Yelp Dataset Challenge -- Infer Categories

Pinxia Ye, Si Liu
pinxia@hawk.iit.edu, sliu89@hawk.iit.edu

***Abstract*— The user generated reviews on Yelp contain abundant information which becomes significant reference in decision making, like dining, shopping and entertainment. In this project, we aim to explore restaurant categories on Yelp by analyzing its review data for the non-intuitive correlations, sub-categories. We leveraged several data preprocessing techniques, such as POS and Language Detection, to remove meaningless information from the original data and well prepared the data for analysis. We focused on mining the similarity among different restaurant categories. Several analysis techniques have been studied and adopted in our project to achieve the goal, such as LSA, tf-idf, K-means clustering, Hierarchical clustering, Multi-Dimensional Scaling. In addition, we also studied different similarity measurement, like Euclidean distance, cosine distance and Jaccard distance. Our project results show that these techniques are highly effective and well recommended.**

Keywords— Text Mining, LSA, tf-idf, POS, Language Detection, Cosine distance, Hierarchical clustering, Multi-Dimensional Scaling; Jaccard distance

I. INTRODUCTION

The user generated reviews on Yelp becomes important foundation in decision making, like dining, shopping and entertainment. In our project, we work on the "infer categories" problem in the Yelp Dataset Challenge [1], which is to explore the non-intuitive correlations between business categories or the reasons behind sub-categories, particularly we are more interested in exploring the correlations between restaurant categories: How well is the Yelp restaurant categorization? Does one category deserve sub-category? Can two different categories be combined into one category?

We studied the approaches that could be used for this problem and designed experiments to verify each approach with regard to data preprocessing and analysis respectively. For data preprocessing, we generally improved the processing with more experiments step by step: (1) extract review data from JSON; (2) use language detection tool to get data in English only; (3) run speech tagger with Stanford Log-linear Part-Of-Speech Tagger [3]; (4) remove small documents with data less than 2KB; (5) remove stop words.

In the first phase of the project, we collected the Yelp restaurant data and preprocessed them, from step 1 to step 3. We produced the 10 most representative words for the 7 restaurant categories we are interested and also visualized the word frequency with word clouds. In the second phase, we focused on the experiments design and implementation to explore the correlations between different categories. We designed two kinds of K-means clustering experiments on reviews: one is within one category; the other is within two categories. The within one category clustering experiment is to explore if there are some subset category could be separated from this one category. The inside two-category is to check whether two categories are well defined.

The results of the K-means clustering experiments are not as good as we expected. There is no clear and meaningful correlation between the categories which should have. We analyzed the reasons for those kinds of results, and summarized two major factors for them: the frequency of many common words, such as, "food", "restaurants", "drink", is large, which highly affect the clustering results; the clustering methods we adopted are not well suitable for this problem.

Therefore, in the third phase, we designed more experiments based on the previous analysis and the results are finally quite promising. The major enhancement includes deeper preprocessing, hierarchical clustering [4] and Multi-Dimensional Scaling (MDS) [7] on all restaurant categories reviews. We achieved desired results from these experiments.

The rest of this report is organized as follows. In Section II, the review data is introduced. Section III displays the experiments performed and summaries the experiment results. Section IV presents our analysis on the experiments. We conclude a summary in Section V. Section VI briefly describe our work distribution.

II. DATA

1. Raw Data

The dataset downloaded from Yelp website are in JSON format, of which the size is around 2.4GB in total, containing information for 77,079 businesses. There are two JSON files contain the information of categories and reviews:

- **yelp_academic_dataset_review.json**
- **yelp_academic_dataset_business.json**

We wrote Java code to extract categories and reviews from those two files. We built one directory for each category one by one and collect all reviews for that category with each document consisting one review.

