CS522 Final Project Report

Yelp Dataset Challlenge -- Infer Categories

Name: Si Liu
CWID: A20334820

Abstract— **The user generated reviews on Yelp contain abundant information, which becomes important reference material in decision making, like dining, shopping and entertainment. In this project, we aim to explore the non-intuitive correlations between restaurant categories on Yelp by analyzing its review data. We learned 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 recommenedd.**

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

I. INTRODUCTION

The user generated reviews on Yelp contain abundant information, which becomes important reference material 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, 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 bunch of 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) run speech tagger with Stanford Log-linear Part-Of-Speech Tagger [3]; (3) use language detection tool to get data in English only; (4) remove small documents with data less than 2KB; (5) remove common words with high frequency.

In the first phase of the project, we collected the Yelp restaurant data to analyze and preprocessed them with POS. 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 inside one- category; the other is inside two-category. The inside 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 experiments are deeper preprocessing, hierarchical clustering [4] and Multi-Dimensional Scaling (MDS) [7] on all restaurant categories reviews. The results of this part are quite attractive and we did make achievement in this part.

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 write 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. Table 1 summarizes the number of reviews in each category. Considering the long list of categories, we only list the categories with the number of reviews more than 1000, which are more worth for mining and analysis.

Table 1 Number of documents for each restaurant category

Category	Number of Documents (Reviews)
Restaurants	24974
Shopping	11162
Food	9194
BeautySpas	6556
HealthMedical	5095
Nightlife	5046
HomeServices	4769
Bars	4298
Automotive	4197
LocalServices	3459
ActiveLife	3090
Fashion	3051
EventPlanningServices	2963
FastFood	2839
Pizza	2649
Mexican	2515
HotelsTravel	2481
AmericanTraditional	2414
Sandwiches	2363
ArtsEntertainment	2261
CoffeeTea	2185
HairSalons	2072
Italian	1836
Burgers	1772
AutoRepair	1710
Doctors	1681
NailSalons	1667
Chinese	1623
AmericanNew	1593
HomeGarden	1578
Pets	1483
FitnessInstruction	1439
Hotels	1426
RealEstate	1421
Grocery	1414
BreakfastBrunch	1369
Dentists	1191
SpecialtyFood	1145
WomensClothing	1132
Bakeries	1108
ProfessionalServices	1018
IceCreamFrozenYogurt	1015

In our project, we are only interested in finding the non-intuitive correlation between different restaurant categories. There are 136 restaurant categories in total, and most of them are for different countries, which are named the same as country names. In addition, the number of categories with reviews less than 100 are more than a half, which are not worth for analysis. Therefore, we firstly considered out 7 famous restaurants categories, 6 of them with the number of reviews larger than 1000: Bars, AmericanTraditional, CoffeeTea, Chinese, SpecialtyFood, and Deserts with 700 reviews. Table 2 lists the 7 categories we study. It includes the number of documents and unique words belongs to each category.

Table 2 Number of documents and unique words for 7 categories

Categories	Number of Documents	Number of Unique Words
Bars	4298	12483
Mexican	2515	5515
AmericanTraditional	2414	6511
CoffeeTea	2185	7146
Chinese	1623	4611
SpecialtyFood	1145	5730
Desserts	700	3128

We designed two kinds of clustering experiments on reviews: K-means clustering and Hierarchical clustering. K-means clustering is performed in two ways: one is inside one- category; the other is inside two-category. The inside one-category clustering experiment is performed on the seven categories to explore if there are some subset category could be separated from this one category. The inside two-category is performed on three pairs of six categories to check whether two categories are well defined. The top 10 words of 5 concepts in each category are also produced to help better understand the classification. The hierarchical clustering experiments are on reviews of all restaurant categories, which should be reconstructed for clustering. All reviews of one category are aggregated into one document with the same name as the category name. In addition, we also performed Multi-Dimensional Scaling (MDS) [7] experiments on all categories as the last phase of the experiment.

III. Experiments

1. *Data Preprocessing*

1.1 Data Tagging

We adopt the Stanford Log-linear Part-Of-Speech Tagger [3] approach to all review documents of 7 categories to filter out all noun words for analysis, since noun words are with more concrete meaning and features.

