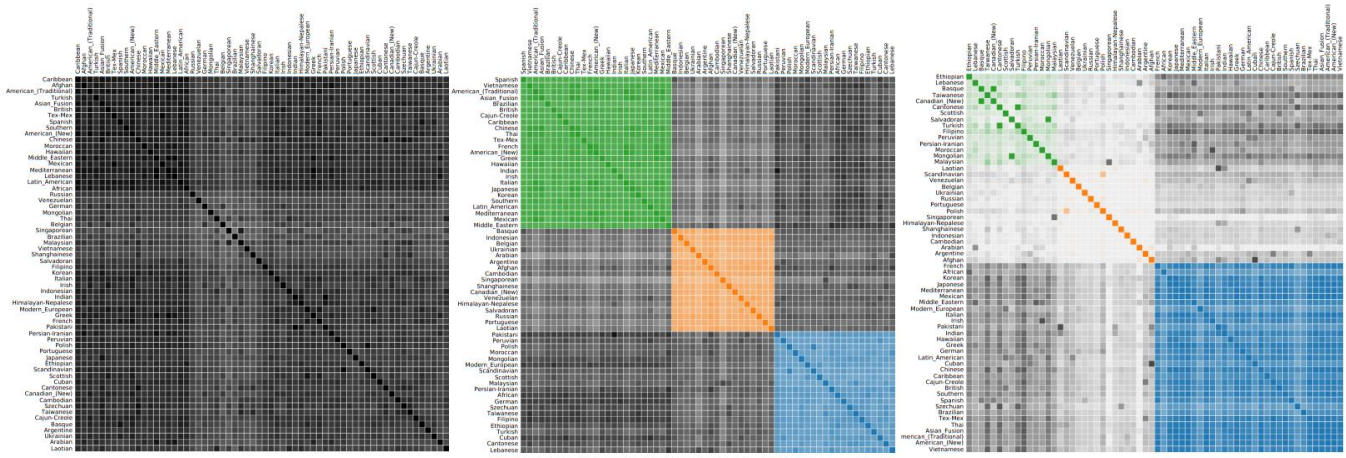


Chart: Cuisine similarity heatmap with clustering.
Dark means similar.
From left to right with different ways of computing similarities: only text representation (unigram), unigram with TFIDF, topics extracted with TFIDF.

And if we look the final result in another way:

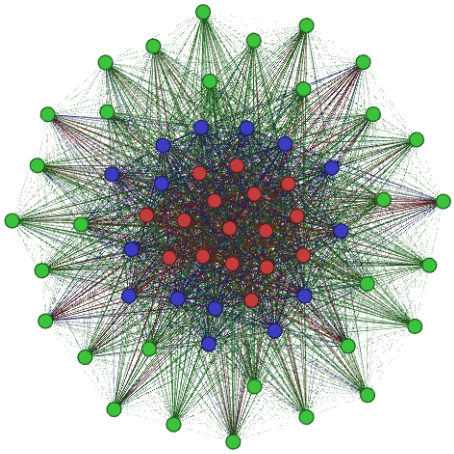


Chart: Cuisine similarity map with clustering.
Shown using Gephi ForceAtlas layout.
Each node denotes a cuisine.

2.1.3 Dish Recognition and review Task III

In this task, it was required to extract dish names from reviews of a certain cuisine.

In 2.1.1, it is understood from the LDA results that people are mainly talking about dishes in their reviews. So it is straightforward to get an okay result with simple topic extraction. However, better quality of dish names is required.

In this task, after comparing the experimenting results among three algorithms, it was concluded that high-quality dish names can be extracted by taking multiple considerations in mine, including using mutual information (as done using **Word2Vec**), and using classification based on manual labeling or common knowledge base (as done using **SegPhrase**), when doing extraction so as to reduce irrelevant noises.

The results are as followed:

Using **ToPMine**, the top phrases include not only dish names, but also common phrases like "food was good", which actually are noises that compromises the final quality.

Using **Word2Vec**, which harnesses the mutual information quantity, the dish names are computed by finding similar terms of known dishes. The top phrases shows better quality but still have some decorative common phrases showing up as noises.

Using **SegPhrase**, which integrates both considerations above and an extra classification, dish names of the best quality are mined. After understanding how the algorithm works, it was concluded that a classification step is needed to reduce irrelevant noise terms, based on either manual labeling or auto labeling using knowledge base.