The total number of categories is 890, with various of number of reviews ranging from 1 to 24974. The "Restaurants" category ranks the top with 24974 reviews. Figure 1 shows the distribution of number of reviews. nDoc is the number of document within each category. It's clear that the tall bar on the right means over 600 categories has less than 200 reviews. Also, on the other side, we have at least 1 category has more than 20 thousand reviews. It is because the categories aren't created equal. They have different level of granularity. For example, the one categories that has the most reviews (over 20k) is "Restaurant" under which there exists 136 sub-categories, like Mexican, American Traditional, Chinese etc. So looking at their category structure, it is a long tail with only a few high-level categories, like restaurant, shopping, health&beauty, and lots of sub-categories, such as sushi bar, fish and chips etc.

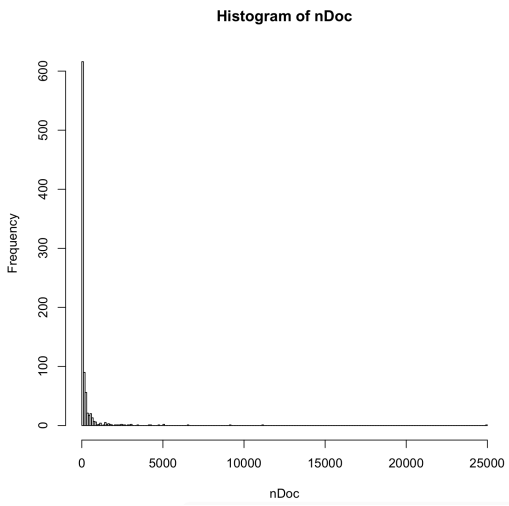

Figure 1 Histogram of the number of reviews of categories (before language filter)

2. *Data Pre-processing*

1) *Language Detection*

The raw review data consists of several languages. In our analysis, we only consider documents in English. Hence, we adopted a language detection tool [2], with 99% percent precision over 53 languages, to detect what kinds of languages in the original review data.

The pie chat in Figure 2 displays the percentages of detected languages percentages in the documents. Out of all restaurant reviews, about 95% are in English, which we assumed it was 100% for all of our previous experiments. The English reviews are about 40K. German, and French reviews take a few thousands each. Additionally, there are hundreds of reviews are written in other languages including European, Asian, and African. With the language detection tool, we filtered out all reviews that are non-English.

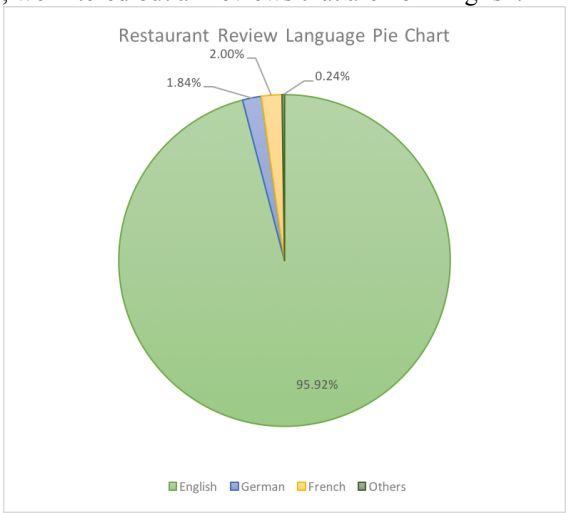

Figure 2 Pie chart of language distribution of restaurant reviews

2) *Data Tagging*

We adopt the Stanford Log-linear Part-Of-Speech Tagger (POS) [3] approach to tag the words in each review. The library is quite easy to use, and the manual explicitly states which tagger we should use, the left 3 words English tagger. That tagger was trained on WallStreet Journal articles with a test accuracy of over 97%, and a generalization accuracy about 90%.

The result of tagging is essentially the original text review plus the tagger append a "/AAAA" postfix to each word and punctuation. "AAAA" here means the type of word. It could be "NN" for nouns singular, "NNS" for nouns plural, "VB" for verb. Below is a sentence of tagged review from American Tradition category:

"Nice/JJ portions/NNS for/IN the/DT food/NN ,/, and/CC strong/JJ drinks/NNS !/. "

Notice that here "portions" is marked as plural noun by "/NNS", and "food" is singular noun as "/NN". The same kind of tag applied to each reviews.

3) *General pre-prpcessing*

We removed the un-important values from the text before processing, such punctuation, stop words, numbers and whitespaces. that can be problematic in running our experiments.

