Data Mining Project - Week 1 - Exploration of Data Set

Data Mining Specialization - Coursera / University of Illinois at Urbana-Champaign

- Author: Michael Onishi
- Date: November, 2019

Introduction

The goal of this task is to explore the Yelp data set to get a sense about what the data look like and their characteristics.

For this project, I used the python programming language along with some libraries such as pandas, sklearn, wordcloud, pyLDAvis.

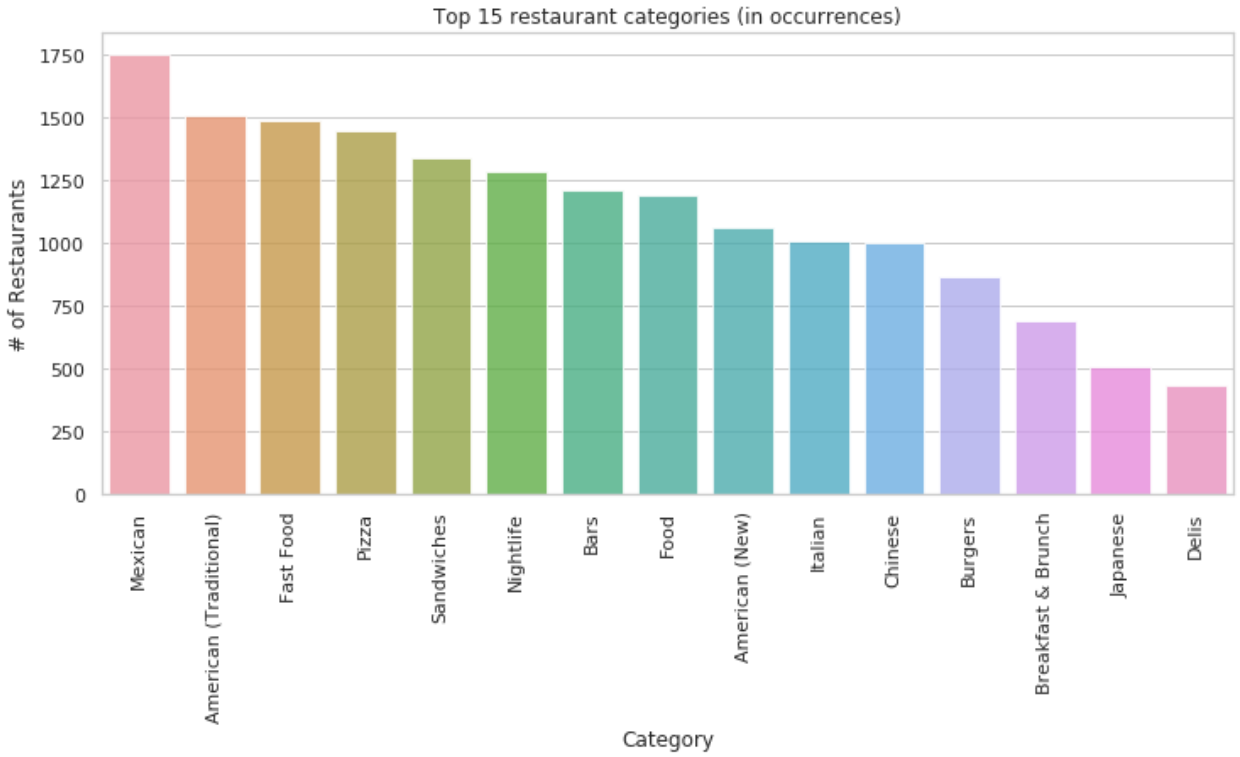
The full notebook with all the steps and details taken here is available at: https://github.com/michaelonishi/coursera-data-mining-specialization/blob/master/c6-data-mining-project/task1/Data_Mining_Project_Week_1_Exploration_of_Data_Set.ipynb

