Restaurant Recommendation Algorithm in Neo4j

Nicolas Chapados, Jake Fuiman, Ethan Gu, Mark Hurley, Daniel Vahey
Northeastern University, Boston, MA, USA

Abstract

For this project, we created a restaurant recommendation engine for the Santa Barbara area that asks a user for the name of a restaurant they like, then outputs ten restaurants our algorithm found to be similar, highly-rated, and nearby. We determined the similarity scores with the algorithm described below, using the categorical data about the type of food, the location, and their normalized rating score. All outputted restaurants have at least a 4 star rating and are at least 10 miles away from the inputted restaurant. As an example, we used the restaurant "Finch & Fork" as an input, and our recommendation engine outputted ten restaurants that all had at least a four star rating, were within ten miles, and shared at least five attributes with "Finch & Fork". We were able to conclude that our Yelp recommendation engine was a success at being able to recommend similar, high quality, easily accessible restaurants to people based on the restaurants they enjoy.

Introduction:

Picking a restaurant can be a very tedious task. Whenever you visit a new restaurant you always run the risk of the service or food being below your standards. In order to avoid this risk, oftentimes a customer will continuously visit the same restaurants over and over simply because they know that they will be content with the results of their experience. The problem with this is that people never get the chance to experiment with other restaurants and they lose the opportunity of new positive experiences. Additionally, newer or smaller restaurants don't get as much business. People understand this tradeoff, but an efficient solution doesn't exist. This is where the recommendation algorithm comes in. There are many factors that go into picking a restaurant, and the algorithm understands this and prioritizes those that are important to its users. It was built using data from thousands of restaurant reviews and the consideration of a restaurant's features such as geography, rating and restaurant type, with the help of Yelp's online datasets.

Methods:

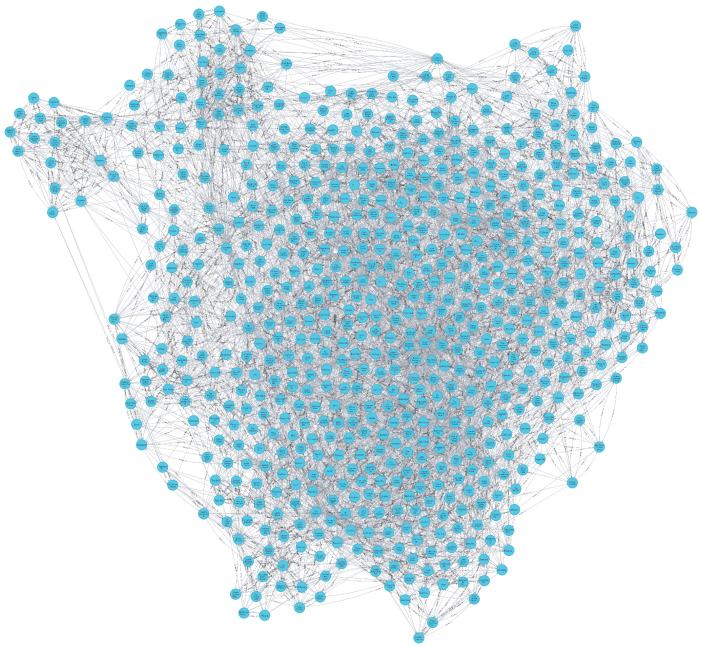
The datasets we decided were necessary for our analysis were a Yelp reviews csv and a Yelp businesses csv that we got from Yelp's website. These datasets contained review and business data for 11 different metropolitan areas. Since each metropolitan area had so many restaurants and businesses, we decided to just focus our analysis on restaurants in the Santa Barbara area.

First, we needed to filter and clean our business data to only include certain columns for all restaurants and food-related businesses in Santa Barbara. After filtering and cleaning our business data, we can then merge with our review dataset through a shared business_id column. Since each restaurant and food-related business has many reviews due to the nature of Yelp, we also needed to do a groupby on business_id and stars as well as a count so that we know how many 0, 1, 2, 3, 4, and 5 star reviews each business has to use in our comparisons as well as gain the ability to find an average rating for entry. Our final, cleaned data contains the following information for each restaurant/food business; address, take-out (T/F), business_id, categories (what kind of restaurant/food), latitude, longitude, name, review count, stars, star_count, and dates (for each review).

Given that the objective of our analysis is to recommend restaurants to a user based on other restaurants that they enjoyed using neo4j, we will also need to compute a distance metric between restaurants to create edges in our database. The first step of this is to determine which restaurants are within 10 miles of each restaurant since we want our recommendations to be easily accessible for a user. As for the similarity score, we wanted to take into account the number of reviews a restaurant/food business received, its star rating, and the categories each restaurant fell under (cuisine, type of restaurant, vegetarian, etc.). We first initialized our similarity score by the number of categories overlapped between restaurants. However, since restaurants can vary greatly in size and popularity the values for number of reviews and star rating can fluctuate greatly and we don't want outliers skewing our recommendations, preventing smaller businesses from showing up. In order to normalize the similarity scores between restaurants, we used the formula:

similarity score = (log(review count)/1.5 * stars)/12

This way we can properly return highly rated restaurants based on star rating, and still take into consideration how many reviews a restaurant received without letting it overpower the similarity score. These are the scores that we stored to use as edges between restaurants within our neo4j database, and this is what our overall graph looked like:



Analysis:

We wanted to be able to provide a user-interface that would allow any given user to use our project.

```
---------------------------------
 Connection to neo4j successful
---------------------------------
This recommendation engine only works for restaurants in the Santa Barbara area. Some good starting points are 'Finch & Fork', 'Su
Casa Fresh Mexican Grill', and 'Chase Restaurant'

Enter a restaurant, or 'exit' to exit: not a restaurant
---------------------------------
That restaurant is not in the database
---------------------------------

Enter a restaurant, or 'exit' to exit: Finch & Fork
---------------------------------
Uncorked Wine Tasting and Kitchen, Rating: 5.0, Distance from Input:  0.72 miles
Barbareño, Rating: 4.5, Distance from Input:  0.16 miles
Finney's Crafthouse, Rating: 4.5, Distance from Input:  0.87 miles
Lure Fish House, Rating: 4.5, Distance from Input:  2.9 miles
Eureka!, Rating: 4.0, Distance from Input:  0.18 miles
Benchmark Eatery, Rating: 4.0, Distance from Input:  0.24 miles
Scarlett Begonia, Rating: 4.0, Distance from Input:  0.31 miles
Jill's Place, Rating: 4.0, Distance from Input:  0.4 miles
Bluewater Grill - Santa Barbara, Rating: 4.0, Distance from Input:  0.93 miles
Yellow Belly, Rating: 4.0, Distance from Input:  1.61 miles
---------------------------------
```

Here is an example of what the front-end of our recommendation engine looks like. Upon connection to our neo4j database, the user is prompted to enter a restaurant that they enjoy (if it is not in the database, the algorithm will prompt the user again). A user would then be returned a list of 10 restaurants that our algorithm deemed similar to the original restaurant, Finch & Fork in this case, along with each restaurant's distance from the inputted restaurant and its star rating on Yelp.

As we can see from our output, our recommendation engine returned 10 restaurants in the vicinity of Finch & Fork. Given that our objective is to recommend highly-rated, similar restaurants, as we can see through our similarity score calculations, it is expected that all of our recommendations have star ratings of 4 and higher. Upon taking a closer look at our recommended results when inputting Finch & Fork:

name	review_count	stars	categories
Uncorked Wine Tasting and Kitchen	146	5.0	{'Wine Tasting Room', 'Tapas/Small Plates', 'A...
Barbareño	405	4.5	{'Diners', 'Specialty Food', 'Sandwiches', 'Am...
Finney's Crafthouse	1047	4.5	{'Gastropubs', 'Nightlife', 'American (New)', ...
Lure Fish House	1453	4.5	{'Breakfast & Brunch', 'Beer', 'Wine Bars', 'S...
Eureka!	961	4.0	{'Burgers', 'Nightlife', 'American (New)', 'Am...
Benchmark Eatery	544	4.0	{'Breakfast & Brunch', 'Vegetarian', 'Seafood'...
Scarlett Begonia	897	4.0	{'Breakfast & Brunch', 'Beer', 'Cafes', 'Food'...
Jill's Place	325	4.0	{'Burgers', 'Delis', 'Steakhouses', 'Sandwiche...
Bluewater Grill - Santa Barbara	768	4.0	{'Breakfast & Brunch', 'Seafood', 'Venues & Ev...
Yellow Belly	337	4.0	{'Breakfast & Brunch', 'Pizza', 'Nightlife', '...

What immediately stands out is the star rating column, where we can see that higher star ratings are recommended frequently as it is a good indicator of a restaurant’s food and service. Though not fully visible in the table above, we can also see that through the categories column, similar restaurants in terms of type and cuisine are recommended. Finch & Fork, our inputted restaurant, falls under the following categories (Breakfast & Brunch, Nightlife, American (New), American (Traditional), Bars, Restaurants). All of the output recommendations share at least 5 of the same attributes, which demonstrates how our recommendation engine ensures that suggested restaurants will have similar food (American) and/or experience (Breakfast & Brunch, Bars, Nightlife) as the inputted restaurant that a user enjoyed.

Conclusions:

Overall, we believe that our Yelp recommendation was a success, and can be a very useful tool for anyone who might want to try out a new restaurant without the risk of blindly choosing. We were able to recommend restaurants similar to and in the general vicinity of a user-inputted restaurant quite successfully. As for where we would expand on this project given more time and data, the first thing would be to expand its capabilities to more major cities and metropolitan areas so that our recommendation engine has more diverse uses for people who may not be near the Santa Barbara area. Another area for growth could be to allow a user to input categories that they are interested in, in the case that they might be new to an area and might not yet have a preferred restaurant or just want to try a restaurant completely new.

Author Contributions:

Nicolas Chapados: Cleaned data, created database, created recommendation algorithm
Jake Fuiman: wrote report
Ethan Gu: wrote report + created visuals
Mark Hurley: wrote report
Daniel Vahey: created main program driver / user interface