Following are three tables shows the top dish names mined by three different algorithms. The analysis above is made based on such results.

dim sum	string beans	pork buns
Chinese food	crispy noodles	noodle soup
fried rice	salt pepper	hot pot
Chinese restaur	broccoli	dim sum
food was good	favorite dish	bbq pork
egg rolls	snow peas	shaved ice
pretty good	beef broccoli	pork belly
orange chicken	moo goo	peking duck
lunch specials	moo shu	chow mein
Panda Express	sesame chicken	beef noodle soup
	egg fu	

Table: Top phrases mined by different algorithms
From left to right: **ToPMine**, **Word2Vec**, **SegPhrase**
It shows that **SegPhrase** produced the best result.

2.1.4 Popular dishes and restaurants, review of Task IV and V

One of the most common functions of data mining is building a recommender system. Here based on Yelp's dataset, and the previously mined dish names, it is straightforward to consider building a recommender system by harnessing the mined dishes, and compute a score based on certain rules. Ultimately, the scores can be used to rank popular dishes (in one cuisine), and popular restaurants (provided the interested dish).

In task 4, I used following formula to compute the score:

$$\log(1 + occurrence) * \frac{(1 + usefulVoteCount) * rate * sentimentScore}{\sum(1 + usefulVoteCount)_i}$$

This score is computed using all reviews that mention one certain dish name. The *sentimentScore* here is computed using sum of sentiment scores of all sentences that mention the certain dish name in one review. The *log* term here is considered as a TF-like modifier to penalize too-frequently mentioned dishes.

The result of such computation is as follows, ranking Chinese cuisine dishes:

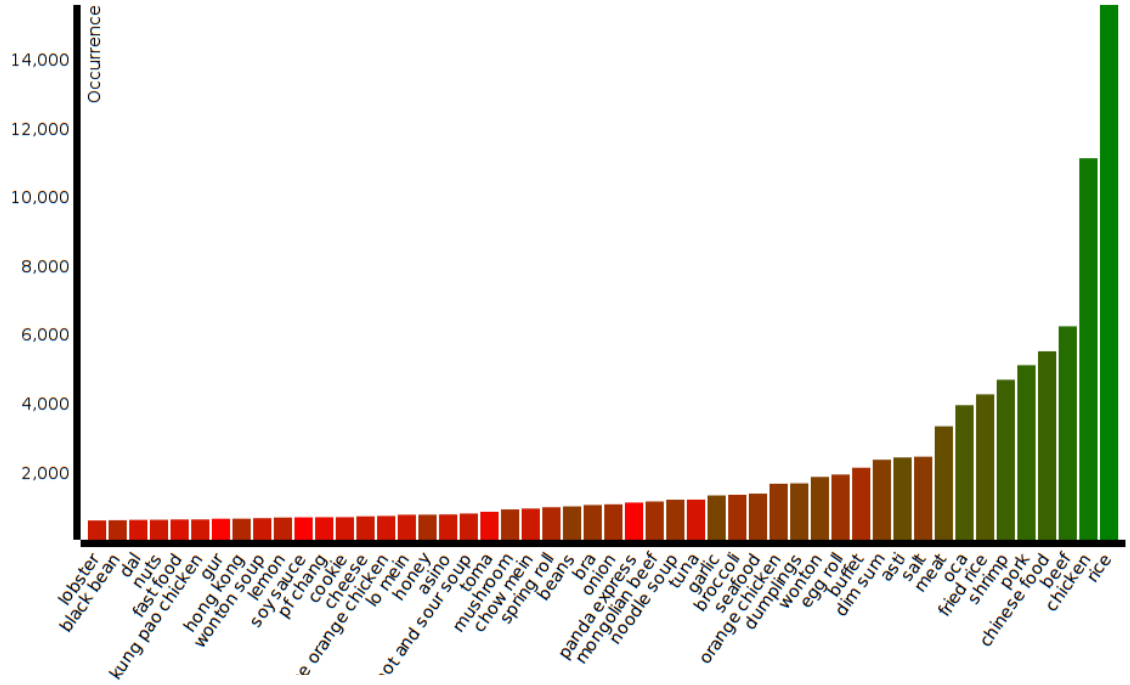


Chart: Dish rank

Y-axis is the occurrence.

Color-axis is the score, green is good.