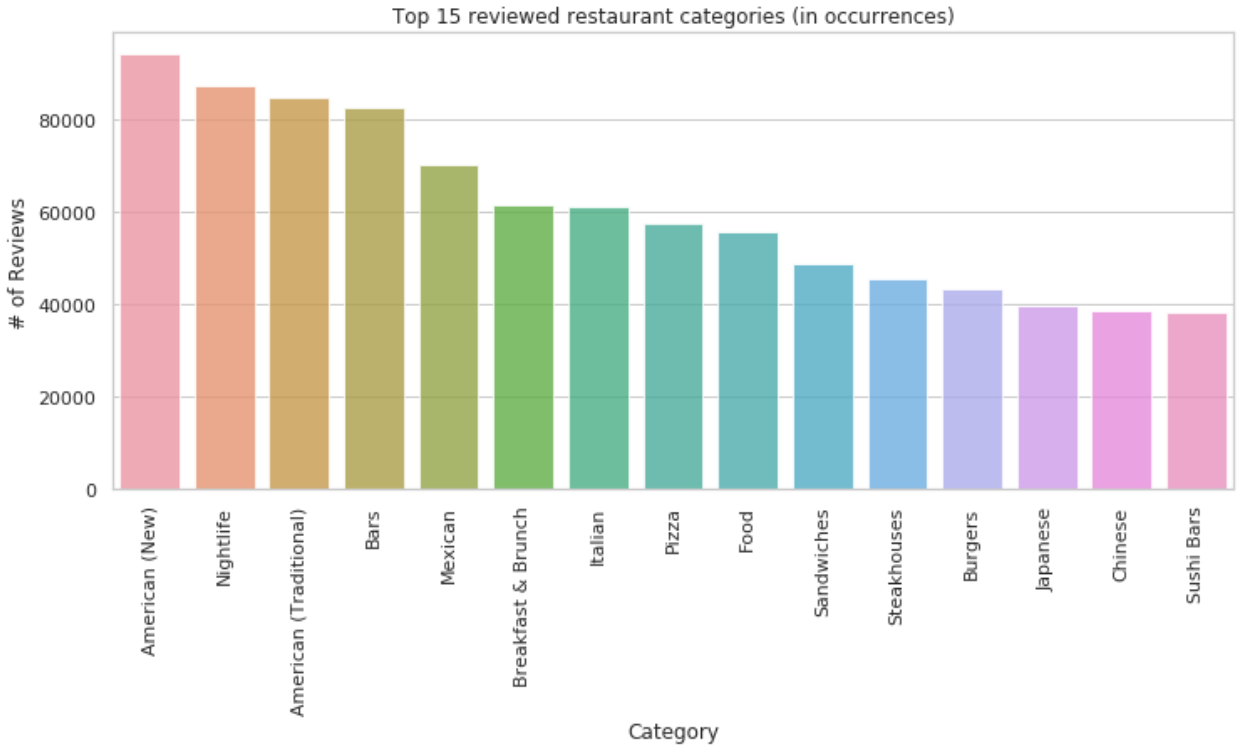
Data Exploration

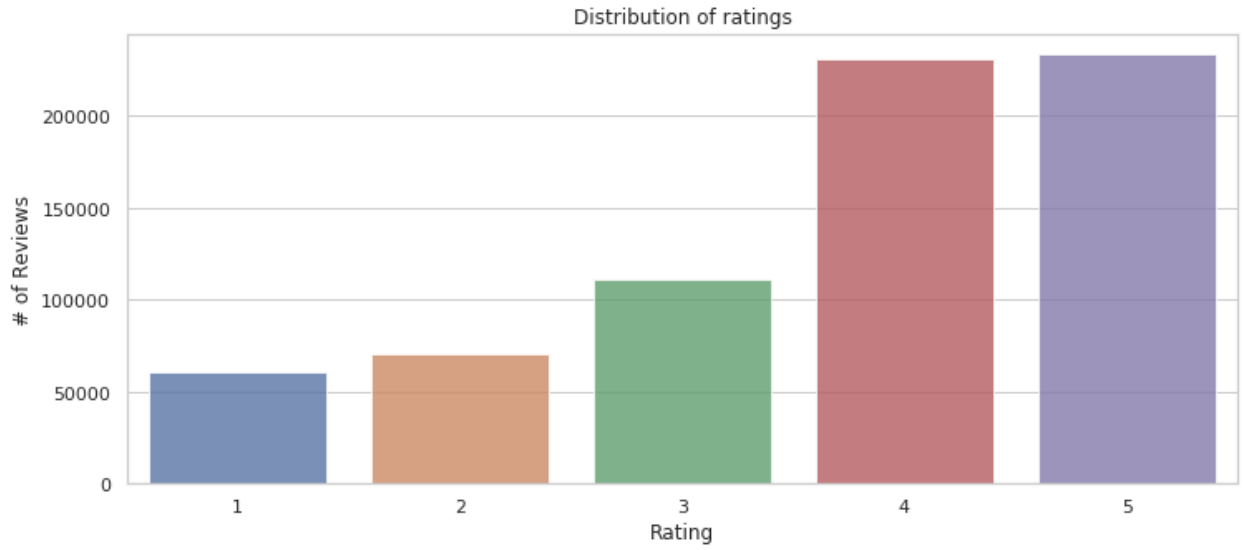
The business dataset is composed of 42153 records. But it contains other types of business besides restaurants. After filtering only restaurants, we have 14303 records. The reviews dataset is composed of 1125458 records, 706646 of them are about restaurants.

Below I plotted the distribution of restaurants in the top 15 categories in number of restaurants and in number of reviews. I also plotted the distribution of ratings.

One interesting thing to note is that “Fast food” restaurants are the 3rd more common, but that category is not among the top 15 reviewed categories. Another interesting fact is that the distribution of ratings is very concentrated in the positive side (4 and 5 stars).


