Data Mining Project Final Report

1 Abstract

In this final report, I will review all the six completed tasks throughout this project. A summarizing that contains useful information also be provided by mining data after reviewing all required tasks.

1.1 Exploration of Dataset, and environment

1.1.1 Exploration of Dataset

The dataset that is used in this project comes from Yelp (*Yelp Dataset*, 2018). It contains a lot of text data such as reviews of restaurants, restaurant information, and users' information. In contains approximately 1.6 million reviews, more than 500000 tips and attributed for around 481000 business attributes. In this project, I only use the reviews and business attributes datasets.

1.1.2 Environment for this project

Windows 10 is used as the operation system for this project. All data processing, analyzing and visualization tasks are completed using python 3.9. Python packages are used to help preprocessing data and get result. NLTK package is used for preprocessing data such as tokenizing and stemming data. Gensim and sklearn packages are used to help analyzing data. Altar, a python visualization tool, are used to generate almost all visualization for the projects.

2 Project Activities

2.1 Task 1: Topic Mining

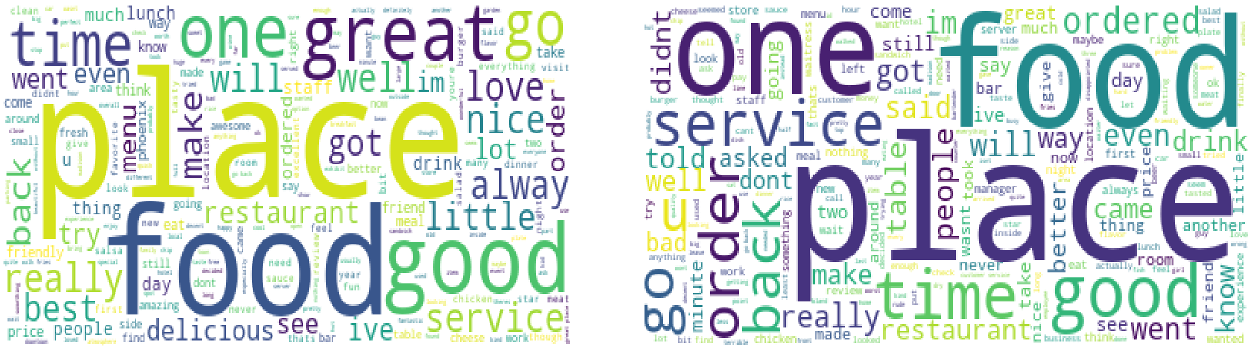
It is important to get a general understanding of the data before going deeper understanding and dig more important information from it. In this task, LDA topic model was used to extract information from all the reviews. Besides, I compared all topic words based on review rating. I extract 10 different topics from review data, each topics contains 15 words.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
place	museum	food	food	service	like	place	food	store	room
great	exhibit	sauce	good	time	just	great	good	good	car
food	indian	good	like	told	time	phoenix	place	like	hotel
good	exhibits	place	place	customer	place	time	great	really	stay
service	science	chicken	just	just	dont	game	fish	place	clean
love	buffet	fresh	cheese	called	people	like	mexican	selection	just
really	native	bbq	got	like	im	garden	like	just	like
friendly	american	meat	ordered	phone	did	desert	chicken	little	nice
staff	kids	thai	order	work	said	really	salsa	great	rooms
like	center	great	service	minutes	didnt	fun	chips	dont	staff
time	naan	like	time	did	know	nice	just	time	did
nice	adults	just	really	said	hair	just	really	madison	service

just	body	pork	dont	pizza	ive	bar	service	grocery	work
ive	planetarium	best	didnt	make	really	good	best	know	im
breakfast	art	delicious	came	know	going	center	sauce	shop	free

Figure 1. Topic Mining for all review data

After analyzing topics based on all review data, I make topic mining using LDA based on review ratings. All reviews are sorted into 2 parts based on their ratings. Those reviews that get ratings larger than 3 are sorted as high-rating records. Other ratings are sorted as low-rating records. After sort all ratings, I make a LDA analysis. Besides, several word-clouds are made based on results I get. Based on the results, we could found out that compared with high-ranking reviews, low-ranking reviews more focus on service and order.