1.2 Remove non-important words

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.

1.3 Create document-term matrix

We created two different the document-term matrix using word frequency as matrix entry and using tf-idf

(term frequency – inverse document frequency as entry) respectively.

1.4 Most representative words

We also produced the 10 most representative words for the 7 categories in the order of their importance as shown in the table below. These words have high relevance to that category.

Table 3 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

1.5 Word clouds

In order to be more visually distinguish which terms appear more frequently in a category. We also visualized the word frequency with word clouds, which is a popular text visualization presenting word frequency, popularity or importance through font size. Majority of word clouds lay out the words in a random, though aesthetically appealing way. They can provide a summarization of large amounts of text in limited space and guide users' attention to discover more related information. Word clouds have been widely used in both business and research. We produced word cloud for each category. The bigger the size of the word in a word cloud, the higher the frequency of that word in the category.

Figure 1 is an example of the word cloud for category "AmericanTraditional". From it we can see that the words "food", "place", "service", "time", "restaurant", are quite striking among all words. It indicates that there are many such kind of words in the review of "AmericanTraditional" category.

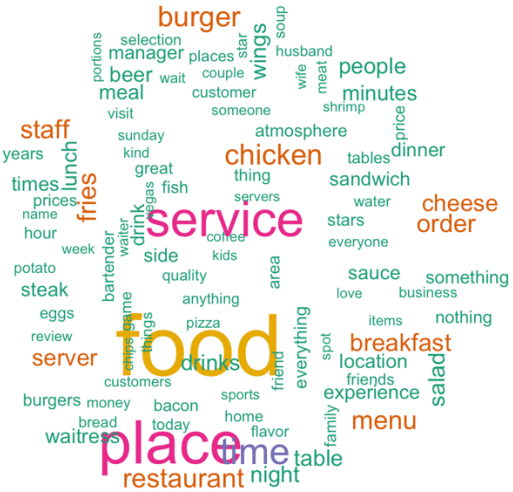

Figure 1 Word Cloud for catrgory "AmericanTraditional"

2. K-means Clustering

K-means clustering is performed in two ways: one is inside one- category; the other is inside two-category. The inside one-category clustering experiment is performed on the seven categories to explore if there are some subset category could be separated from this one category. The inside two-category is performed on three pairs of six categories to check whether two categories are well defined. The top 10 words of 5 concepts in each category are also produced to help better understand the classification.

1) One-category clustering

We conducted the one-category clustering on the 7 famous restaurants categories respectively, 6 of them with the number of reviews larger than 1000: Bars, AmericanTraditional, CoffeeTea, Chinese, SpecialtyFood, and Deserts with 700 reviews.

In general, the results of this one-category are not as good as we expected, because it doesn't show any useful information to tell if one category can be subset. This kind of results is explainable, we can draw the conclusion that each category is well defined and there doesn't exist any good subset option for the category. Figure 2 is an example of the clustering result of "Mexican" category.

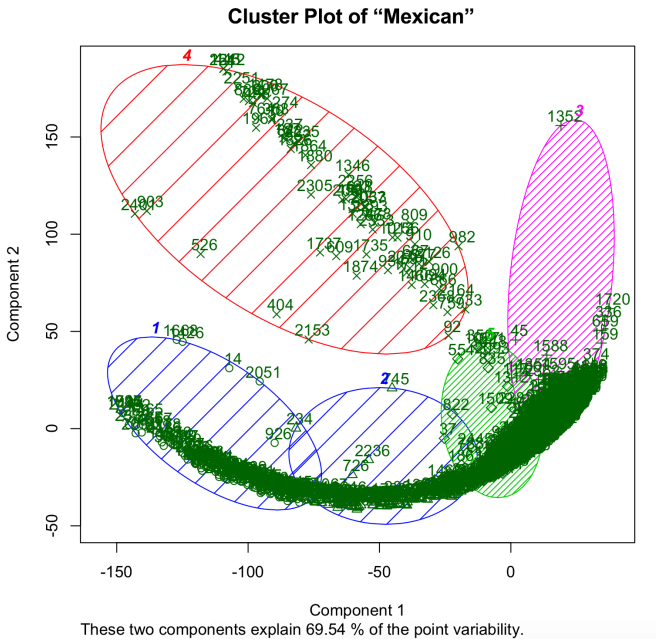

Figure 2 K-means clustering on "Mexican" cagtegory

We also produced top 10 words of 5 concepts in each category. This result can show what kinds of words appear in the category frequently. Below is the list of top 10 words of “Mexican” category:

[1] "business" "location" "place" "food" "service" "tacos" "staff" "taco" "time" "burrito"

[2] "location" "business" "place" "tacos" "food" "service" "burrito" "taco" "staff" "time"

[3] "hanshiktaco" "escabeche" "pozole" "brian" "rude" "gordo" "raspados" "cocktails" "tomatoes" "flan"

[4] "tolteca" "burrito" "staff" "breakfast" "customers" "waitress" "rafa" "realz" "tacos" "staple"

[5] "location" "tacos" "place" "burrito" "asada" "carne" "food" "staff" "salsa" "great"

We can see from the words list above that there are many common words, such as “business”, “location”, “place”, “food”, “time”, appear as the top words in the review of “Mexican” category. In fact, all the other categories have the similar results, which contain a large amount of common words.

2) Two-category clustering

We conducted the two-category clustering on the three restaurants category pairs respectively: AmericanTraditional vs. AmericanNew, Italian vs. Pizza, Javanese vs. SushiBars. Figure 3 is the clustering results of category pair Japanese and SushiBars. We still can not clearly explain the information in the figure.

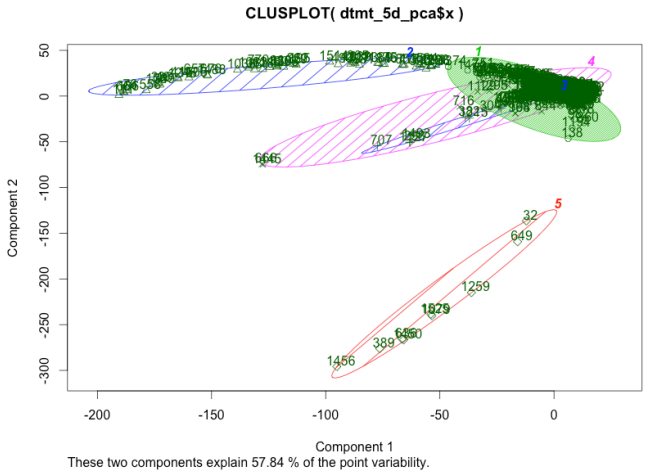

Figure 3 K-means clustering on "Japanese-SushiBars" categories

We also produced top 10 words of 5 concepts in each category pair. Below is the list of top 10 words of “Japanese-SushiBars” category:

[1] "business" "sushi" "place" "food" "service" "time" "restaurant" "roll" "staff" "chicken"

[2] "business" "mark" "sushi" "food" "place" "service" "china" "roll" "time" "staff"

[3] "salty" "entrée" "york" "sauces" "vegetables" "meat" “steak" "clay" "plant" "snow"

[4] "complaints" "average" "upside" "shooters" "oyster" "pineapple" "convenience" "cake" "panda" "express"

[5] "aubrey" "service" "place" "chicken" "rice" "roll" "teriyaki" "sushi" "lunch" "love"

We can see from the words list above that there are still many common words, such as “meat”, “rice”, “place”, “service”, “food”, appear as the top words in the review of “Japanese-SushiBars” category. The same phenomena for the others.

3. Hierarchical clustering

The results of the previous 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 this part, we designed more experiments based on the previous analysis and the results are finally quite promising. The major experiments are deeper preprocessing and hierarchical clustering [4] on all restaurant categories reviews.

In the deeper preprocessing part, three kinds of deep preprocessing are performed: language detection, small documents removal and common words removal. For language detection, we adopted a language detection tool [5] to detect what kinds of different languages in the documents and removed those documents which are not in English. For

small documents removal, we erased review documents with size less than 2KB, which doesn't contain convincing information. For common words removal, we created a list of common words which don't distinguish one category from another and removed them.

