"I Want To Start A Business": Getting Recommendation on Starting New Businesses Based on Yelp Data

Project Final Report

Rajkumar, Balaji Ambresh
balaji.ambresh@nym.hush.com
(05929421)

Fouladighaleh, Sadjad
sadjad@cs.stanford.edu
(06043889)

Ghiyasian, Bahareh
bghiyasi@stanford.edu

December 2015

1 Introduction

Starting a new business can be a challenging task. Aside from logistic issues, such as facilities and supplies, the preliminary question of *'what business to start?'* seems to be one of the most important matters that should be addressed.

There are various parameters that can affect a business' success. Competing businesses, social structure, cultural considerations and marketing techniques are some of the factors that can contribute to the success or collapse of a business. As the number of these variables increases, it gets swiftly harder to find a dependable answer to this question. However, paying attention to the customers' behavior can help us to have a better insight on this question. For example, if people from a neighborhood travel a long distance to go to a bookstore, one can understand that starting a bookstore closer to those people has a great potential of success.

In this project, we are going to answer the following question:

> *Given an area in a city, what new businesses in that area are capable of attracting most customers?*

For us to be able to answer this question, we first need to clarify the terminology used in this definition. We start by defining *neighborhood*. A neighborhood is usually used to refer to *'the area or region around or near some place or thing'* or *'a district or locality, often with reference to its character or inhabitants'*. In the case of this project, we refer to a neighborhood as *a set of businesses that a group of people tend to go to almost exclusively*. Please note that this notion of neighborhoods is different from the *actual* neighborhoods in a city. However, we expect people prefer businesses that are located closer to them. Therefore, we believe that the neighborhoods by

our definition should coincide with actual neighborhoods in the area. Finding these neighborhoods is a vital part of understanding customers' behavior[1].

By a new business, we mean the category of that business, not the brand name. Therefore, our system will suggest terms like *coffee shop* or *hair salon*, instead of *Starbucks* or *Supercuts*.

We make suggestions for new businesses in two different categories. The first category is the non-existent businesses in the current neighborhood that people need to go to another neighborhood for those. However, this might not cover all of the potentially-successful businesses. For example, in a particular neighborhood people might not be interested in Fusion Food, simply because there is no Fusion Food restaurant around. If we see that a business is popular in adjacent neighborhoods, we can suggest that as a new neighborhood. This kind of suggestions form the second category of our output.

Other categories of suggestions are possible, as well, but we decided not to explore them in this work. For example, imagine a neighborhood with an average-rated, but popular coffee shop. A coffee shop with better service and without the negative points of the first coffee shop has the potential of being successful. By analyzing the reviews, we can extract the characteristics that can make this new coffee shop successful. Such suggestions are outside of the scope of our work.

2 The Dataset

Yelp provides a dataset of about $61,000$ business as a part of Yelp Dataset Challenge. The major characteristics of this dataset, as appeared on the website, is as follows:

1. 1.6M reviews and 500K tips by 366K users for 61K businesses,
2. 481K business attributes, e.g., hours, parking availability, ambiance,
3. Social network of 366K users for a total of 2.9M social edges,
4. Aggregated check-ins over time for each of the 61K businesses.

These data are in JSON format, so we decided to use MongoDB as the storage. Fortunately, MongoDB supports spatial queries over data, which is very useful to us.

We will treat each city in the Yelp dataset separately. Table 1 shows the cities with most number of businesses in this dataset.

3 Extracting Neighborhoods

As we discussed earlier, understanding customers' behavior is crucial. To do so, we decided to group businesses into clusters that we call *neighborhoods*. Each neighborhood is a group of businesses that a group of customers visit almost exclusively. Although there is no notion of location in this definition, we expect businesses in a neighborhood to be close together, geographically. The reason is that people tend to go to businesses that are close to them, rather than travel a farther distance.

The Yelp dataset, actually contains a neighborhood property for some businesses, which is a list of actual neighborhoods that a business belongs to. However, there are several problems associated

[1] We have to emphasis on the fact that the neighborhood in this definition is a notion to group businesses together and is different from the *dictionary* definition of a neighborhood.