Word cloud for high-ranking records

Word cloud for low-ranking records

Figure 2. Word-cloud for different ranking records

2.2 Task 2: Cuisine clustering and review

Most of the restaurants are categorized by cuisines. The target of this task is to find out similarities of different cuisines and clustered topics using different algorithms based on reviews the restaurant received.

I pick 50 categories from the given sample set of reviews randomly. After preprocessing data using proper methods, I count the similarities between selected categories using cosine distance method. To improve the cuisine results, TF-IDF vectorizer and LDA model was used on the review texts of each category, and then computed cosine similarity between categories with the vectors returned by LDA model. In this part, 10 topics are generated via LDA model.

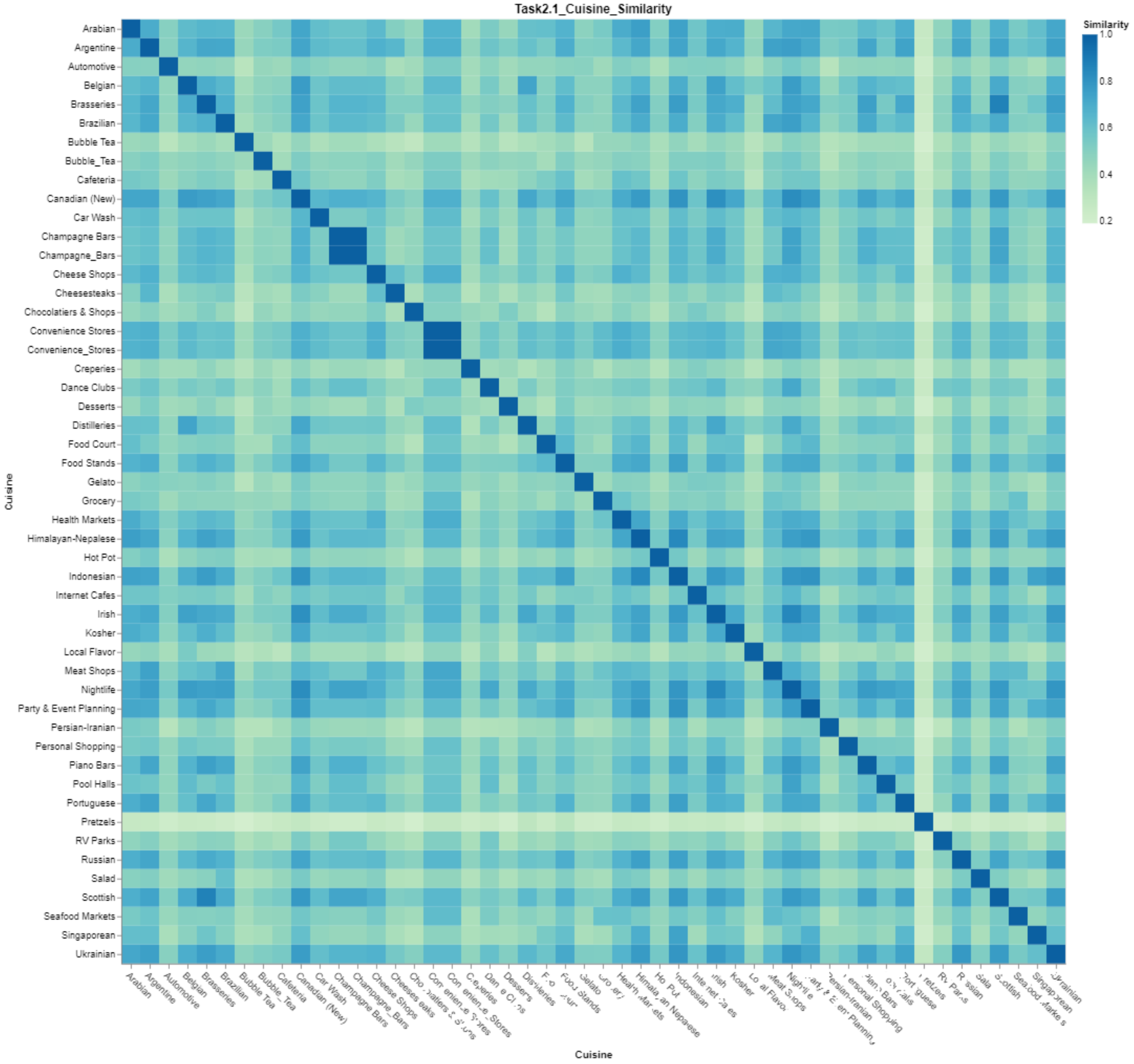