In the hierarchical clustering part, we conducted several experiments with different parameters before clustering, such as use document term matrix and use tf-idf matrix, Euclidean distance and Cosine similarity [6]. The results indicate a big difference between Euclidean distance and Cosine similarity.

The hierarchical clustering experiments are on reviews of all restaurant categories, which should be reconstructed for clustering. All reviews of one category are aggregated into one document with the same name as the category name. Thus, the data is composed of documents of categories, of which the total number is 133.

In addition, we also performed Multi-Dimensional Scaling (MDS) [7] experiments on all categories as the last phase of the experiment. It is to visualize the level of similarity of individual cases of a dataset.

3.1 Deeper preprocessing

1) Language Detection

In this first part, we adopted a language detection tool [2] to detect what kinds of different languages in the documents and removed those documents which are not in English. The pie chat below displays the percentages of detected languages percentage in the documents.

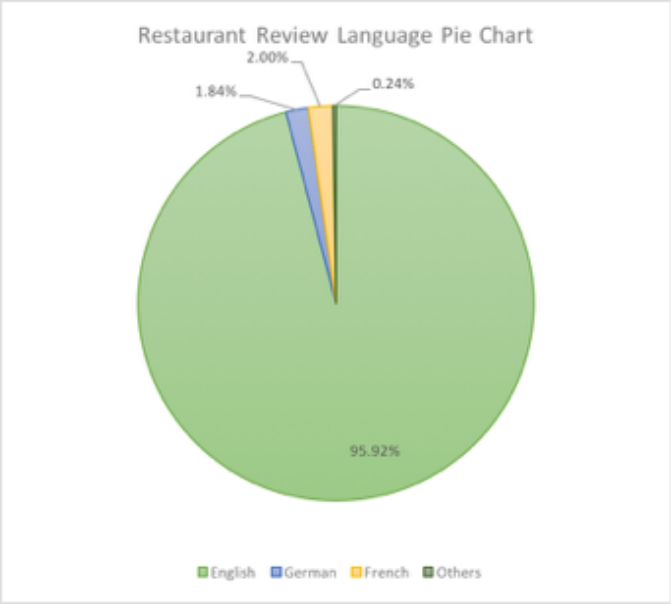

Figure 4 Pie char of language distribution of restaurant reviews

As we can see from the chat, there are mainly three kinds of languages: English, German and French. The other languages only counts 0.24%. Since English is the only one language interested, we removed the other language text in documents.

2) Remove small documents

Small documents with fewer text can not tell any meaningful features, so we deleted documents with size less than 2KB. After this step, there are totally 98 documents (categories) left for further processing.

3) Common words removal

We created a list of common words which don't distinguish one category from another and removed them. 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"

3.2 Hierarchical Clustering

Hierarchical clustering 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. There are multiple distances we can use, with different results, such as Euclidean distance and cosine distance.

The hierarchical clustering experiments are conducted on the data preprocessed above. We tried mainly four 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. The results indicate that the cosine distance similarity method is more effective for this problem.

Note: Since the results figures are too large to display clearly in the report, we only cut sub-figures with representative information from the original figures.

1) Euclidean distance

In this part, we adopt Euclidean distance to measure the similarity of different categories. We performed the hierarchical clustering using Document Term Matrix and tf-idf matrix representatively. We use the log value of the height parameter to plot the results of hierarchical clustering, because many of the original values are too small to be seen, which cannot be analyzed.

i. Use Document Term Matrix

Figure 5 is part of the results of hierarchical clustering using Document Term Matrix. We can see that several categories we interested are together, such as "Chinese" and "Szechuan", "AsianFusion" and "Korean". The closer between two categories, the more similar they are. They are more correlated to each other. This step has made much better progress comparing to the K-means clustering experiments. However their distance are still far. In order to explore whether the tf-idf would improve the results of Euclidean distance, we further conducted experiments using tf-idf.

Figure 5 Heriarchical clustering with Eulidean distance use DTM

ii. Use tf-idf

Figure 6 represents the results of hierarchical clustering using tf-idf matrix. Comparing to the results of using document term matrix, we couldn't see any big improvement from the above figures. The distance between two similar categories are still large, even though they are together, such as "Japanese" and "SushiBars". In addition, some similar categories are separated further from each other than the previous experiment, "AmericanNew" and "AmericanTraditional", for instance.