Table 1: Top cities based on number of businesses in Yelp dataset

City	# of businesses	# of categories
Las Vegas, NV	13600	664
Phoenix, AZ	8410	620
Scottsdale, AZ	4039	547
Edinburgh	2930	352
Pittsburgh, PA	2724	418
Mesa, AZ	2347	462
Tempe, AZ	2258	475
Henderson, NV	2130	434
Montreal	1870	297
Chandler, AZ	1867	430

with using this property to group businesses together. First of all, a lot of businesses in Yelp dataset does not have a neighborhood assigned to them. Additionally, these neighborhoods don't necessarily cover the whole city. One other important point is that the actual neighborhoods do not capture the customers behavior properly.

To extract the neighborhoods of our own definition, we decided to use community detection. To do so, first we need to form a graph of businesses, where an edge between two business A and B means that these two businesses have (enough) common customers.

The basic way to understand if two businesses have a common customer is by using the user reviews. If user u wrote (good) reviews for both business A and business B, we can consider that user as a common customer. We limit these edges only between businesses toa single city, and only consider reviews with rating ≥ 3.

As the edges are only between businesses in the same city, we create the graph and find communities separately for each one. For example, for *Pittsburgh, PA*, the graph has the following properties:

- 2,724 nodes (businesses),
- 454,053 edges,
- 112 nodes with degree zero.

By considering *tips* as another signal, we can add more edges to this graph. If a user u writes tips for a business A we take that person as a customer of business A. One important point here is that not all of the tips are positive.

To overcome this problem with negative tips, we need some kind of sentiment analysis. Tips are usually shorter than reviews and are comprises of a few short sentences. We used Sentiment140 to exclude negative tips from this.

In our preliminary tests, we found out that this graph structure is not enough to extract neighborhoods as we expected. For example, there are people who wrote reviews for businesses in different parts of the city. Therefore, changes need to be done to strengthen the community structure of the graph. One way to fix this is to remove edges between businesses that are located far from each other (we are currently using a threshold of 2km). Applying this will result in a graph with $296,623$ edges.

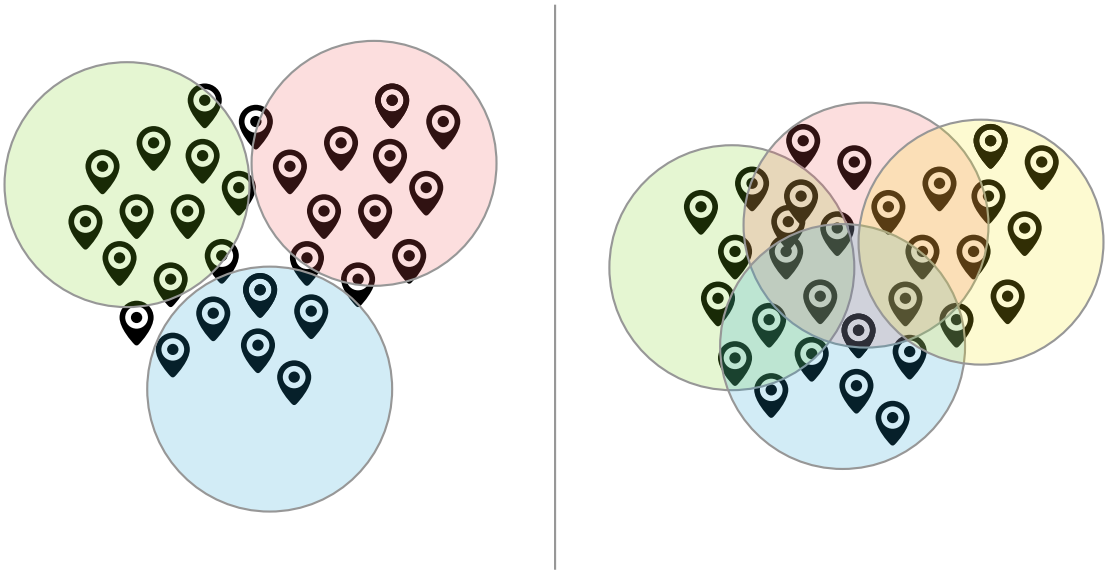

Figure 1: What we expected from neighborhoods (left) and what they actually could be like.