III. EXPERIMENTS

Our experiments focus on restaurants categories and reviews. First we experimented with K-means clustering within one category in order to explore the composition of categories, which will answer whether they could be split into multiple sub-categories. Also, we tested clustering for two categories, especially the ones that are intuitively connected. We found out from these experiments that the categories are tightly clustered with occasional outliers, which leads to our next experiment - hierarchical clustering. The reviews under same categories are concatenated into one document. We then discussed different approaches of similarity measurement which is later used to hierarchical clustering and multi-dimensional scaling.

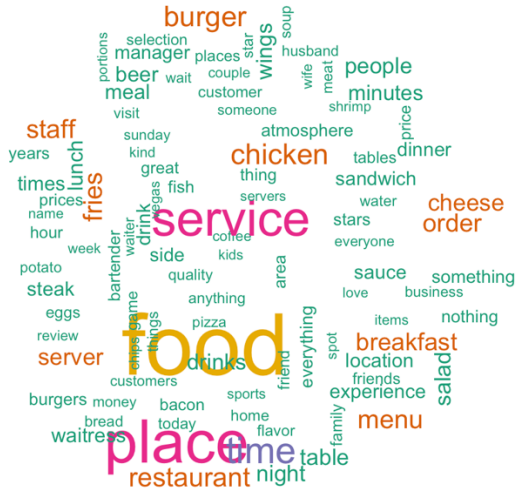

Figure 3 Word Cloud for catrgory "AmericanTraditional"

1. K-means Clustering

We reviewed the representative words from our Document-Term Matrix. Figure 3 shows the word cloud for one of them, while Table 1 presents number of documents and frequent words for 7 large restaurant categories. The part of speech tagger has done a very good job of extracting features - the nouns we extracted from those documents matches to typically what people talk about restaurants, such as location, service, food. However, the words that distinguishes each category are not necessarily obvious. So we think TF-IDF may be a good idea here to eliminate the "common factors".

1) One-category clustering

From Figure 2 and Table 1, we can observe that first some words like "place", "service", "food" exists in almost every category. We think that makes a lot of sense because those aspects are what people review of a restaurant. This result motivated us to deep preprocessing. Here we can see the POS tagging works. It reveals the topic of reviews very well.

Table 1 Most representative words for 7 categories

Categories	Most representative words
Bars	place food service time drinks night staff people beer drink
Mexican	food place service tacos taco time burrito salsa chicken restaurant
AmericanTraditional	food place service time chicken burger restaurant menu fries staff
CoffeeTea	coffee place starbucks time service location staff food people drink
Chinese	food place chicken service rice restaurant time order soup beef
SpecialtyFood	place store food chocolate service time staff selection meat prices
Desserts	place cake chocolate cream service food flavors time coffee staff

Now look at the words unique to each cloud, for example, burger and fries for American Traditional, burrito and tacos for Mexican, which actually talks about the uniqueness of each category, although we think the result should do better than that when we dig into the data deeper in the later experiments.

One of our goal is to help Yelp decide whether a category can be split into multiple sub-categories. To achieve that, we think Tf-idf on single categories and then clustering would work. We run tf-idf and cluster in R. We first reduce the tf-idf matrix dimension to 5 and then using K-means to cluster the reviews into 5 clusters.

Figure 3 presents a typical K-means clustering result (5 clusters). They aren't readable from the plot. The clusters aren't distinguishable, also the few outliers are caused by the number of nouns in those documents are too few to analyze. Almost all categories plot similarly, except for Mexican, from which we discovered the non-English reviews. After the non-English reviews are filtered out, the plot became similar to the figure above.