Figure 6 Heriarchical clustering with Eulidean distance using tf-idf

2) Cosine distance

Cosine similarity gives a useful measure of how similar two documents are likely to be in terms of their subject matter. This technique is widely used to measure cohesion within clusters in the field of data mining.

i. Use DocumentTermMatrix

Figure 7 represents the results of hierarchical clustering using Document Term Matrix. From it, we can see that most categories with high similarities are connected together with short distances. The results finally become much more reasonable comparing to the results of using Euclidean distance.

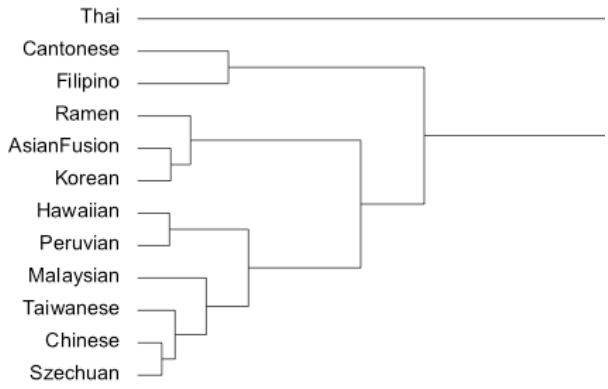

Figure 7 Heriarchical clustering with cosine distance using DTM

In order to explore whether the tf-idf would improve the results of cosine distance, we further conducted experiments using tf-idf.

ii. Use tf-idf

The figures below represent the results of hierarchical clustering using tf-idf. From it, we can see that most categories with high similarities are connected together with much shorter distances than using document term matrix. For example, "Japanese" and "SushiBars", "Chinese" and "Thai". This is, again, a big gain comparing to the results of Euclidean distance.

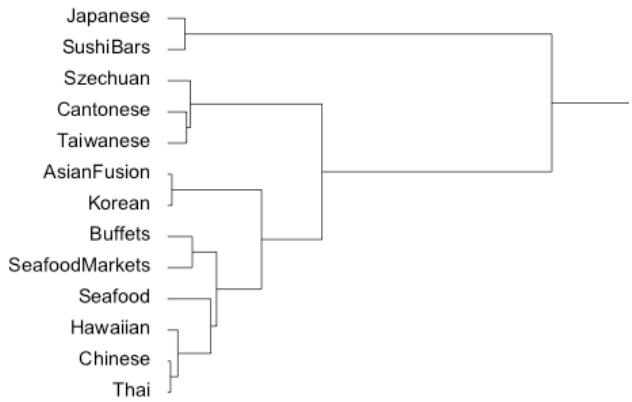

Figure 8 Heriarchical clustering with cosine distance using tf-idf

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 9 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.

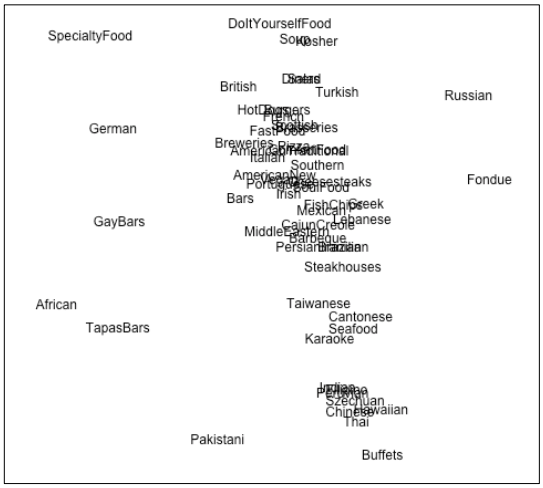

Figure 9 MDS with cosine distance on all categories

Figure 10 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.