Figure 3. Cuisine Similarity using cosine distance method

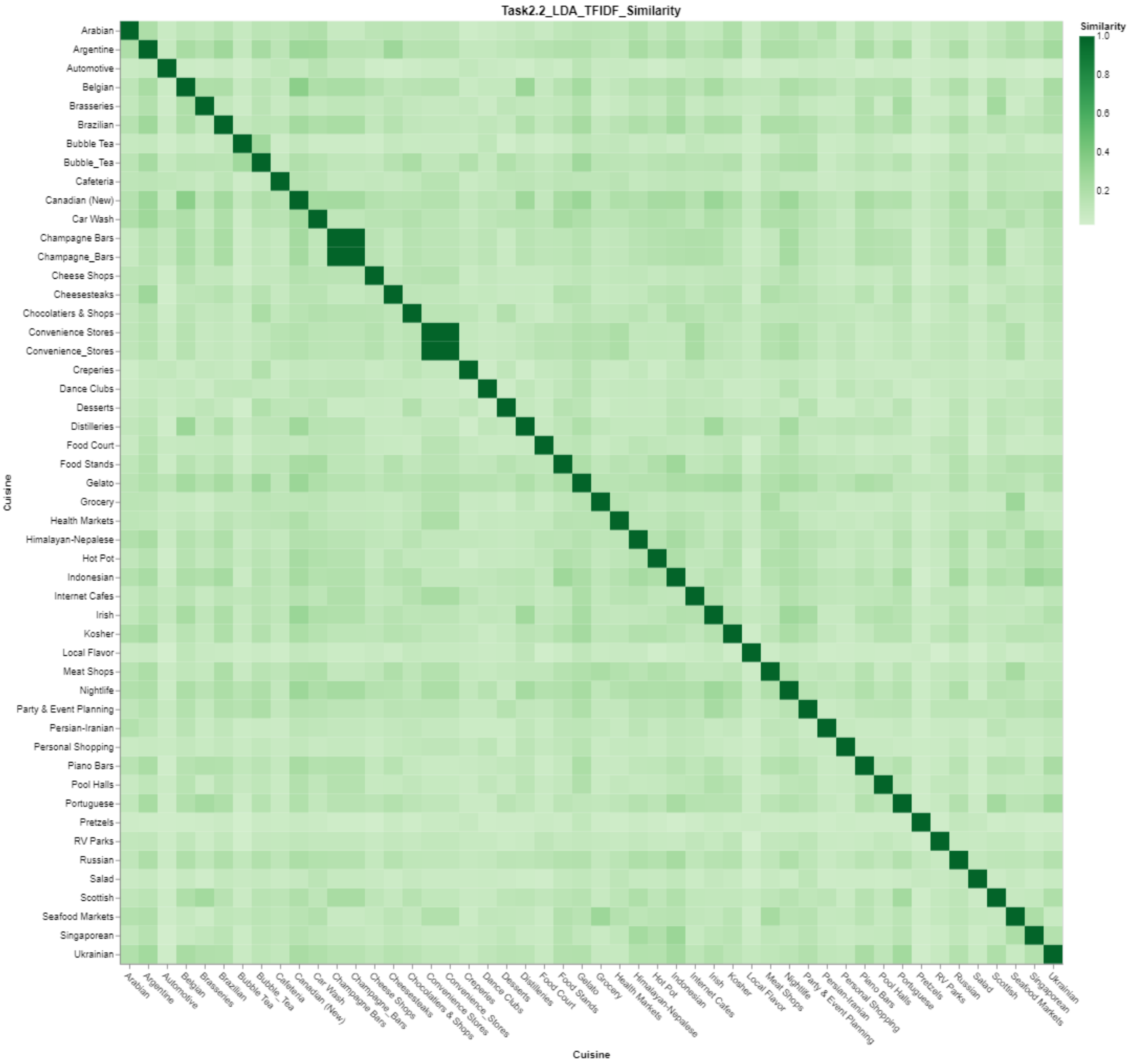


Figure 4. Cuisine Similarity using TF-IDF methods

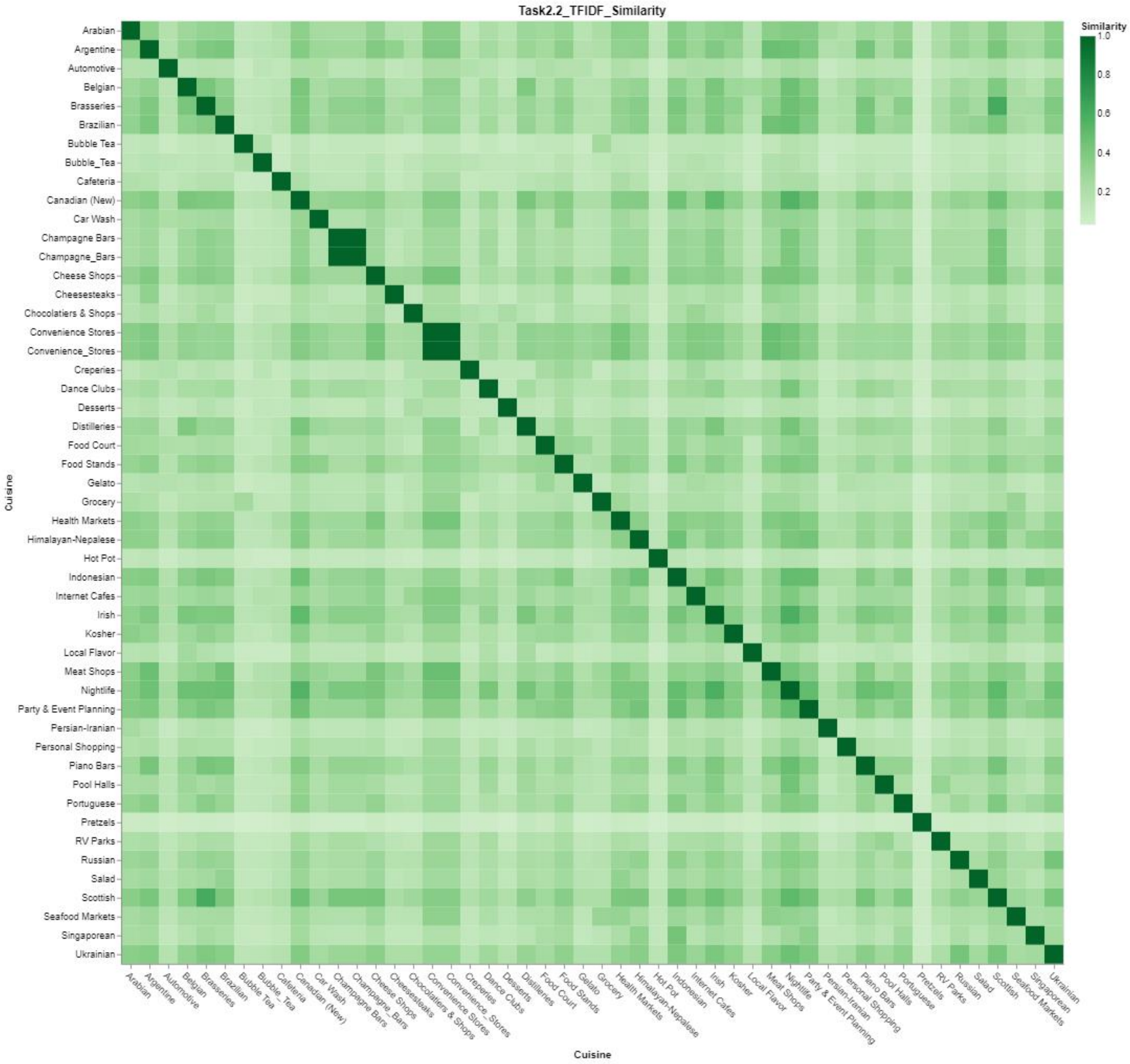


Figure 5. Cuisine Similarity using LDA_TF_IDF methods

Finally, to compare the function between different clustering algorithms and different clustering numbers, I use LDA and LSA to generate different number of topics and compared then using grid heat map.