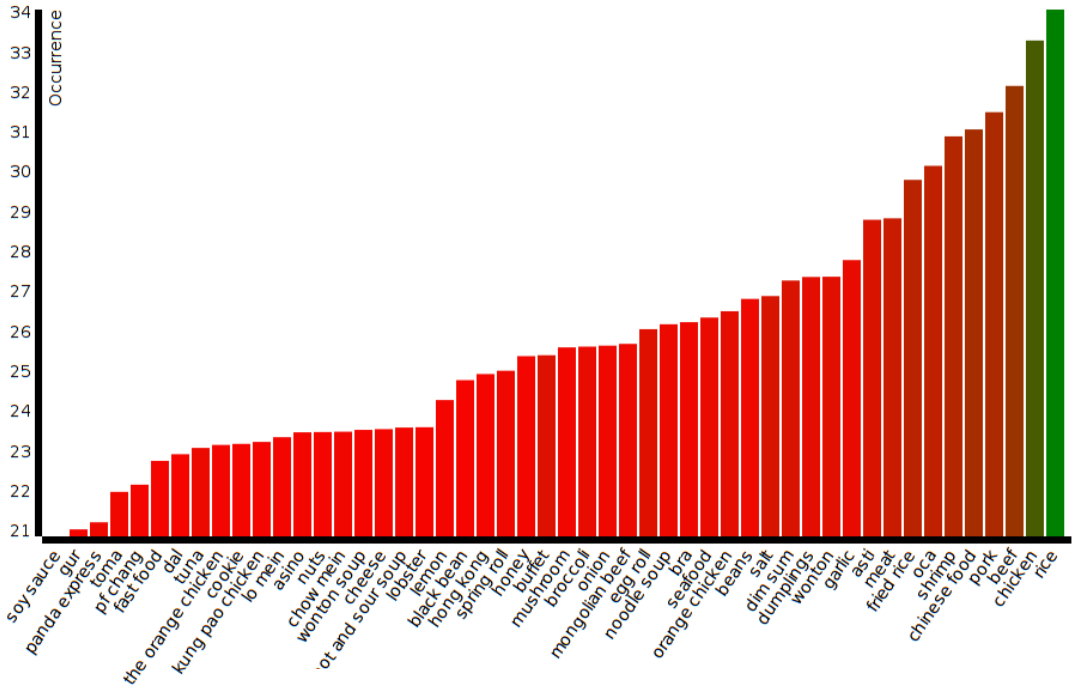


Chart: Dish rank

Y-axis is the score.

Color-axis is the occurrence, green is high.

It is shown that the score (the goodness of a dish) and the occurrence (the popularity) are generally correlated, though there are some disorder among any position of the ranks.

In task 5, similarly, using the same formula, but computed in a different way, a rank of restaurants with given dish name is computed. The computation is done by firstly segment reviews that mention the given dish name by restaurants. And the score is then computed upon the reviews under each restaurant.

The final result is shown below, with “Beef” as the interested dish.

Generally, popularity and quality still correlates, though even more obviously the disorder can be seen.

