Business Recommendation Engine

Elvin Cheng, Julia Favis, Jason Fujii, Kristina Fujimoto, David Pogrebitskiy
Northeastern University, Boston, MA, USA
https://github.com/pogrebitskiy/Business_Recommend

Abstract

In this project, our group created a business recommendation engine using data from Yelp, a popular user-review website of various businesses. The goal of our business recommendation engine is to be able to input certain businesses (restaurants, chiropractors, auto-repair, etc.) that are liked by the user and our engine will recommend similar businesses based on metrics calculated from factors such as reviews and distance. We were able to accomplish this goal by cleaning the data in python using PyMongo then using MongoDB to build a network for the recommendation engine. Our hypothesis is that this engine will be able to recommend well rated/reviewed businesses to the liking of the user successfully. We were able to create our engine with thorough analysis of the reviews and compatibility scores between other businesses and the searched business. The business recommendation engine is able to take in a business ID and a distance value and provide the user with all the businesses in that area sorted by rating of the business and distance.

Introduction

Online business review platforms have grown significantly in popularity over the past few years. There are significant high-quality businesses waiting to be discovered, but it can be difficult for users to effectively filter through the abundance of information and choose which one would best suit their needs. In this project, our group wanted to address this prompt by creating a business recommendation system that could provide users recommendations based on a selected business/establishment that the user inputs. Our working hypothesis is that by analyzing the Yelp review dataset, we could create an insightful engine that created recommendations to users based on similarity factors and reviews. Furthermore, we hypothesized that this engine would be predominantly more useful than typical ‘similar businesses’ recommendations online. Some reasons as to why this project is significant include: the enhanced recommendations, the increased efficiency, the scoring of businesses, and data analysis capability for business owners. Our recommendations are enhanced because our engine recommends based on individual preferences, accounting for similar business categories and proximity, while emphasizing highly-upvoted reviews and total review counts. Furthermore, our recommendation engine helps users save time from scrolling through countless similar businesses, instead determining similar recommendations that are well rated and sorting them by overall score which is determined by types of reviews and helpfulness of reviews. Finally, business owners can use our engine to understand user preferences and trends, and determine which factors can help attract more customers. They can also utilize our engine to look at their comprehensive score which takes into account their review count, review type, and usefulness of reviews. Our primary focus when building our business recommendation engine was to create an easier way to find well rated and reviewed businesses similar to ones that users are familiar with.

Methods

We acquired our data from Yelp through their review dataset, which can be accessed as public files from their website. From that data, we imported business.json into a Business collection and review.json into a Review collection in a MongoDB database called Business_reccomend. We imported them as they are so that we could do the preprocessing via MongoDB queries and aggregations.

Our data pre-processing was done in Python via PyCharm utilizing PyMongo. Our first pre-processing step was to calculate the adjusted usefulness score for each review, and add it to the review's entry in the Review collection. The adjusted usefulness score is calculated by determining if the rating of the business is above or below 3 stars; subtracting the log of the number of upvotes if it's below 3 stars to lower the score, and adding the log of the number of upvotes if it's above 3 stars to raise the score. Our next step was to add a credibility score to each review, calculated by multiplying the average adjusted score of the dataset by the review count, and dividing that by the review count + 5 (to account for if a review count was 0). We also aggregated the given longitude and latitude fields into a coordinate field in order to gain the ability to do geospatial analysis. We utilized secondary indexing to create a coord_2dsphere index for the business collection and a business_id index for the review collection.

Analysis

Our data consists of about 7,000,000 reviews and about 150,000 businesses from across the U.S. and Canada. Each business contains its MongoDB generated ObjectID, its Yelp business_id, name, address, city, state, postal code, latitude and longitude (distribution in Fig 3), star rating (distribution in Fig 4), review count, is_open boolean, attributes (which consists of various boolean values based on business type, i.e. BusinessAcceptsCreditCards for a restaurant or ByAppointmentOnly for a doctor), categories, hours, average_adj_score (distribution in Fig 1), and credibility_score (distribution in Fig 2).