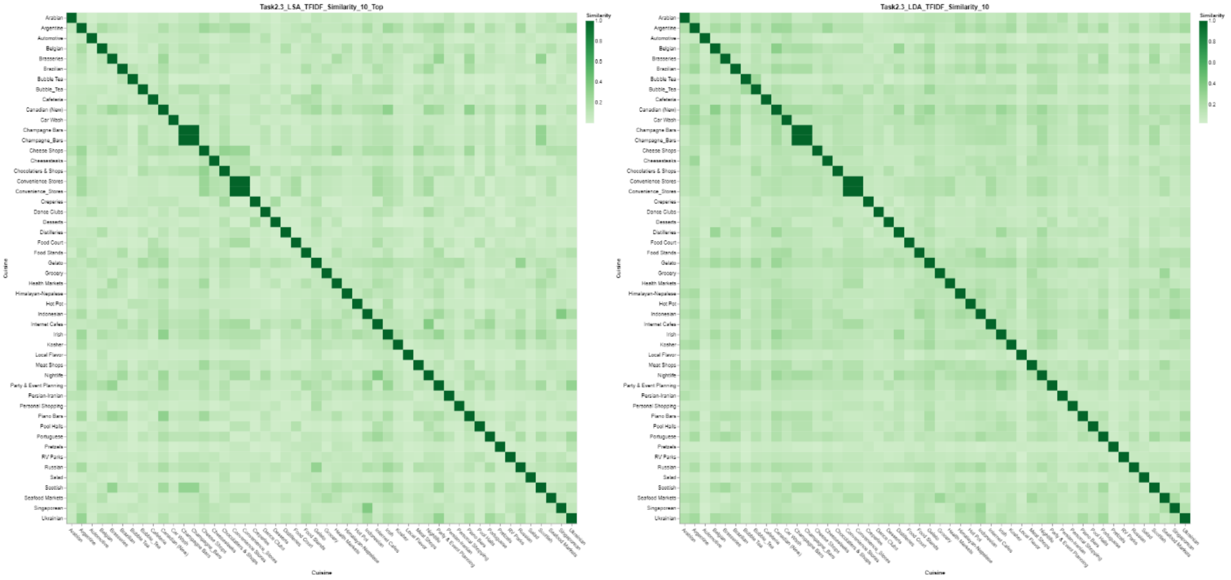


Figure 6. 10 Topic similarity using LDA and LSA

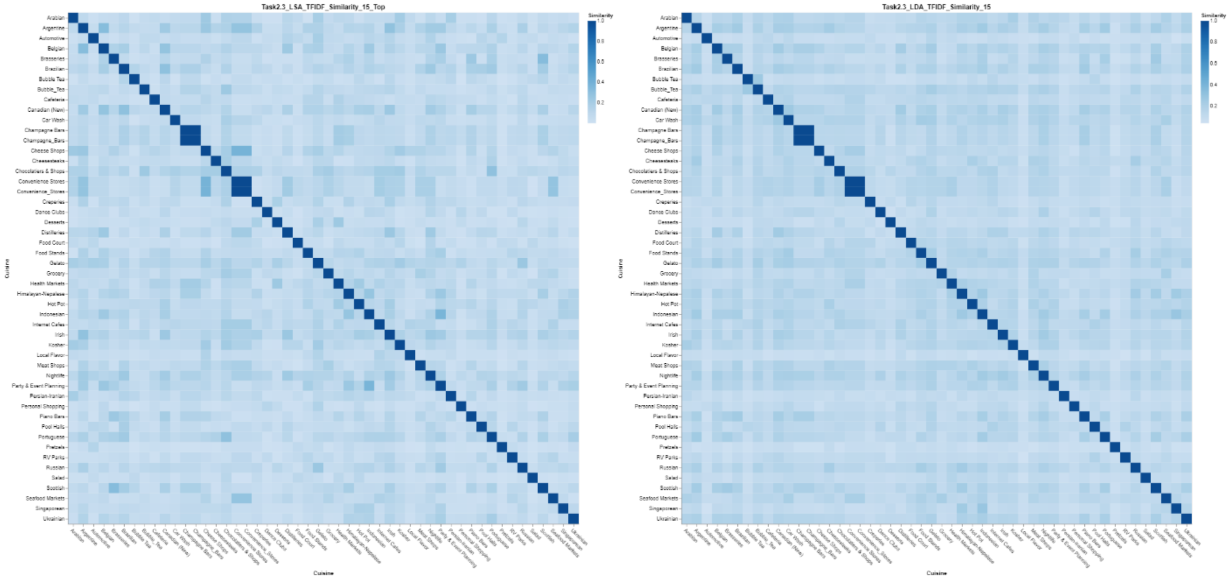


Figure 7. 15 Topic similarity using LDA and LSA

2.3 Task 3: Dish Recognition and Review

This task requires to extract dish name from reviews of a certain cuisine. I choose extract names for Chinses cuisine. I use Word2Vec in this task to mind dish names in this section. The top phrases show better quality.

Stir fire 0.39408854

Bell pepper 0.39146996

Cream cauce 0.37390384

Garlic sauce 0.37300736

Clay put 0.34477097

Bean sauce 0.34379196

Thai style 0.33260161

Singapore noodle 0.31919032

Black pepper 0.31781009

Melted mouth 0.30877316

Figure 8. Top phrases mined by Word2Vec

2.4 Task 4 & 5: Popular dishes and restaurant

Tasks 4 and 5 focus on finding the best dishes and restaurants in a selected category. Like task3, Chinese cuisines is selected as my research target. Based on Yelp dataset and cuisine names provided, reviews that contains target name are aggregated and counted their average ratings.

Task 4 shows the top 100 mentioned dishes in Chinese Restaurant. X-axis shows the name of each cuisine, y-axis shows the times this dish is mentioned in all the reviews. Colors shows the average ratings of each dish. The lighter the color, the higher the average rating of the dish.

It is easy to find out fried rice, dim sum, and egg roll are the top 3 most popular dishes in Chinese cuisines. Besides, the average ratings of top 10 most popular Chinese dishes are higher than 3.4. One of the interesting findings of task 4 is about soy sauce. Soy sauce is one of the most important condiments in Chinese cuisine. What surprises me is people mentioned it in reviews. Moreover, most reviews mentioned soy sauce have relatively lower ratings than other reviews.