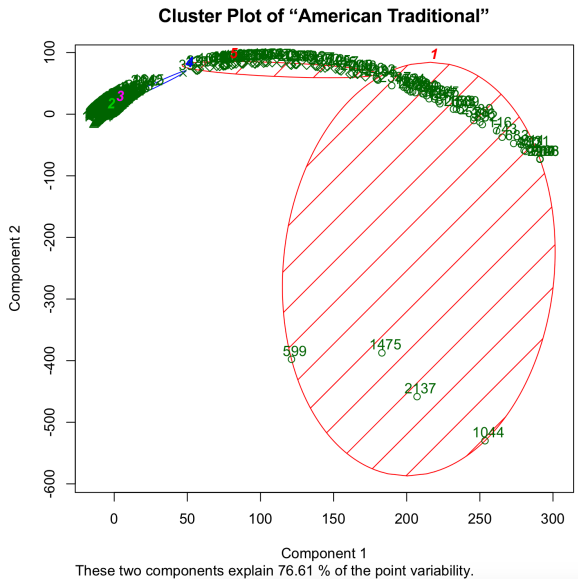

Figure 4 K-means clustering on TF-IDF inside "American Traditional" cagtegory

2) Two-category clustering

After we tried TF-IDF on single category, and the results aren't not ideal, we tried TF-IDF on two categories that we think is similar, but Yelp has decided to separate them, such as "American Traditional" and "American New", "Japanese" and "SushiBar", "Italian" and "Pizza". We want to find evidences by clustering the review that they do belong to separate categories. Figure 4 presents the clustering result on American Traditional and American New. This result is quite similar to the ones in previous experiments. The clusters aren't observable. These two categories' reviews are quite similar after all.

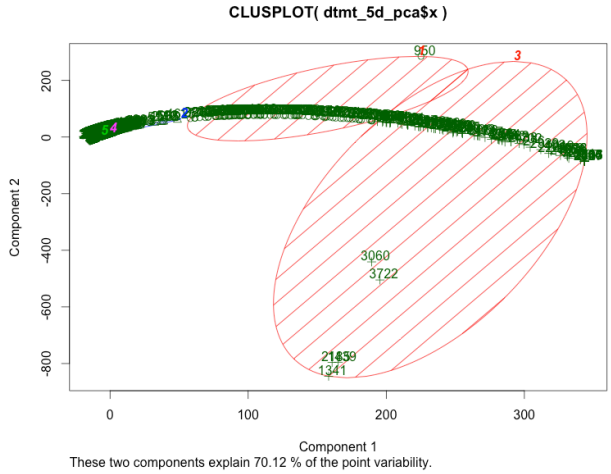

Figure 5 K-means clustering on TF-IDF of "American Traditional" and "American New"

2. Category Similarity

From the previous experiments, we gained the knowledge that each categories are tight clusters. From this point on, we concatenated on all the reviews in one category

into one file, so that each category reviews are included within one file. Then we name that file with its category's name, and put all restaurants category files into the same folder. This step is based on our finding from the precious clustering experiments that every category is tight clusters that could be abstracted as a dot in the concept space of all reviews.

1) Common Words Removal

In our single category exploration, we found out the top frequent words for almost all categories contains generic subjects like location, staff. We experiment with a manual list of stop words to substitute the tf-idf methods. The result actually makes more sense compared to the above. Our own list of stop words are selected from the single category top words list. We went back to our first experiments and take the words that doesn't help to identify categories, such as, "service", "location", "business", "food" and etc. The stop words list has 14 words. We removed those words on top of standard English stop words.

Below is the list of common words we removed in this part.

"food", "drink", "restaurant", "business" , "service" , "staff", "service" , "great" , "location", "place" , "time" , "atmosphere", "best" , "store" , "price"

2) Documents Check and Removal

Revisiting the previous experiments, we found out that some categories only contains less than 100 hundred words before stop words removal, which means it may only contain a few characteristic words. Also, some categories aren't necessarily restaurants, like "FruitsandVeggies" that is a grocery store, maybe offering buffet. After removing these categories, we have 98 categories left which has a document size more than 2KB, and a strong tie to restaurants.

3) Euclidean Distance, Cosine Similarity, Extended-Jaccard

The methods in the previous adopt Euclidean distance in default. Since there are still lots of unrelated categories clustered, we researched in other similarity measurements to explore the possibility of improving the results. In the later experiments, we tried cosine similarity matrix and Jaccard distance in the measurement methods and the results are really making sense.