Each review has its MongoDB generated ObjectID, Yelp review_id, Yelp user_id, Yelp business_id, star rating, usefulness rating, funniness rating, coolness rating, the text, date, and adjusted score. These images will visualize the distribution of different metrics from our data.

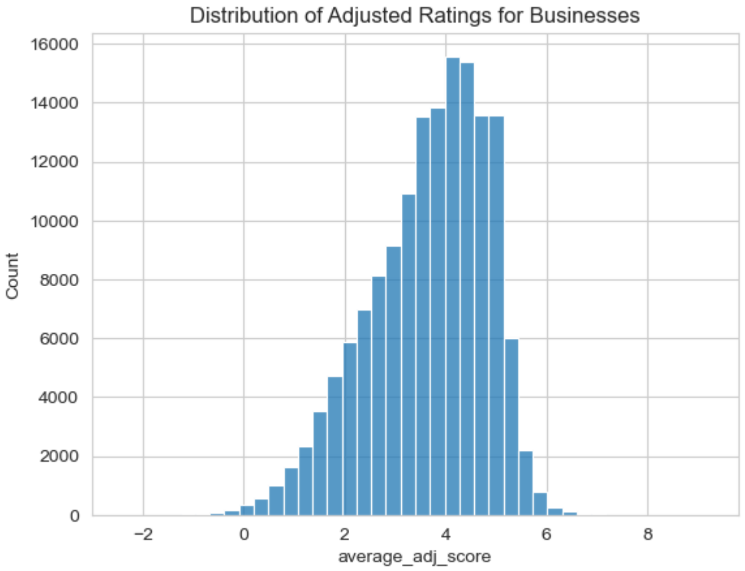
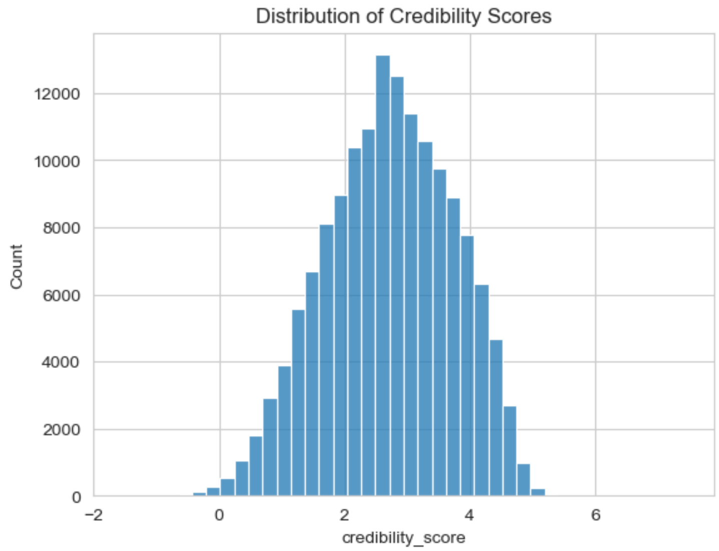


Fig. 1

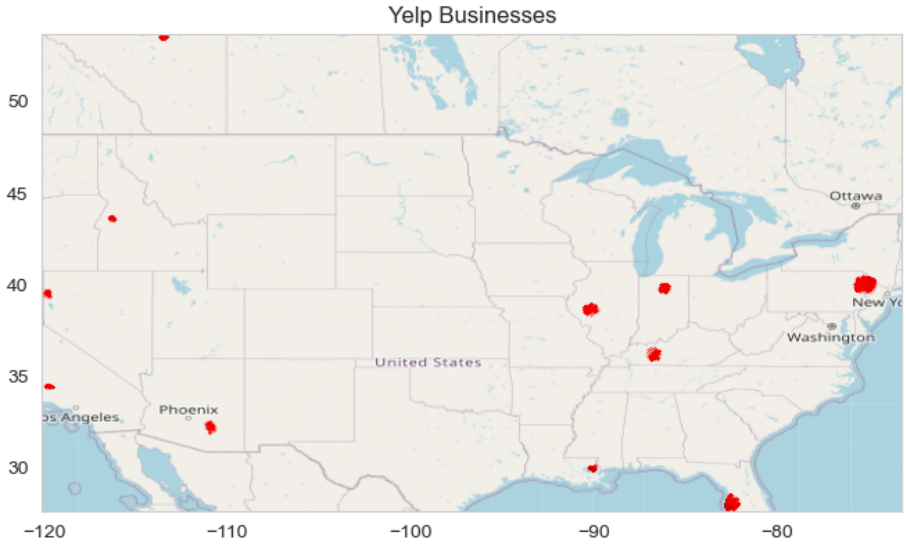
This visualization shows our calculated adjusted rating for each business in the dataset. This adjusted rating is calculated using our formula to weigh each review differently based on its usefulness and stars.