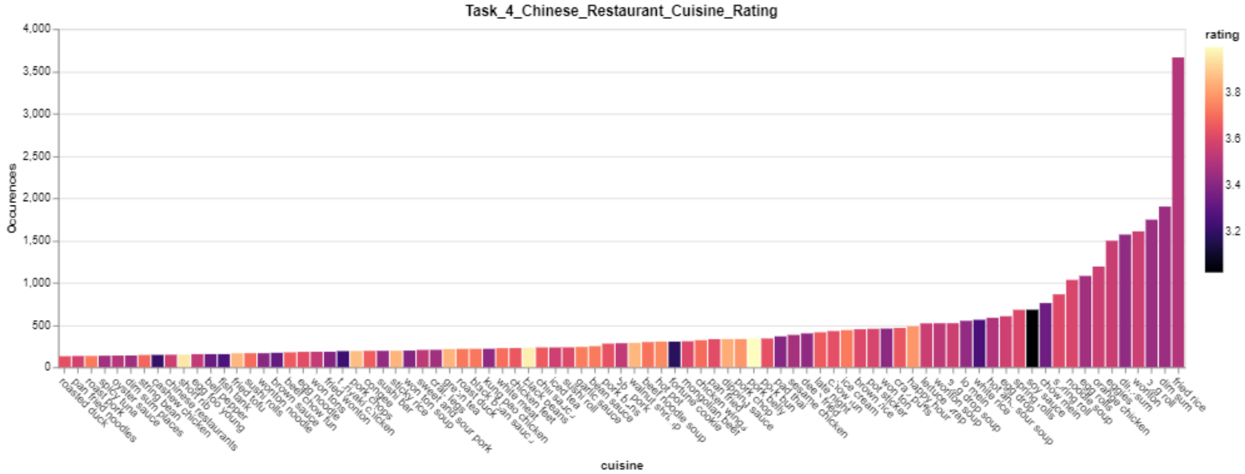


Figure 9. Chinese Cuisine Rating

Task 5 shows the 50 popular Chinese restaurants. X-axis shows the name of each restaurant. Y-axis shows the times this restaurant is mentioned in all the reviews contains fried rice. Colors shows the average ratings of each restaurant. The lighter the color, the higher the average rating of this restaurant.

The reviews towards Chino Bandido mentioned fried rice the most times. Besides, the average restaurant rating of Chino Bandido is around 4, a relatively high rating. We may be able to infer that Chino Bandido offers tasty Chinese fried rice. Besides, Sam Woo BBQ Restaurant may serve nasty Chinese fried rice since they got a low average rating.

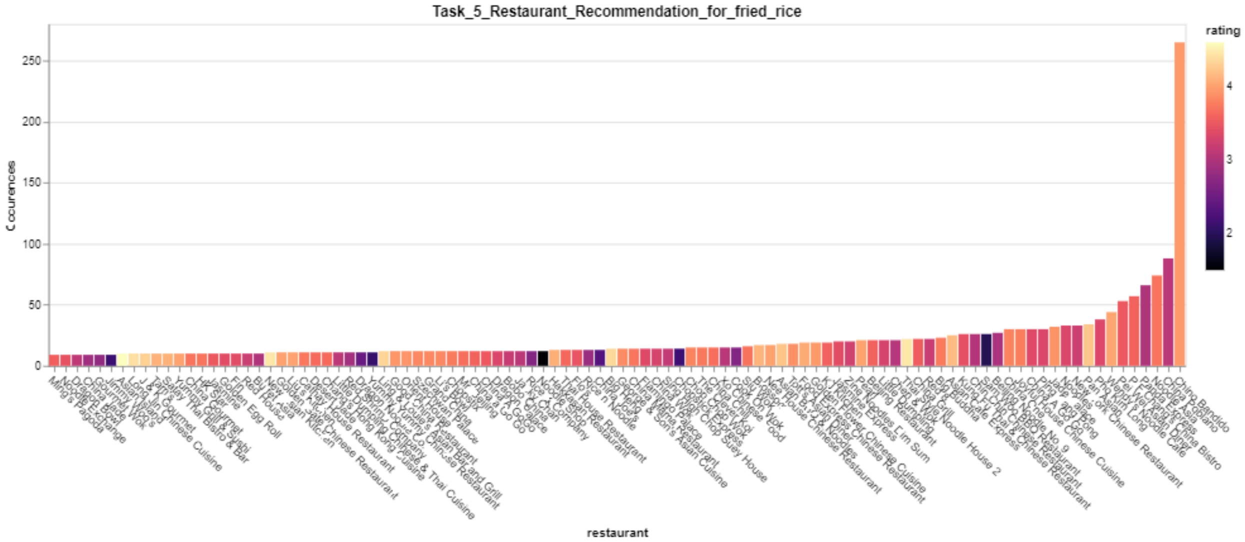


Figure 10. Chinese Restaurant Rating

2.5 Task 6: Prediction and restaurant sorting

The results whether a restaurant would pass health inspection are predicted in this task. The factors such as information of review texts, restaurant cuisines, locations, rating and review numbers are used to help predict.

In this task, I tried 7 different classification algorithms, including LogisticRegression, SVC, GaussianNB, KNN, DecisionTree, RandomForest, and XGB. I use all additional features including cuisine types, location, number of reviews, and the average rating to generate classification model. Cuisine types and location are preprocessed as mentioned in Data Preprocess part. After getting f1-scores of each algorithm, I visualize them in a plot. X-axis shows the test size set when train the model. Y-axis shows the scores of each algorithm.