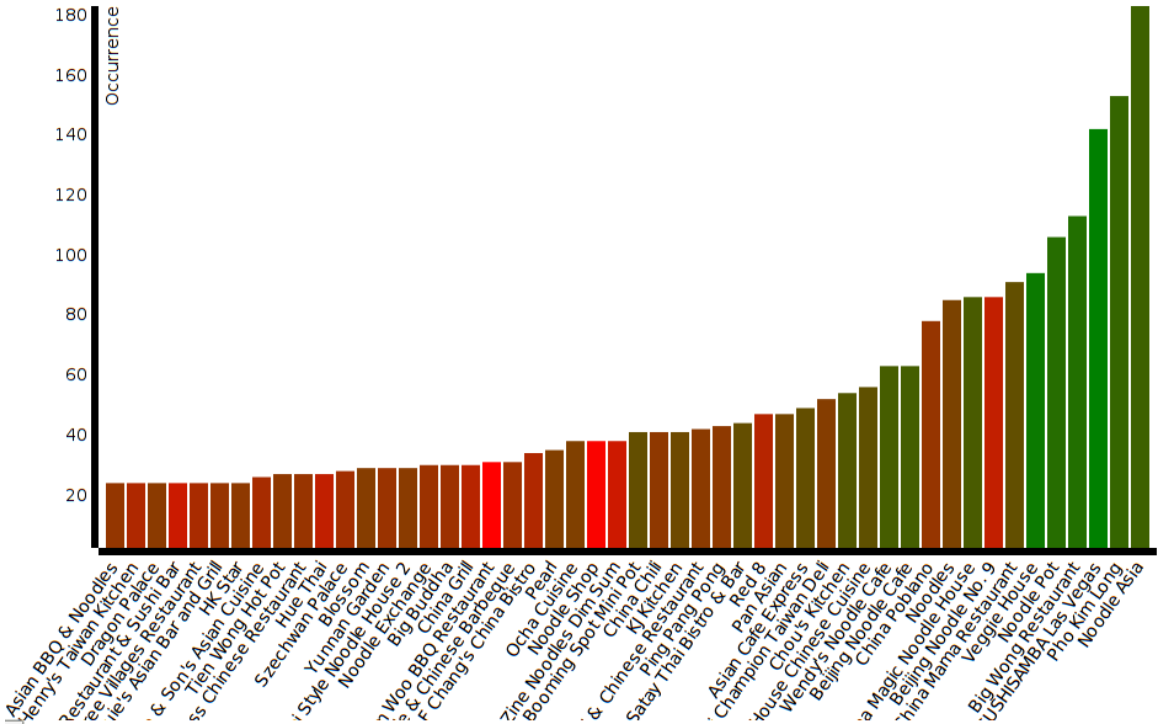


Chart: Restaurant rank, dish being “Beef”
Y-axis is the occurrence of “Beef”.
Color-axis is the score, green is good.

2.1.5 Hygiene prediction and review of Task VI

In this task, hygiene condition is predicted for customers to make dining decisions, using the information of review texts, restaurant cuisines, locations, rating and review count.

It is intuitive that people may not only talk about dishes, but also hygiene conditions in reviews. These mentioning not only include straightforward phrases (like dirty or tidy), but also some certain terms (like bathroom, it is imaginable that only when the condition is really bad, reviewer may want to mention it). Also cuisine types also should reflect the hygiene condition, which might be affected by how the dishes are typically made in the certain cuisines.

Algorithms

In this task, in terms of the classification algorithm, I tried multiple ensemble algorithms (SGDClassifier, RandomForestClassifier, ExtraTreesClassifier, GradientBoostingClassifier), and logistic regression (LogisticRegression), and blending of all above. Finally, the best performing result was produced by a blending of above (logistic regression with RandomForestClassifier and ExtraTreesClassifier with different parameters).

The blending approach is essentially a way to combine multiple independent classifiers by firstly computing the results of each classifiers by probabilistic figures, and secondly doing an extra step of higher-level classification, based on the probabilistic figures retrieved by previous steps.

Features

In terms of the features, I tried using the additional information only, the additional information with LDA topics extracted from uni-gram and bi-gram input, the additional information with top terms with high mutual information (w.r.t. the hygiene condition).

Before extracting the topics or computing mutual information, I preprocessed the texts by removing symbols like "&#\d+;", and case-lowering, stemming, and stop-words removal.

Results

I finally achieved a 0.5559 F1 submission score using blending of RandomForestClassifier and ExtraTreesClassifier without much parameter tuning, with features being 300 topics extracted by LDA and additional information.

It is worthwhile to note that with additional information alone, using the same parameters and same algorithm, I was able to achieve 0.551 F1 submission score already.

2.2 More Explorations

2.2.1 The questions

Generally, all these sub-tasks focus on providing suggestions to the customers. I would like to explore more in the perspective of the restaurants, by finding out what factors are mostly cared by the customers.

In particular, the fame of a restaurant is more likely to be affected by highly-influential customers. At the same time, when a reviewer in Yelp says that she is coming to the restaurant again, her words are more worthwhile and thus potentially more influential.

Therefore, I would like to explore the factors cared by highly-influential customers and return customers.

2.2.2 Factors cared by highly-influential customers

2.2.2.1 User network of Yelp

Yelp has its own SNS (social network service) which enables customers to add friends and thereby form a user network. In this paper, it is suggested that such a network is built only to share information. Therefore, highly-influential customers in such a network potentially have more effect on a restaurant's rating.

If we explore how the network of Yelp looks like, we may find the degree number (defined as the number of friends one has) is distributed as follows:

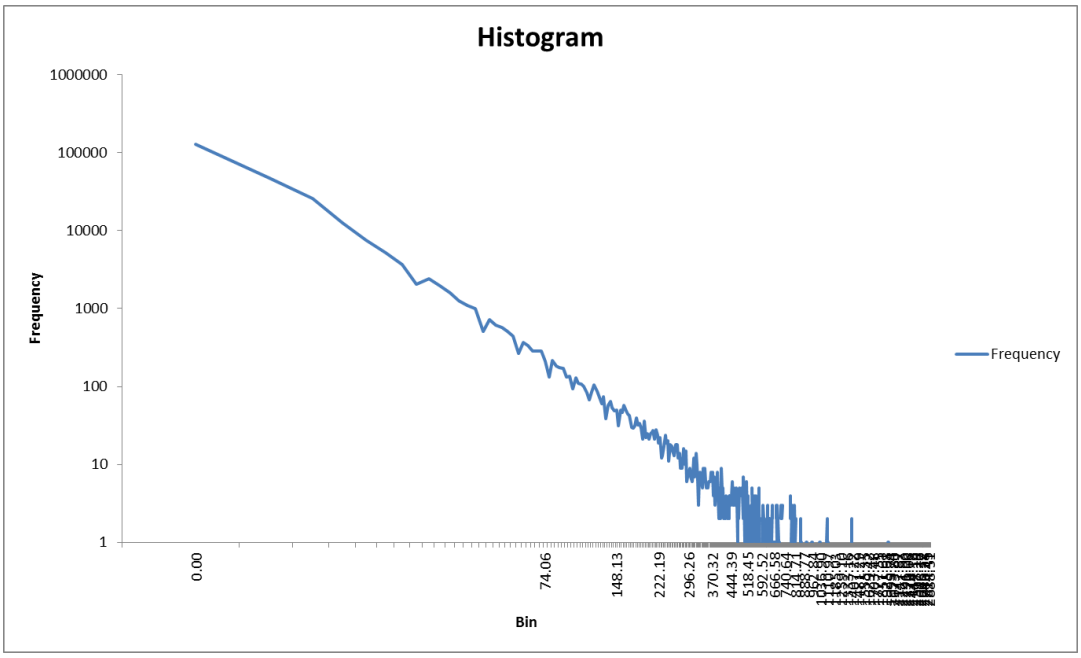


Chart: User degree distribution of Yelp SNS
It is a power-law network, i.e. small-world network

As is shown, this network is a small-world network, which has few high-degree hub users. These hub users are considered highly-influential in this report, since they are essential to passing information from one point (community) to another among the network.

And following chart shows the review count of these hub users (degree > 200), which is highly different (of more counts than) from the samples of all users:

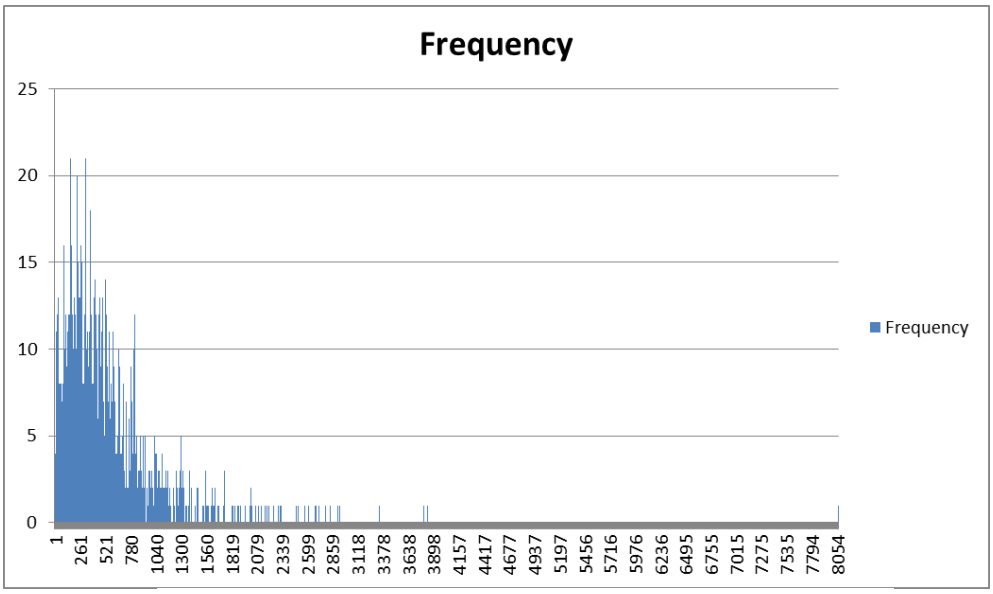


Chart: Review count of hub users (degree > 200)