We originally thought that all restaurant reviews should have some level of similarity, because they all belong to restaurant after all. However, see the measurement distribution, Figure 6 and Figure 7. We tried several parameters for dimension reduction, the magic number turned out to be 20. If the dimension is too few, all points are jammed together. When no dimensional reduction, almost all cosine values are close to 1, which means all categories are too far away from each other.

We can see that extended-Jaccard is stricter than cosine similarity, probably because it is literal term matching that ignores the synonyms. The clustering results, however, from both distances are similar to each other, which are both far better than Euclidean distance results. So the following experiments are based on cosine similarity.

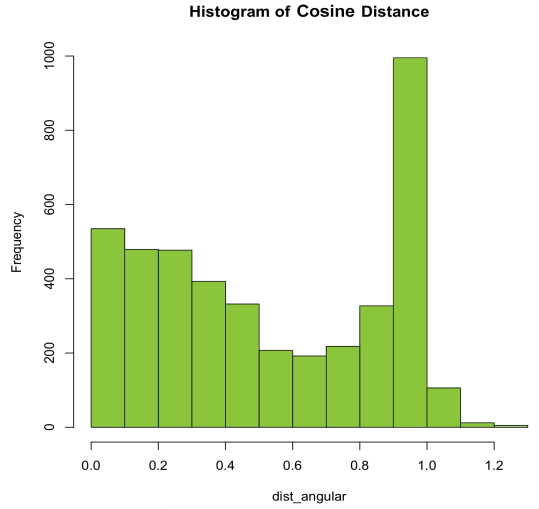

Figure 3 Histogram of Cosine similarity for all categories

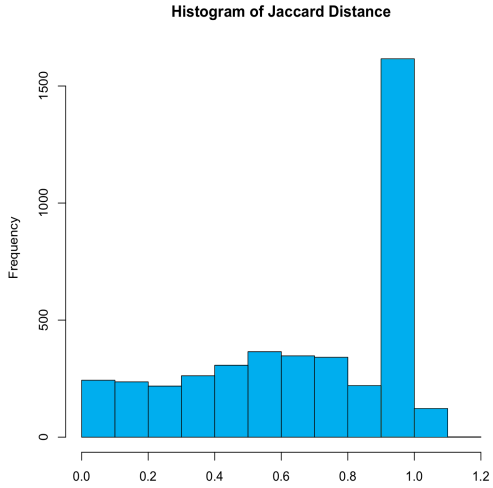

Figure 4 Histogram of Jaccard similarity for all categories

3. Hierarchical clustering

Hierarchical clustering [4] is another kind of clustering method different from K-means. In hierarchical clustering, the process requires a distance matrix, and the processes creates a cluster with the two closest points after evaluating all the points and re-evaluates the distance with the rest of the points and the new cluster. As we discussed before, we adopted cosine distance to measure the distance in this part.

We tried several ways to perform the hierarchical clustering to achieve more explainable results, including two different matrices, document term matrix and tf-idf matrix, and two similarity methods, Euclidean distance and cosine distance. We first applied hierarchical clustering on top of the common words removed document-term matrix. After the additional processing mentioned in previous section, this experiment included 95 categories, and achieved their intuitive correlations. Then, we tried tf-idf on top of our own stop words, and cosine similarity. The result is slightly

different from the one without tf-idf. See Figure 8 and Figure 9 for comparison.

Figure 8 shows a part of results of hierarchical cluster with dimensional reduced document term matrix, while Figure 9 shows the same data only with tf-idf matrix. We can observe the distances between categories are reduced. Both pairs are clustered in the experiments, just the tf-idf data shows they are almost identical.

4. *Multi-Dimensional Scaling*

Multi-Dimensional scaling (MDS) [7] is a means of visualizing the level of similarity of individual cases of a dataset. It refers to a set of related ordination techniques used in information visualization, in particular to display the information contained in a distance matrix. An MDS algorithm aims to place each object in N-dimensional space such that the between-object distances are preserved as well as possible. In this part of the experiment, we performed MDS on all the restaurant data with respect to cosine distance and extended-Jaccard distance respectively.