Top 15 restaurant categories (in occurrences)
# of Restaurants
1750
1500
1250
1000
750
500
250
0
Mexican
American (Traditional)
Fast Food
Pizza
Sandwiches
Nightlife
Bars
Food
American (New)
Italian
Chinese
Burgers
Breakfast & Brunch
Japanese
Delis
Category


Top 15 reviewed restaurant categories (in occurrences)
# of Reviews
80000
60000
40000
20000
0
American (New)
Nightlife
American (Traditional)
Bars
Mexican
Breakfast & Brunch
Italian
Pizza
Food
Sandwiches
Steakhouses
Burgers
Japanese
Chinese
Sushi Bars
Category



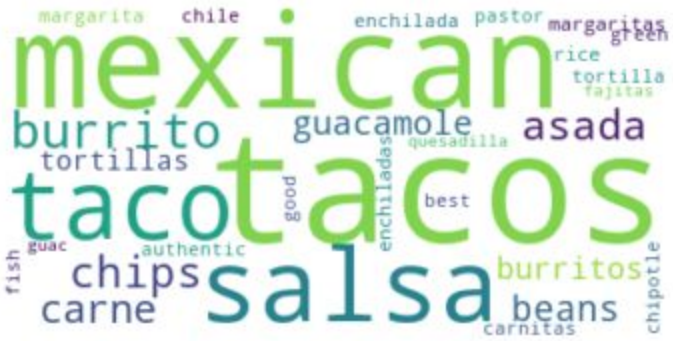
Task 1.1

Use a topic model (e.g., PLSA or LDA) to extract topics from all the review text (or a large sample of them) and visualize the topics to understand what people have talked about in these reviews.

Here I applied the LDA topic model with 10 topics. I used the full restaurant review database (706646 reviews) and the input was from a TF-IDF vectorizer limited in 20000 features (words). I had to apply that limit because the available machine was not able to complete the LDA algorithm with much more than 20000 features (memory limitation). I also applied some text normalization as making every word lower case and removing http links from the reviews.

The 10 topics are shown below with word clouds, with font size proportional to word relevance on the topic. It is very interesting to note that almost all of the topics automatically generated were somehow semantic related.

TOPIC 1 - mexican food

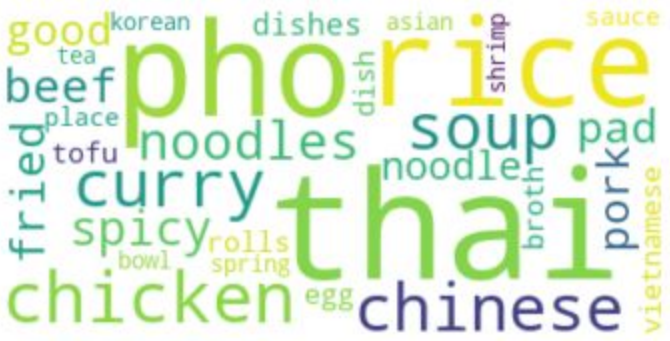


TOPIC 2 - good service / happy hour

TOPIC 3 - greek food?

TOPIC 4 - pizza / italian!

TOPIC 5 - asian food



TOPIC 6 - burgers

TOPIC 7 - good service. Related to topic 2

TOPIC 8 - japanese food

TOPIC 9 - brunch / breakfast

TOPIC 10 - classic restaurants

Task 1.2

Do the same for two subsets of reviews that are interesting to compare (e.g., positive vs. negative reviews for a particular cuisine or restaurant), and visually compare the topics extracted from the two subsets to help

understand the similarity and differences between these topics extracted from the two subsets. You can form these two subsets in any way that you think is interesting.

For this task, I wanted to compare positive vs. negative reviews from italian restaurants. So I filtered the dataset with records containing the "Italian" category and I used only 1 star and 5 stars rating for filtering negative and positive reviews respectively. Because the dataset is already very specific this time, I reduced the number of topics from 10 to 5.

One very interesting result is that in positive reviews people like to nominate the staff that is doing great, showing that people really makes a difference in the business. It is not only the food. In the negative reviews, I was very surprised to see possible problems with minority groups, that can turn out to be a very serious issue.

Below are the word clouds for negative and positive reviews for italian restaurants:

Negative

Topic 1 - sanitary problems / small portions

Topic 2 - problem with minority groups?

Topic 3 - maybe only bad food

Topic 4 - bad service

Topic 5 - sanitary problems?

Positive

Topic 1 - Excellent waiters or chefs nominated?

Topic 2 - Good food and good service

Topic 3 - good waitresses or chefs?

Topic 4 - beer with giardiniera appears to be a good combination :)

Topic 5 - more italian names cited

Conclusion

This project was very interesting, mainly because it was required to handle a real world big dataset showing many challenges on how to deal with large number of unstructured data. The results I got was definitely useful, because it could give good insights to the data with minimal human intervention.

In the whole database, the model grouped semantic related topics showing many distinct aspects of the dataset such as cuisine dishes and restaurant types even with no more information other than the review texts. When the data was filtered in a very contrasting way (negative vs positive reviews), other aspects are shown that could give insights on what are the main aspects that lead customers to really like or dislike a place.