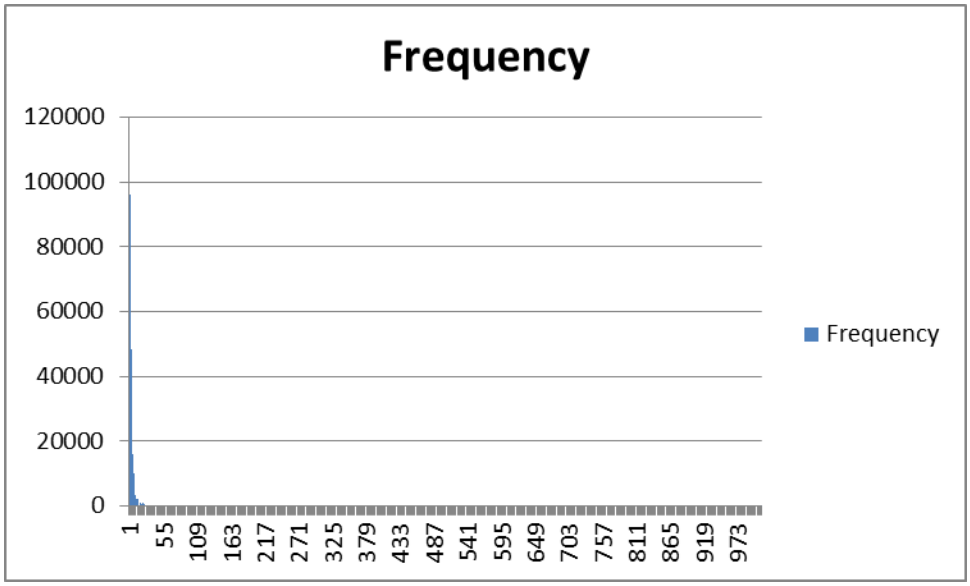


Chart: Review count of all users

So from these charts, it shows that hub users write more reviews. Therefore with more reviews from the same person, it is also intuitive that the focus of such users may be better extracted, since they provides more samples, provided that one person's focus does not change.

2.2.2.2 Topics of hub users

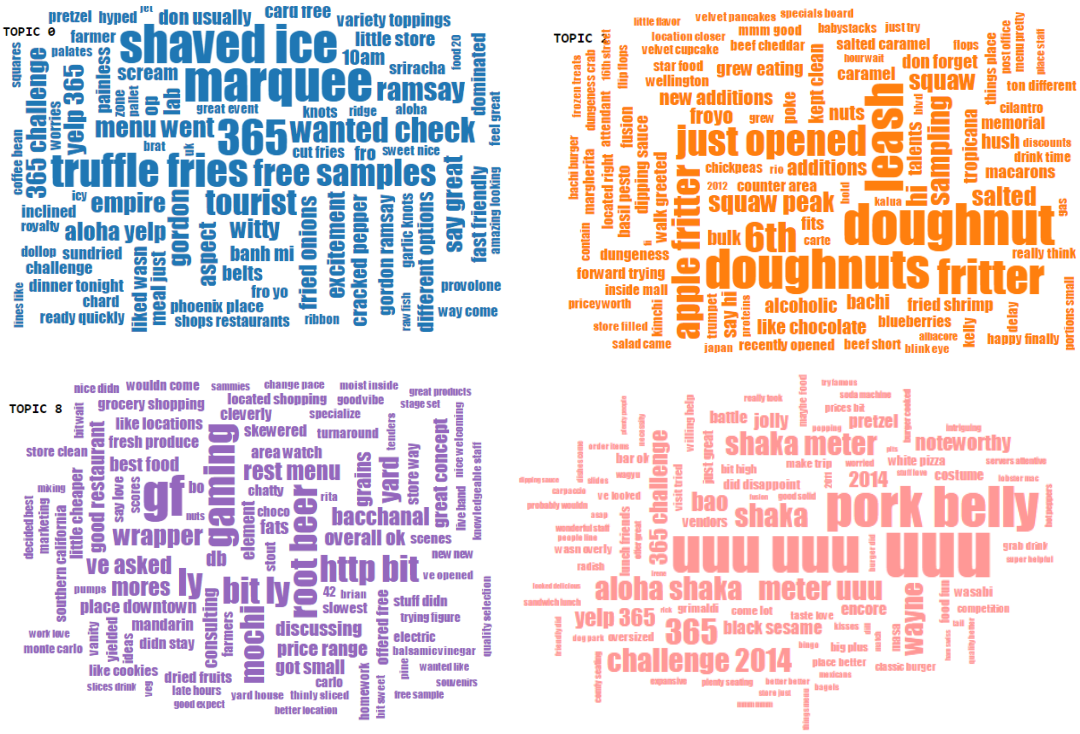


Chart: Review topics of hub users (4 out of 10)

