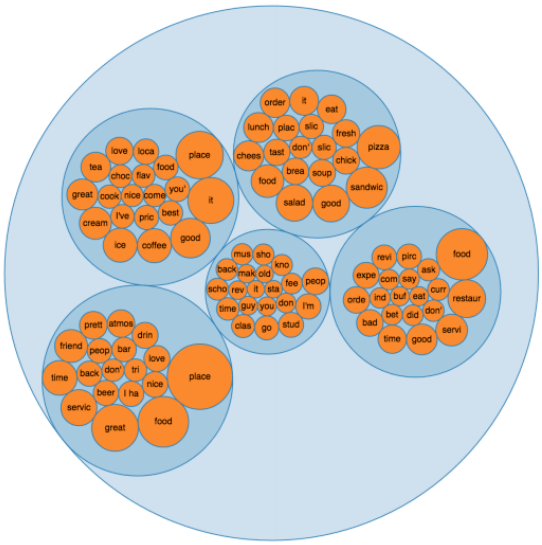
Data Mining Capstone

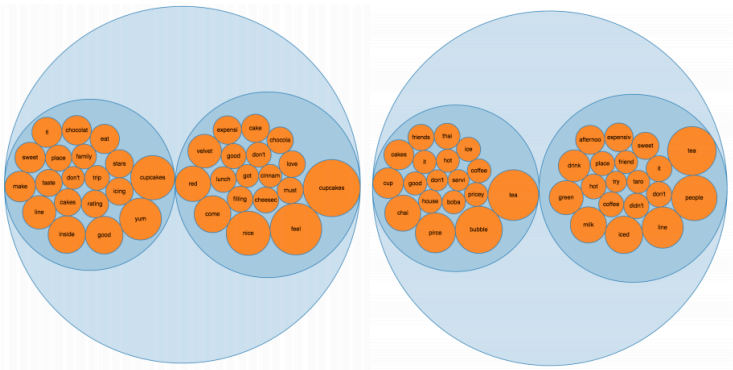
Final Report

- **Summary of Specific Tasks:**

  - **Task #1**
    In task one, restaurant topics were generated from the review text by PLSA in Meta. For example, the upper right circle has ‘pizza’, ‘salad ‘and ‘sandwich’ appearing together. And we can also see that the major things that people care about are the quality of the food, location, service and atmosphere.



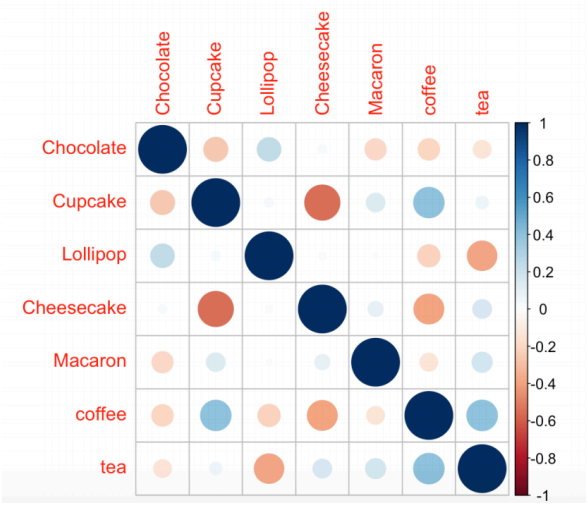
Also, I subset the ‘dessert’ related categories and compared between ‘cupcake’ and ‘tea’ topics. After applying PLSA to these two subset of reviews, I obtained the following visualization.



From the visualization we can see that the cupcake topics are more associated with the cupcakes' flavor: there are chocolate, red velvet, cinnamon, filling those key words that appear a lot; on the other hand, we can see that cupcake works both for 'family' and 'friends'. For the topics that related to 'tea', however, we can see that one of the top key words is 'bubble'/'boba' which is out of my expectation: I thought that tea may refer to the traditional tea house and didn't expect that bubble tea should be this popular! And also, we can see 'friend' appear a lot rather than family. So maybe eating cupcake is a better choice to spend time with family rather han having a bubble tea.

- **Task #2**

  Task two is about Cuisine Clustering and Map Construction. First I plotted the cuisine map (using similarity matrix of the cuisines). To improve it, I firstly added parameter True, and then I used the cosine similarity to compute the similarities between the two kinds of cuisines.
  In task two I used a new cuisine dataset and here is the improved result:



  For the last subtask, I used last week's cupcake case for clustering attempt. In this section, K-Means cluster analytics were applied and from the result we can observe that Chocolate, Cupcake and Cheesecake are considered to be in the same cluster whereas coffee and tea are considered to be in the same one – which makes sense.

