Task_4&5_Report

Introduction

This report includes two parts. Task 4, the first part of this report, shows the 100 frequent mentioned Chinese cuisines on Yelp cuisine dataset. Task 5 shows all Chinese restaurants in dataset which get reviews about 'fried rice'.

Python is used in these tasks. Panda's package is used to process data. Nltk is used for tokenizing and stemming original data. Altar is used to visualize my results.

Steps

Data Sets

Three different datasets are used in these tasks. 'yelp_academic_dataset_review.json'(review datasets) dataset provides ratings and reviews. 'yelp_academic_dataset_business.json'(business datasets) is used to get restaurants' name for task 5. I uses the Chinese Cuisine list I got from task 3 as the cuisine list for task 4.

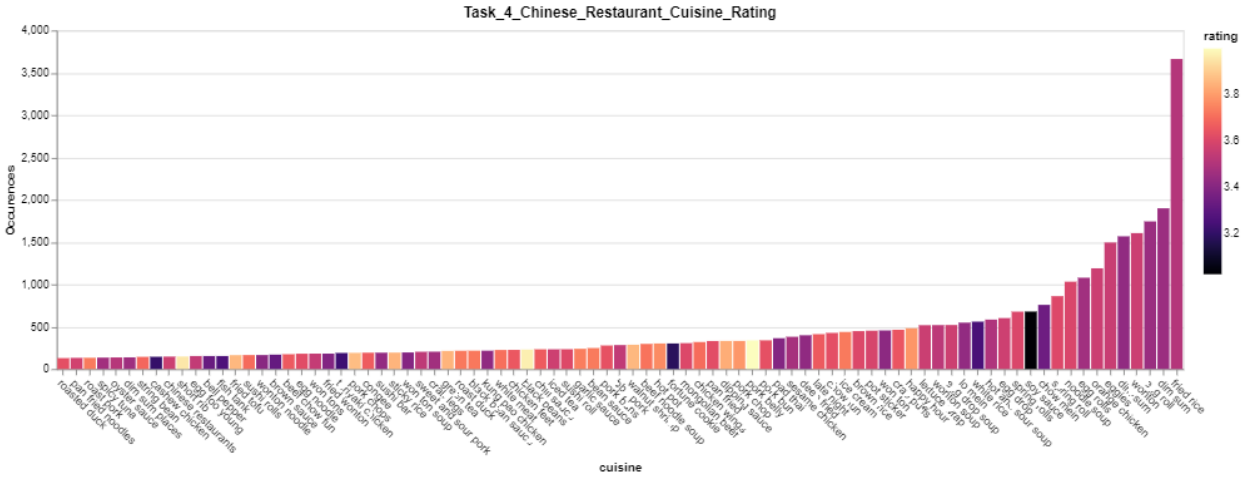
Data preprocessing

After getting data from datasets and preprocess data using nltk with the same steps as Task2 and Task 3. I get 3 cleaned text data sets stored in Pandas. Since my research focus on Chinese restaurant and Chinese cuisine, I selected restaurants whose categories contains 'Chinese' from business dataset. Cleaned review datasets and business dataset is combined using merge function provided by Pandas. After deleting irrelevant columns, I finally got a dataset with only 3 columns, restaurant name, review text, and rating. Each rows represent a specific review. The dataset demo is as following:

	stars	text	name
0	5	I really like both Chinese restaurants in town...	Chang Jiang Chinese Kitchen
1	3	Above average takeout with friendly staff. The...	Chang Jiang Chinese Kitchen
2	4	We order from Chang Jiang often and have never...	Chang Jiang Chinese Kitchen
3	3	Good enough for carry-out in McFarland on a co...	Main Moon Chinese Restaurant
4	5	Best Chinese food madison! I've tried them all...	Main Moon Chinese Restaurant
...	...	...	...

Task 4: Mining Popular Dishes

Task 4 Visualization:



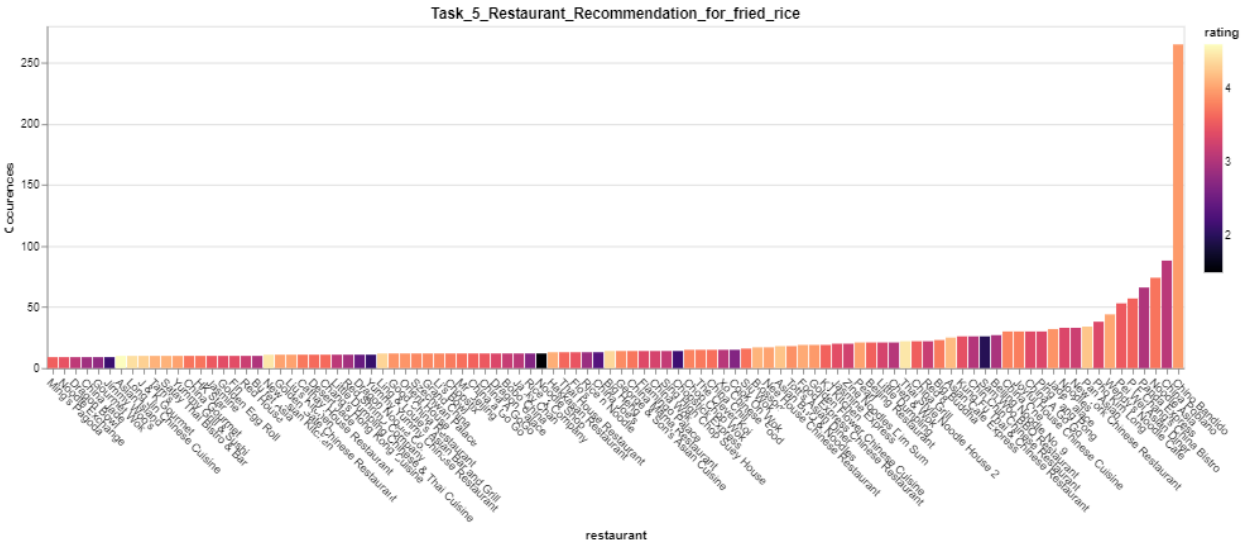
Explanation and Findings

This visualization shows the 100 mentioned cuisines 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.

Task 5: Restaurant Recommendation

Task 5 Visualization:



Explanation and Findings

This visualization 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.

Conclusion

I made popular Chinese dishes mining and Chinese restaurant recommendation in this report. Fried rice, dim sum, and egg roll are the top 3 most popular dishes in Chinese cuisines. Chino Bandido might be a idea restaurant to try Chinese fried rice. Though I get clear conclusion in these two tasks based on my own research, the solutions may not be useful in some extreme cases. My research is based on large volumes of data. There is a possibility of large deviations when the amount of data is insufficient. The further research needs to focus on cuisine mining and restaurant recommendation with small samples.