Fig. 2



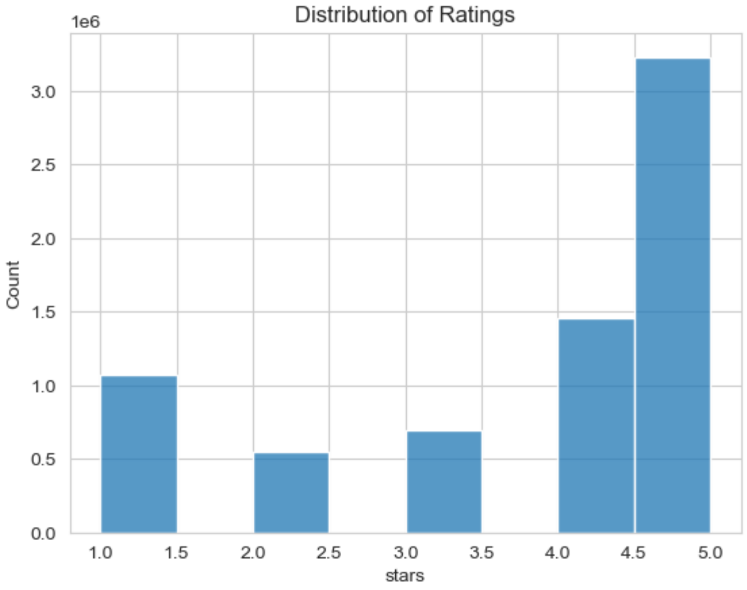
This visualization shows us the distribution of the credibility scores of each business in our Yelp dataset. It is normally distributed, which is expected, since most businesses will be rated quite well with less of them rated poorly or extremely well. As we can see from this graph, the average credibility score for businesses is around 3.

Fig. 3



This visualization shows the distribution of our dataset across the United States and Canada. As seen based on the map, there is only data for businesses in select cities.