- **Task #3**

  Task 3 is about Dish Recognition. I chose the Chinese cuisine category. During the manual tagging process, I deleted the phrases that are not related to Chinese cuisine (like 'the place in TX' etc.) also, I changed the 0s in the positive phrase

lines to 1s. To generate the reviews for Chinese cuisine, I used R to extract the reviews. For extending the names of Cuisines, I used TopMine tool from UIUC. The parameters set in the model were:

• Minimum support: 5 (minimum times a phrase candidate should appear in the corpus to be significant)

• Max pattern: 4 (max a phrase to be, because the dish name is seldom longer than 4 words)

• Number of topics: 5

• Gibbs Sampling iterations: 500

• Thresh (significance): 4 (the significance of a phrase. Equivalent to a z-score)

• Topic Model: 2

Chinese food	2063	
fried rice	1658	
Chinese restaurant		1452
egg rolls	1440	
orange chicken	1160	
lunch specials	966	
food is good	936	
Panda Express	900	
hot and sour soup		801
Mongolian beef	771	
pretty good	762	
dim sum 589		
crab puffs	540	
Chinese place	533	
love this place	526	
egg drop soup	516	
good food	514	
chow mein	459	
spring rolls	445	
Kung Pao Chicken		419
great food	409	
lo mein 401		
wonton soup	400	
soy sauce	374	
noodle soup	373	
highly recommend		372
Las Vegas	371	
food is great	371	
sesame chicken	364	
hot pot 354		

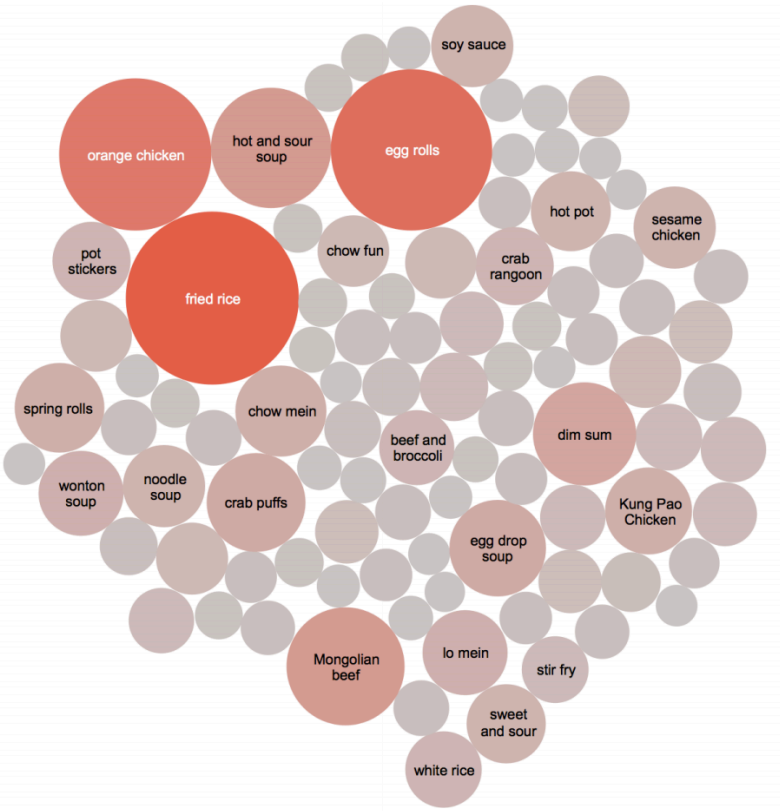
(Top frequent phrases)

Actually the result surprised me: I didn't expect that the stemming for phrases could be that accurate to some dishes. For example, because I am from Asia so I know that there are some dishes that are quite popular locally though I never knew about their English names; after getting the dictionary from the TopMine, I found that there are many dishes that I had thought could be very rare in the U.S that turned out to be quite popular (frequent). For example, the water chestnuts. Based on that, I think the result pretty makes sense.

- **Task #4&#5**

  These two topics are about popular dishes and restaurant recommendation. In Task 3, I analyzed the Chinese cuisine categories (a subset of 20,000 record) and

outputted a set of candidate dish names by TopMine. For task 4, first I completed the positive/negative analysis using sentiment analysis in R to get the positive reviewed dishes out of the candidate dish names; then I sorted the names by the frequency of occurrence and visualized them in Tableau. Here is the plot: larger size and brighter in red illustrates higher frequency of the positive reviews of a specific dish.



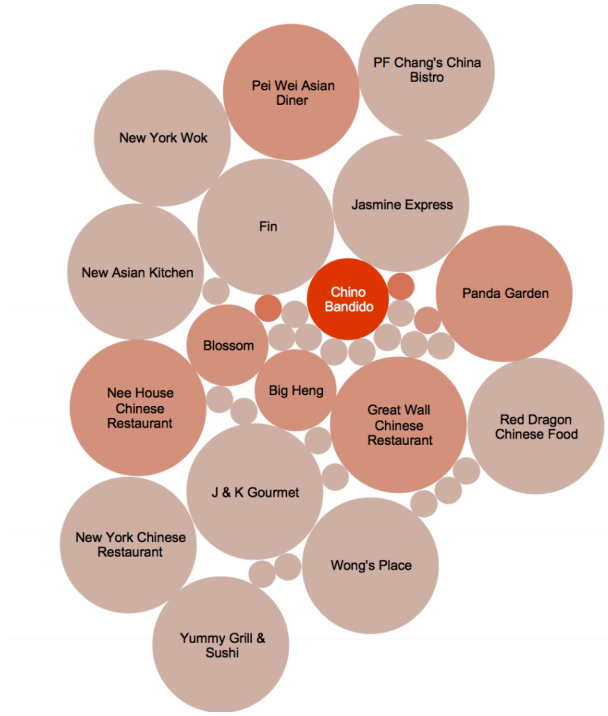
Top 3 dishes are fired rice, egg rolls and orange chicken