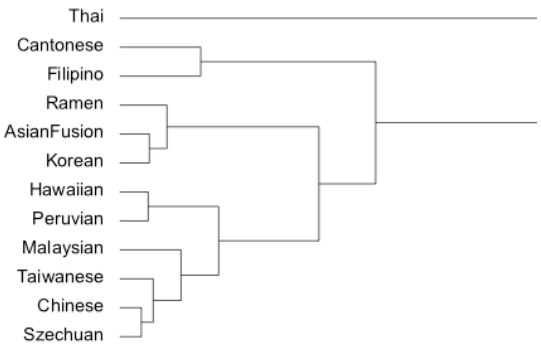

Figure 5 Heriarchical clustering with cosine distance using DTM

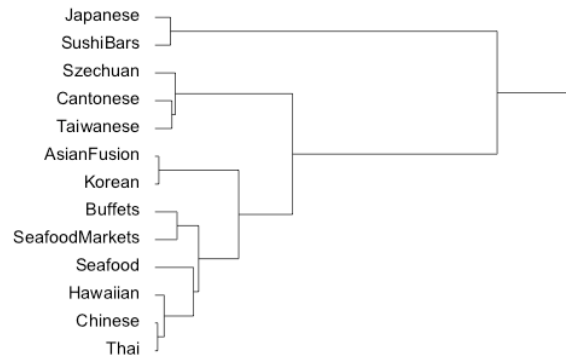

Figure 6 Heriarchical clustering with cosine distance using tf-idf

Figure 10 is part of the result of MDS using cosine distance, which visually display the distance of categories. Categories stay quite close together are with high correlation to each other. For instance, "Szechuan" and "Chinese", "Taiwanese" and "Cantonese", "AmericanTraditional" and "AmericanNew". On the other hand, if two categories stay quite far from each other, their similarity is low. For example, "Russian" and "African" in the figure.

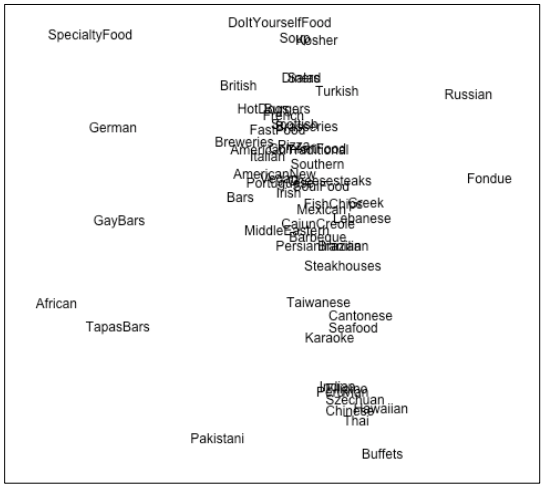

Figure 7 MDS with cosine distance on all categories

Figure 11 is part of the result of MDS using extended-Jaccard distance. The results is similar to the cosine distance, but difference exists. For example, "Taiwanese" and "Cantonese" don't stay as close as cosine distance. Different measurement should have different result, but in general, their results are reasonable and acceptable.

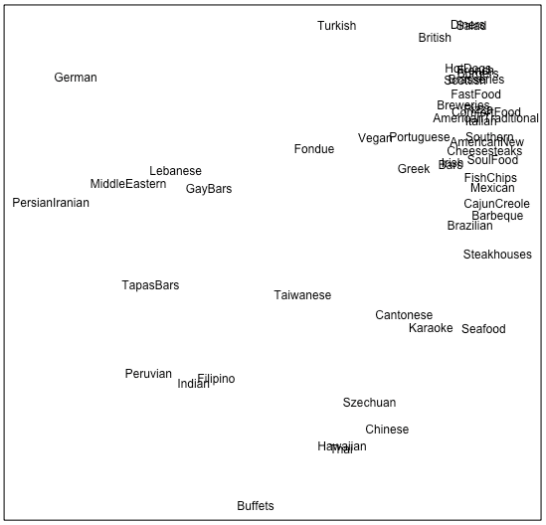

Figure 8 MDS with extended-Jaccard distance on all categories

IV. Analysis

Most of the analysis has been done along with the experiment results. In this part, we would like to further analyze the factors of our experiment and results.