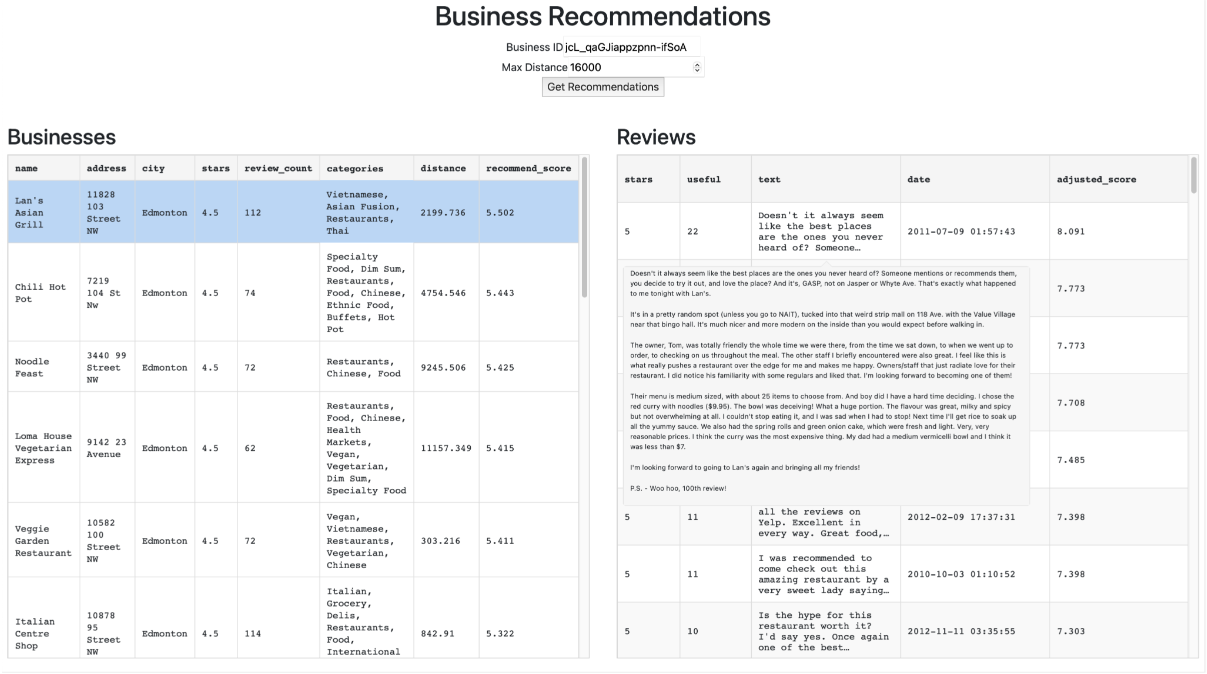
Fig. 4



This visualization shows the distribution of each business' overall star rating. As seen, most restaurants receive pretty high ratings with a typical 4-5 stars. Only a few ratings are negative at 1-1.5 and 2-2.5 stars.

The outcomes of our project include the functionality of generating recommended businesses given the business ID of any business in the dataset. The recommendations are based on our recommend_score, which is calculated by increasing the credibility score by 0.5 multiplied by how many categories the business has in common with each business in the location; the more the categories in common, the higher the score. In our UI (Fig. 5), a user can enter a unique business ID from the dataset and generate recommendations, from the list of recommendations, the user can select a specific business and see the most relevant reviews for it. If the user hovers over the row for a specific recommendation, they can see the text of the review/

Fig. 5



Business Recommendations

Business ID jcL_qaGJiappzpnn-ifSoA
Max Distance 16000

Businesses

name	address	city	stars	review_count	categories	distance	recommend_score
Lan's Asian Grill	11828 103 Street NW	Edmonton	4.5	112	Vietnamese, Asian Fusion, Restaurants, Thai	2199.736	5.502
Chili Hot Pot	7219 104 St Nw	Edmonton	4.5	74	Specialty Food, Dim Sum, Restaurants, Food, Chinese, Ethnic Food, Buffets, Hot Pot	4754.546	5.443
Noodle Feast	3440 99 Street NW	Edmonton	4.5	72	Restaurants, Chinese, Food	9245.506	5.425
Loma House Vegetarian Express	9142 23 Avenue	Edmonton	4.5	62	Restaurants, Food, Chinese, Health Markets, Vegan, Vegetarian, Dim Sum, Specialty Food	11157.349	5.415
Veggie Garden Restaurant	10582 100 Street NW	Edmonton	4.5	72	Vegan, Vietnamese, Restaurants, Vegetarian, Chinese	303.216	5.411
Italian Centre Shop	10878 95 Street NW	Edmonton	4.5	114	Italian, Grocery, Delis, Restaurants, Food, International	842.91	5.322

Reviews

stars	useful	text	date	adjusted_score
5	22	Doesn't it always seem like the best places are the ones you never heard of? Someone…	2011-07-09 01:57:43	8.091
				7.773
				7.773
				7.708
				7.485
5	11	all the reviews on Yelp. Excellent in every way. Great food,…	2012-02-09 17:37:31	7.398
5	11	I was recommended to come check out this amazing restaurant by a very sweet lady saying…	2010-10-03 01:10:52	7.398
5	10	Is the hype for this restaurant worth it? I'd say yes. Once again one of the best…	2012-11-11 03:35:55	7.303

Doesn't it always seem like the best places are the ones you never heard of? Someone mentions or recommends them, you decide to try it out, and love the place? And it's, GASP, not on Jasper or Whyte Ave. That's exactly what happened to me tonight with Lan's.