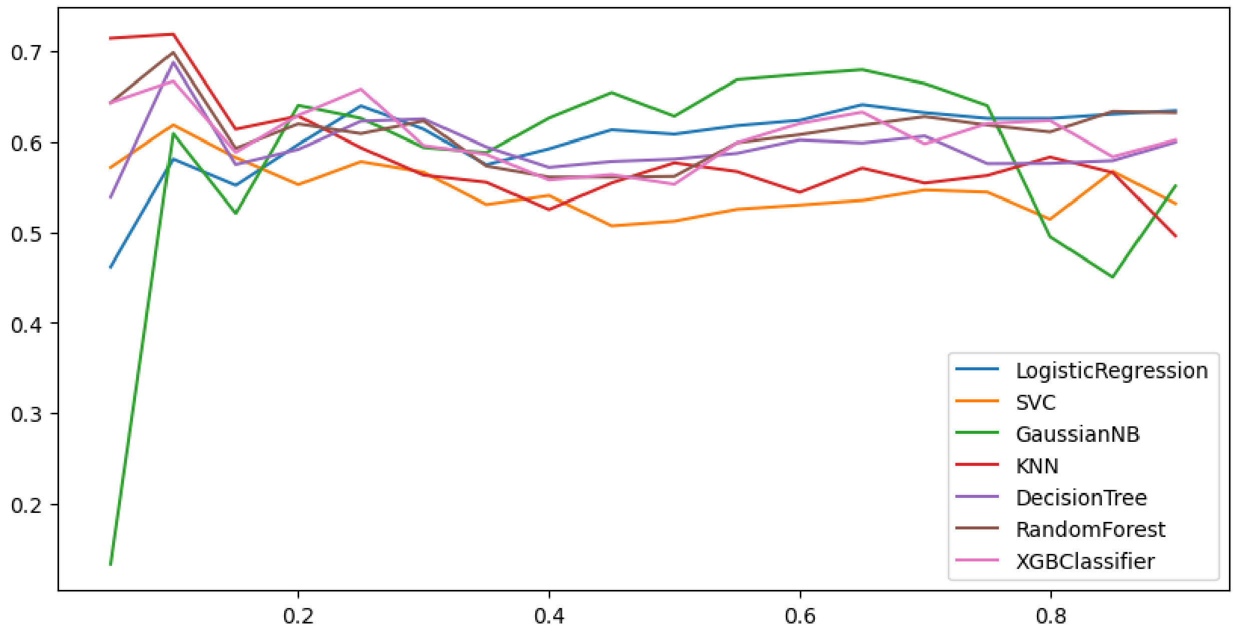


Figure 11. F-1 scores of each algorithm used in this task

As can be seen, GaussianNB has a better performance in this task when we set test size between 0.4 and 0.6. Besides, KNN have the best performance in all selected algorithms when test size is set as 0.1.

2.7 Project Highlights

This project contains 6 sub-tasks of customer cares. The information retrieved form these sub-tasks are helpful to the customers, restaurants, and rating websites.

2.7.1 Usefulness of Results

Task 6 is the most useful sub-tasks I produced in this project. Both restaurants owner and users would benefit from this task. Customers would make a wiser dining decision based on provided information such as reviews comment, ratings, and even the location and cuisine. Restaurant owner could know the demands and reviews from internet. Such a retrieving information are helpful to restaurant owners to make better judgement to attract customers and pass their healthy inspections.

2.7.2 Novelty of Exploration

The classification algorithms in task 6 was modified to get better performance. Apart from normal classification such as logistic regression and decision tree, a new classification algorithm, XGB, is adopted.

2.7.3 Contribution of New Knowledge

Filter features and use the most appropriate features would help get better performance. In task 6, using filtered features would have higher train set F1 scores.

3 Conclusion

In this report, a review of sub-tasks is provided. All these tasks and explorations covers multiple techniques such as topic mining, clustering, feature extraction and engineering.

Results from all sub-tasks shows the features and characteristics of reviews people published in Yelp reviews. These information are helpful for customers and restaurants to receive and provide high quality restaurant service and delicious dishes.