1. *Imbalanced number of category reviews*

For the review data used in our projects, the number of reviews of each restaurant varies from1 to 24974, which is a big range. It indicates that categories with less review data would not be analyzed precisely, which could result in mistakes in the analysis of correlation. Thus, we can only predicate the correlation among different categories based on our experiment results but make precise conclusion.

2. Achieved intuitive correlations

The biggest achievement of our experiments is that the intuitive correlations between different categories are identified. From both hierarchical clustering and MDS, many intuitive correlations are clearly displayed. For example, "Szechuan" and "Chinese", "Taiwanese" and "Cantonese", "AmericanTraditional" and "AmericanNew, "Japanese" and "Sushibars, "" and "Italian" and "Pizza".

3. Small Categories Evaluation

In the challenge description, one of the questions from Yelp is whether Szechuan deserve to be a separate category rather than Chinese. First of all, it is a much smaller category than Chinese. Their document term matrix cosine similarity is 0.02, but the tf-idf cosine similarity is 0.04. Figure 8 and figure 9 compares the different positions in the document term matrix and tf-idf matrix hierarchical cluster.

Although Chinese and Szechuan look alike in the document-term matrix, they show different characters in tf-idf, which quite matches to the fact in reality. Szechuan is a branch of Chinese food, specialized in hot and spicy dishes. It has quite a few famous dishes you may find in most of Chinese restaurant like Ma Po Tofu, but in general, Szechuan dishes are almost always hot and spicy. Due to its characters are shown in the reviews, Szechuan deserves to become a separate category.

4. Non-intuitive correlations

Another question we want to answer in the challenge is the non-intuitive correlations like "Karaoke" and "Korean". We think Karaoke is an interesting category that worths a closer look.

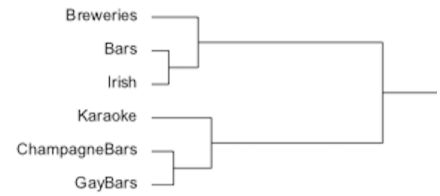

Figure 12 "Karaoke" in document term matrix hierarchical cluster.

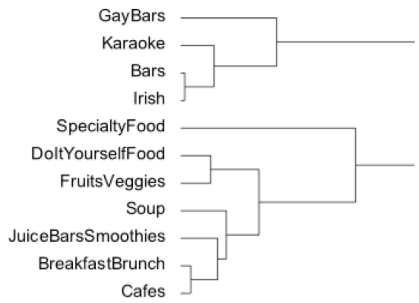

Figure 13 "Karaoke" in tf-idf hierarchical cluster.

First, Figure 12 and 13 shows Karaoke's positions in hierarchical clustering. Karaoke is clustered with bars, which is obvious and intuitive, but karaoke is also clustered with gay bars. At first, we thought this was caused by the drink part, then we went back to the reviews and discovered gay bars reviews did mention more "karaoke" than the generic bar, and searching online we found out karaoke is a big topic in LGBT community. Maybe this correlation is intuitive for some people, but it turned out to be a surprise for us.

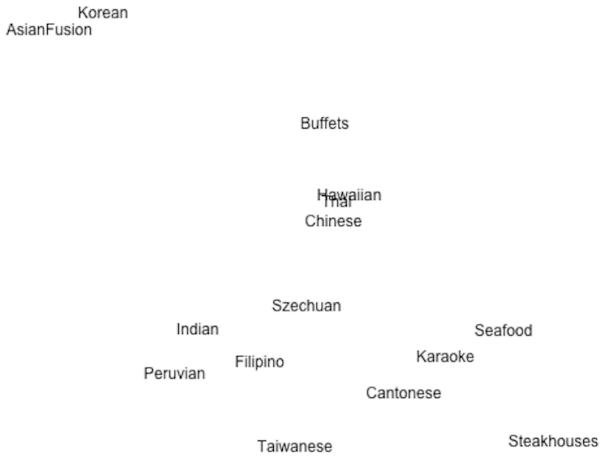

Figure 14 "Karaoke" in Multi-Dimensional Scaling (partial).