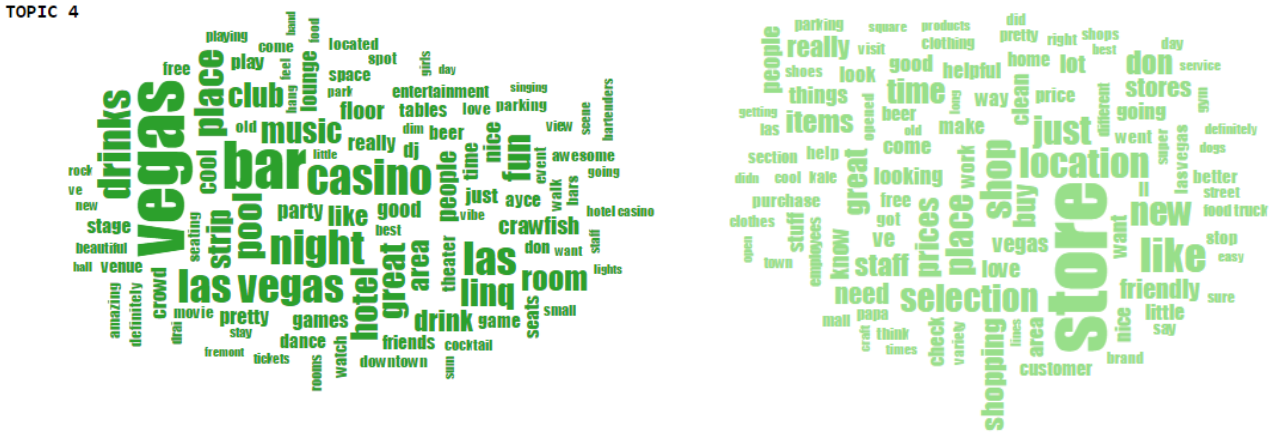


Chart: Continue: Review topics of hub users (2 out of 10)

So above are the word clouds of six topics out of ten extracted topics, using LDA and TFIDF with bigram and unigram, from all the reviews of hub users (degree > 200).

Empirically, fewer topics about dish names are seen. As a replacement,

- "Yelp 365 challenge" is a very frequent term in one of these topics, showing that this particular activity is indeed a strong boost to review count from users who are interested. So, **advice to restaurants: cooperate with Yelp, or similar services that provide such a review platform.**
- "just opened" is very frequent in one topic, showing that a restaurant possibly should grasp the precious time when it just starts to serve, because that's when hub users may come into. **Advice to restaurants: build good impression from the beginning.**
- Peripheral services such as "bar", "casino", "store", "tourism" are also frequent in multiple topics, showing that they may be reasons for hub users to come**. Advice to restaurants: do locate in popular areas.**
- "bit ly" mostly indicats that users are posting pictures. So, **advice to restaurants is: make things beautiful!**

2.2.2.3 Return users in Yelp

Return users are more likely to better reflect a restaurant's quality.

Following is the distribution of return reviews in Yelp (review made by the same user on same restaurants for more than once):

Chart: Return customers.
X-axis shows how many times one user reviewed on one restaurant. Y-axis is the count of such users.

So there are quite abundance of return users in Yelp, and therefore the data mined from such data subset should be meaningful.

2.2.2.4 Topics of return users

It is beneficial to firstly understand what return users are talking about.

Below are word clouds of five topics out of ten extracted topics, using LDA and TFIDF with bigram and unigram, from all the return reviews (including the first review of a return reviews set).

It is notable that although some of the topics are still about dishes, a higher portion of the topics are mentioning **service, portion**. And interestingly (and maybe sadly), one special topic is about "closed", indicating that users want to go to the restaurant but it is closed.

Figure: Return customers' topics (5 out of 10)

2.2.3 Summary

In this set of exploration, I focused the topic mining onto a subset of customers who are important in terms of influences, which are supported by analyzing the data from the Yelp data set.

Result shows that, other than analyzing all customers, focusing on such important customers returns a very different focus of topics, including Yelp activities, restaurant opening time, peripheral services, general services, and so on.

3 Project Highlights

In this capstone project, I have completed 6 sub-tasks and extra explorations of key customers' cares. The information retrieved from these sub-tasks are helpful to the customers, while the information retrieved from the extra explorations are helpful to the restaurants.

3.1 Usefulness of Results

The result of the 6 sub-tasks are very helpful to the users, since essentially these information retrieved includes dish popularity and goodness (task 4 & 5), and restaurant hygiene condition (task 6). In fact, intuitively these factors are what customers care about very much when they make dining decisions.