It's in a pretty random spot (unless you go to NAIT), tucked into that weird strip mall on 118 Ave. with the Value Village near that bingo hall. It's much nicer and more modern on the inside than you would expect before walking in.

The owner, Tom, was totally friendly the whole time we were there, from the time we sat down, to when we went up to order, to checking on us throughout the meal. The other staff I briefly encountered were also great. I feel like this is what really pushes a restaurant over the edge for me and makes me happy. Owners/staff that just radiate love for their restaurant. I did notice his familiarity with some regulars and liked that. I'm looking forward to becoming one of them!

Their menu is medium sized, with about 25 items to choose from. And boy did I have a hard time deciding. I chose the red curry with noodles ($9.95). The bowl was deceiving! What a huge portion. The flavour was great, milky and spicy but not overwhelming at all. I couldn't stop eating it, and I was sad when I had to stop! Next time I'll get rice to soak up all the yummy sauce. We also had the spring rolls and green onion cake, which were fresh and light. Very, very reasonable prices. I think the curry was the most expensive thing. My dad had a medium vermicelli bowl and I think it was less than $7.

I'm looking forward to going to Lan's again and bringing all my friends!

P.S. - Woo hoo, 100th review!

Conclusions

Our business recommendation engine does a great recommendation job. By taking in a business ID and distance value, the engine is able to locate all the businesses that match the category field of the business searched, sort by a comprehensive score that takes into account the business' reviews, and then return the list of all the businesses within the specified distance range. We were able to successfully test our business recommendation engine and received some useful businesses. In addition, we created a user interface that makes it more friendly for users to use. There is an input field for business ID and distance. When the user clicks on search, the business recommendation engine outputs all its recommendations to the user on this interface. While our project provides a comprehensive analysis of each business by providing an overall score that takes into account all the reviews, our recommendation engine also has some limitations. One limitation that we had in our engine was that we currently must input a business ID into the UI or the MongoDB script rather than a business name. This prevents typical users looking for businesses from using it without our help as it requires us to search the business ID that interests the user because they are specific to this dataset. In the future, we would work on improving to make our business recommendation engine more user-friendly by allowing individuals to search by inputting a business name or address rather than an ID. This will make our engine more useful to a wider variety of people. A limitation we would have with using business names as a search in our engine would be chain restaurants such as McDonald's. These types of businesses have the same name at different locations so this would cause an issue with finding similar businesses within the distance measurement entered by the user. Since the business ID allows for targeting only one location of chain restaurants, using the business name would cause confusion with which location to select for the recommendation analysis. Another limitation of our business recommendation engine is the diversity of our data. Our data is only focused in certain cities/states of the United States and Canada while in other areas there is only one business recorded. This provides a limit to where our users can be located and in the future we would implement a broader dataset that covers many more areas throughout the United States and Canada. This would allow more users around the region to use our engine and provide these users with more thorough recommendations. One thing about our business recommendation engine is that it recommends businesses that are of similar type to the business of interest. For example, if you input a health clinic/spa type of business, the recommendation engine will provide other health clinic/spa businesses in the same area. This allows for making accurate recommendations because when you are looking for another business that is similar to the one of interest, you wouldn't want to get businesses that aren't related to your needs. For example, if you entered a health clinic/spa, but the recommendation engine returns healthy restaurants, that wouldn't be useful to the user. In the future, we would want to expand past these capabilities. While our recommendation engine completes the tasks by including businesses similar to the input business, we would want to have possible expansion to recommend businesses from other genres that would be of interest to the user. For example, if a user inputs a bookstore, we would want to not only include other bookstores but possibly businesses such as libraries and cafes. This would require computing the similarity of businesses using the category field and finding a formula that would best fit the data.

Author Contributions

We collectively worked together to pick the datasource carefully, as we wanted more numerical values in our data for our analysis. Once we selected the Yelp dataset, we decided as a group on our approach for the recommendation engine and how we would clean the dataset. Then, David and Elvin primarily did the data cleaning and PyMongo querying while Jason, Julia, and Kristina worked primarily on writing the report and creating the presentation.

References

Yelp open dataset. Yelp Dataset. (n.d.). Retrieved April 6, 2023, from https://www.yelp.com/dataset