Additionally, Figure 14 presents Karaoke's position in Multi-Dimensional Scaling (enlarged). It is within the Asian quarter, surrounded by Cantonese, Seafood, and Szechuan. Korean is not in the cropped map. This actually shows some level of relationship between karaoke and asian cuisines. Given the origin of the word "karaoke", it is quite expected that it can be found in asian restaurants.

We first looked into the cosine similarity matrix, and reviewed the similarities of Karaoke vs other categories, then went back to the original Yelp business data, where we can see how the business are categorized. We counted the business with Karaoke category tag and also the number of occurrence for other types. Chart 15 shows the top occurrence categories plus some similar categories with Karaoke.

Category	Occurrence Count	Cosine Distance	Tf-idf Cosine Distance
Karaoke	122	-	-
Restaurants	55	-	-
Bars	82	0.03	0.04
American Traditional	13	0.26	0.07
Gay Bars	8	0.04	0.12
Asian Fusion	7	0.41	0.27
Korean	5	0.42	0.25
Filipino	5	0.36	0.21
Chinese	2	0.50	0.15

Chart 15 Occurrence of other tagged categories of Karaoke businesses, and those categories' similarities with Karaoke.

The first column are category names. The second column shows the occurrence count of each category. “Karaoke” business was tagged 122 in the dataset. Out of those 122 businesses, 55 of them are restaurants, 82 of them are bars. For bars, and the restaurants categories below, we also list the cosine distance of document term matrix, and the tf-df cosine distance.

We can see that bars has the closest distance among all. Surprisingly, American traditional has higher number of occurrence, compared to the Asian cuisines, probably due to the food at bars. This insight we got from hierarchical cluster. Also, we can see that the Asian cuisines despite of a far distance in the third column, they actually shows a near distance in tf-idf, which was pointed out in MDS. In Figure 14, “Filipino” is on the right of “Karaoke”.

We should also consider the occurrence of category types is not necessarily an accurate measurement. The reason is businesses has different number of reviews. Some business may be strong in number, but the reviews doesn’t always reflect that. Also, Yelp’s category tagging is not 100% accurate. For example, when we searched gay karaoke in yelp.com, a fair amount of the returned results are only marked as gay bar, but with key word “karaoke” in the review text.

So far, we have achieved exploring correlations of Karaoke. If we apply the same method, we could analyze any categories. But also note this method only works when users are sufficiently enthusiastic about the service such that they leave reviews, which is the source of our data.

V. Conclusion

In this project, we reviewed and reused the data analysis methods studied in class, such as latent semantic analysis ,SVD, tf-idf and K-means clustering. Thanks to this project training, all those techniques are further understood and enhanced in our mind. In addition, we utilized powerful data preprocessing tools, Part-Of-Speech tagging and language detection, which must benefit us in the future study and work.

We also experimented that two other data mining techniques: hierarchical clustering and Multi-Dimensional Scaling. Hierarchical clustering is an effective technique to explore the correlation between different restaurant categories. It allows us to see the groupings of similar restaurant categories in the plot visually and finally produced well grouped results for our problem. Multi-Dimensional Scaling is also a good method to visualize the level of similarity of individual cases of a dataset based on different similarity measurement. We are sure that both the hierarchical clustering technique and MDS will also be helpful to us in the future study and work.

References

[1] https://www.yelp.com/dataset_challenge
[2] “Mining Opinion Features in Customer Reviews”, Hu and Liu, 2004
[3] http://nlp.stanford.edu/software/tagger.shtml
[4] http://beyondvalence.blogspot.com/2013/12/cluster-analysis-hierarchical-modeling.html
[5] https://github.com/shuyo/language-detection
[6] https://en.wikipedia.org/wiki/Cosine_similarity
[7] https://en.wikipedia.org/wiki/Multidimensional_scaling
[8] https://en.wikipedia.org/wiki/Jaccard_index