While the result of the extra explorations provide the information of what influential users, who have abundant connections on Yelp's social network service, really care about, what pushed them to come to the reviewed restaurants, and what return users care about. Since the fame of a restaurant is generally built up by its customers, and such customers are influential among their networks, retrieving such information are very helpful to restaurant owners to make better judgment to attract more people coming in.

3.2 Novelty of Exploration

Throughout the sub-tasks, multiple new techniques were used to do the explorations:

- In Task 6, when doing feature engineering, multiple methods of extracting features were tried, including LDA topic extraction, filtering top terms with high mutual information, and word stemming. In particular, using top terms with high mutual information only boosted the train F1 score by 0.07, from 0.61 to 0.68, while by 0.004 on the test set.
- In Task 6, the classification algorithm was modified to use blending technique, which essentially does a 2-fold classification, by firstly doing classifications using different independent classifiers, and then doing higher-level classification using the results provided by earlier classifiers. The result, without any tuning of parameters, achieved 0.5553 as the submit F1 score.
- In extra exploration in the final report, harnessing the knowledge of social network analysis helped to narrow down the exploration data space, into focusing on key customers who are influential among their networks. Final result, as analyzed by topic extraction, shows that such a group of users are indeed talking about different things, compared with the general mass. These newly emerged topics supply a good source of factors that restaurant owners may want to take care of, if they want to attract influential customers.
- In extra exploration in the final report, by better understanding the input data, return customers are identified as the key customers as well, thereby a subspace of review samples can be extracted. Topic analysis upon such review corpora indicated that these people are talking about things different from the mass as well. Suggestions are thereby given to the restaurants to attract return customers.

3.3 Contribution of New Knowledge

In this final report, as well as earlier reports of sub-tasks, following new knowledge can be learned:

- It is advisable to firstly filter features, by means such as computing the mutual information between the features and the labels. Results in Task 6 shows that using mutual-information-filtered top words contributed to higher train set F1 scores.
- Key customers that have high connections in their networks, are usually doing much more reviews than the general mass, and according to the network theory, they are more influential. Studying their topics reveals a brand new scope of focuses compared with the mass, which can be supplied to restaurant owners to improve their services, so as to attract these people.
- Key customers that return are also talking about different things as compared to the mass. And it is also advisable to restaurant owners that services are important to attract people coming back.

- In a higher level, with a dataset provided with network information, it is advisable to data analyzers to harness such data to either narrow down the samples, or use them as new features when doing feature engineering. The network information can be harnessed as it is (as in this final report demonstrates), or built up artificially using extra information (such as a network of restaurants co-reviewed by customers).

4 Conclusion

In this final report, a review of earlier sub-tasks is provided, in addition to extra explorations made upon key customers.

All these tasks and explorations harnessed multiple data mining techniques, including topic mining, clustering, feature extraction and engineering. In addition to these, external knowledge such as network theories is applied.

Results from the sub-tasks provide a very good insight of what people are talking about in Yelp reviews, and ranks of dishes in cuisines, ranks of restaurants of certain dishes, and ranks of hygiene condition inferred by review corpora. These pieces of information are helpful to customers to make dining decisions.

Results from the explorations in the final report demonstrate that key customers can be extracted by their significance in network, and their review values. Also topic mining upon these samples reveals that they do focus on things other than dishes. Such information are useful to the restaurant owners if they want to attract such key customers.

5 Acknowledgement

In addition to the reference documents listed below, I'd like to thank all the professors and TAs offering this series of course: Professor Han, Professor Zhai, and Professor Hart. The courseware and the assignments are really very helpful for me to build up data mining skills from the scratch. The final project is also awesome, in that I really gained much from hands-on experiments doing data analyses.

Also I'd like to thank all the colleagues, especially those who help to review my reports, and those who are very active in the forum. Your sharing really helped me a lot, though I do not talk much on the forum^_^

I know that this report is not the best I can do, due to the time constraints. And there are lots of findings only residing in my computer python codes. What I do want to express here is that this experience is awesome itself, and I do treasure it very much.

Finally, BIG thanks to Coursera for providing such a helpful platform for further studying beyond my college life!

References:

- "Mining Quality Phrases from Massive Text Corpora" Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, Jiawei Han, University of Illinois, Urbana Champaign.
- "On the Efficiency of Social Recommender Networks." Felix W. Princeton University.
- "Oversampling with Bigram Multinomial Naive Bayes to Predict Yelp Review Star Classes" Kevin Hung and Henry Qiu, University of California, San Diego.
- "Clustered Layout Word Cloud for User Generated Review." Ji Wang, Jian Zhao, Sheng Guo, Chris North. Virginia Tech and University of Toronto.