For task 5, I subset the Chinese cuisine in Las Vegas, NV and Phoenix, AZ, and merged with the positive review information from the whole review dataset. R was used for analysis. I got:

1) The count of positive reviews that are related to fried rice for each restaurant;
2) The address of each restaurant;
3) The average rating of each restaurant.

Also, I visualized the count of reviews and average rating for these Chinese restaurants. The size of circle illustrates the rating (stars): the larger the circle is, the higher rating the restaurant has: for instance, Fin, J&K Gourmet both have an average of 5 stars rating; and Big Heng has a 4.5 stars on average. And the brightness of color illustrates the count of reviews: for example, the brightest

one Chino Bandido has 6 positive reviews about fried rice and it has the brightest color among these restaurants.



One interesting result is that Chinese restaurants tend to close to each other. My guess is that they are very likely to be in Chinatown in those two cities together. I also go to google to check whether the rating matches: it did match. So maybe I would go for Panda Garden for fried rice if I am visiting Phoenix someday.

- **Task #6**

  In this task, several models using R were built to predict the hygiene conditions of the restaurants in the given dataset. Additionally, I merged external dataset to provide more potential predictors by the zip code information in the original dataset.

  For the text processing part, I used R to extract the key categories for each restaurant.

  For the modeling part, I used Random Forest and Gradient Boosted Tree model to build the classification. The predictors I used to build the models including bi-categorical variables for each class: whether the restaurant's categories contain American/Sandwiches/Chinese/Pizza /Traditional/Japanese/New /Mexican/Food/ Breakfast/Brunch/Italian/Vietnamese/Fast/Thai /Seafood/Bars/Sushi/Burgers/Delis/Cafes/Mediterranean/Barbeque/Asian/Fusion /Indian/Greek/Cantonese/Vegetarian; also, by the zip code I generated the

latitude and longitude of the location (and found that the dataset is subset from WA area) – I only included the categories that have more than 300 records in the dataset. Other numeric variables include number of reviews, ratings.
I used the dataset which has the result (0/1) as the training and testing set by a random 70%-30% split by the class. Based on the misclassification rate from the testing set, the gbm is 37.8% and the random forest one is 34.2%. So I used the Random Forest model and achieved approximately F1 score of 0.53.
Findings: Based on the importance of fit (appendix) we can see that the top important predictors include: number of reviews, ratings, longitude and latitude. For the specific category, Chinese restaurant and Thai restaurant are more likely to be categorized as 'not pass'.

- **Usefulness of Results:**

  There are several key points I found that my results are useful:

  - The correlation of desserts could provide the store owners how to pair the desserts together to increase revenue/profits. For example, people love to have cheesecake and tea together, so there could be a combo when buying them together to make it more popular; also, they could pair those desserts that are not sold that well and apply coupons.

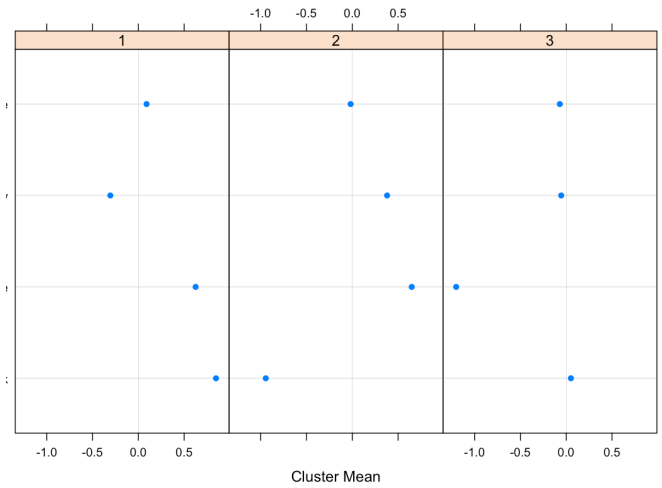
  - As a graduate student originally from China, I found myself a lot of time searching for good Asian food on yelp before. After completing task 4&5, I found some good restaurants for Chinese food with high ratings as well as star foods. Would definitely give it a try when I am visiting the area.

  - The classification prediction would give a potential result for the hygiene condition check and thus could give a warning to those restaurants with 'not pass' prediction.

- **Novelty of Exploration**

  - Clustering Result visualization

I used R (merged with cluster analytics code from NU analytics data mining class) for Clustering data visualization in the previous task report:



- Topic model cluster visualization
  I used d3.js for the first task and I personally like the way how the circles present relations between clusters: http://qinita.github.io/DataMiningMOOC-Capstone-Week1/

Thank you very much for reading this report! ☺