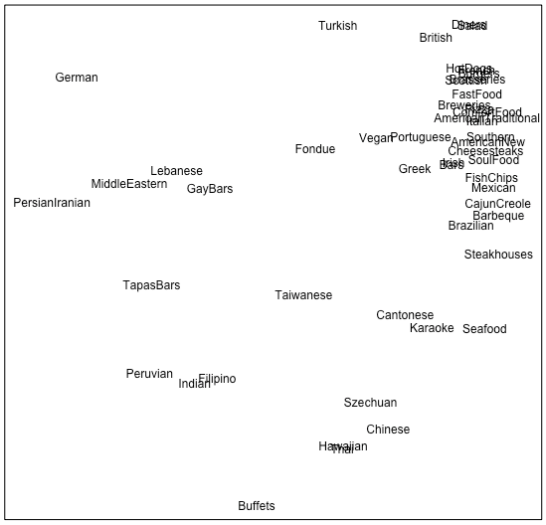

Figure 10 MDS with extended-Jaccard distance on all categories

5. Result Summary

We summarized the results of our experiments using histograms based on the two major measure distances throughout our project, cosine distance and Jaccard distance.

Figure 11 is the result of cosine distance. Figure 12 is the result of the Jaccard distance. Although two histograms are different, both of them reveal that categories between 0.9 and 1.0 occupy the most, and very few categories fall into the range larger than 1.2, which means the current restaurant categorization of the Yelp is quite reasonable and there is less chance to re-categorize them.

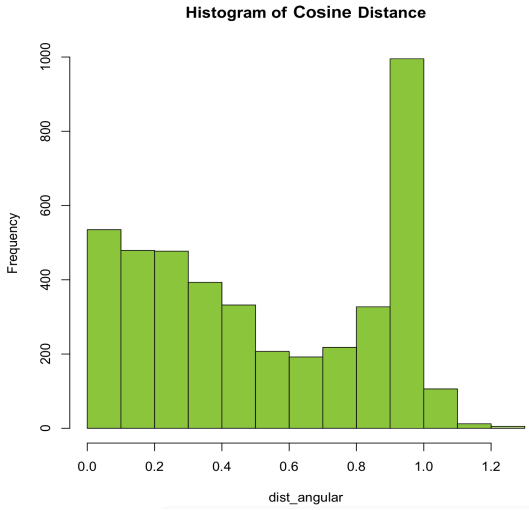

Figure 11 Histogram of Cosine similarity for all categories

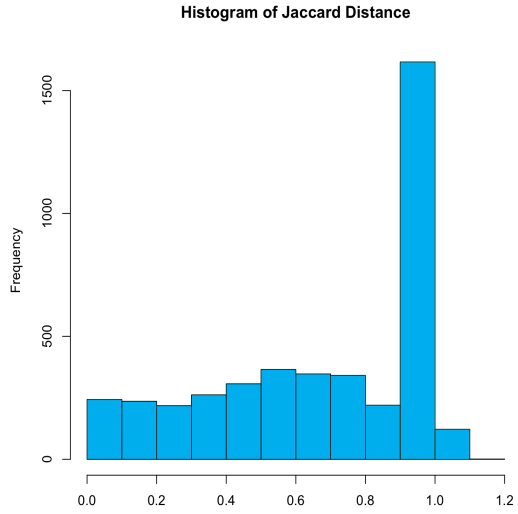

Figure 12 Histogram of Jaccard similarity for all categories

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. Explored non-intuitive correlations

The results of our experiments also reveal some non-intuitive correlations between categories. Some of the results are beyond our expectations. For example, “GarBars” and “Karaoke” almost stay together with each other in most cases. This drove our interests in exploring their relations further. Then we searched online and did find that many Karaoke offer Gay Bars in Chicago.

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 learnt useful data preprocessing tools, Part-Of-Speech tagging and language detection, which must benefit us in the future study and work.

From this project , we also leant 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.

VI. Wrok Distribution

We have two team members working together on this project. One is Pinxia Ye, the other is me (Si Liu). During the project, we worked closely together most time and discussed every step to proceed together. For different similarity measurement, we studied them together and chose the ones used in the experiment. The general work distribution of us is as following:

Pinxia Ye:

1) Extract none words using POS
2) Identify different languages in the review using language detector
3) Perform K-means clustering in side one-category
4) Perform MDS on all restaurant categories

Si Liu:

1) Extract review data from the raw JSON file
2) Preprocess the review data for clustering experiments
3) Perform K-means clustering inside two-category
4) Perform Hierarchical clustering on all restaurant categories

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