3.1 Results on extracting neighborhoods

Unfortunately, our efforts to extract neighborhood were not successful. At first, we thought that it could be a result of extra edges inside the graph. So, we tested different ideas to prune the businesses graph, such as removing edges between far businesses and increasing the required number of common reviews to create an edge between businesses. We also tried different algorithms for community detection such as Clauset-Newman-Moore (which tries to maximize modularity) and Girvan-Newman (which is based on betweenness centrality), but we couldn't extract a good community structure from this graph and most of the nodes (%80-95) ended up in a same community.

This problem is inherent in how we defined neighborhoods. We expected to be able to group businesses as Figure 1-left, but the customers' behavior is much more similar to Figure 1-right. Two customers who are living one or two blocks away from each other can have a set of common businesses that they visit, in addition to sets of businesses that only one of them visits. Also, people might visit businesses not only close to where they live, but also close to where they work or go to school.

The benefit of extracting such clusters is that you can actually observe the behavior of a certain group of customers and find the missing pieces in those clusters. However, it seems that the set of current data in Yelp dataset is not enough for that. The only way to associate a user to a business is to use the review and tip data, and it's not known that if the review or tip is based on a frequent or a one-time visit. On the other hand, if the check-in data per user is available, it is possible to exploit that data to extract so much more information about customers behavior.

As the community detection does not yield a good result for our work, and generating suggestions is separate from community detection part, we decided to use a notion of dynamic neighborhood. The user of our system can define a point for starting their new business, and we can assume a radius R around that point as the neighborhood. All of the businesses in the city can be grouped into two different sets: inside and outside of that neighborhood.

4 Suggesting Businesses

After extracting the neighborhoods in the graph, it is now time to suggest new businesses for a specific neighborhood. As we discussed before, the suggestions come in two different categories. The first category is based on missing businesses in the neighborhood which there is a demand for them. The second category is formed from the businesses that are successful in other neighborhoods. The final decision of choosing a business from these two lists is a problem that we do not address in this work.

4.1 First Category: Missing Categories in the Neighborhood

It is worth mentioning that from the previous part, we assigned each business to a neighborhood. Additionally, we know businesses favored by a customer based on reviews and tips. We assign each customer to the neighborhood that he/she has the most reviews and tips about businesses in that neighborhood.

Now, to determine the popular businesses outside of a given neighborhood, we form a graph of customers and business categories. Customers who belong to the neighborhood are considered. Categories of businesses they go to within their neighborhood are considered to the local categories. We find the categories for all of the businesses the local people go to, then we remove the categories that already exist in the neighborhood. In addition to these category nodes, we also add customers assigned to the same neighborhood. For each business category outside of the neighborhood that a customer visits, we add an edge between the customer and that business category. In other words, this graph shows what kind of businesses are visited outside of the neighborhood by the customers. In another variant of this algorithm, we also included friends of a customer as customers for a business. We demonstrated the results for both modes. Business categories then are sorted by their in-degree and it forms the list of suggested categories.

As an example, we tested this algorithm on a certain neighborhood (Capitol) in Madison, WI. Table 2 shows the results of running this algorithm on that neighborhood.

Table 2: Top businesses located outside of the Capitol neighborhood in Madison, WI that are frequently visited by people belonging to that neighborhood.

Rank	Category
1	Restaurants
2	Food
3	Nightlife
4	Bars
5	American (New)
6	Breakfast & Brunch
7	American (Traditional)
8	Tea
9	Bakeries
10	Shopping

4.2 Second Category: Top Categories in Other Neighborhoods

For the second category of suggestions, we plan to use a basic approach. We form a graph similar to previous part, except for all of the customers and business categories within a city. By excluding the categories that have customers from the current neighborhood, we will have a list of business categories that have no customers in the current neighborhood, but are popular outside of the neighborhood. They form the second category of our suggestions for starting a new business.

5 Evaluation

This section focuses on evaluating the goodness of recommending the most popular categories in terms of precision and recall. Baseline signal is the check-in count, which means that businesses with more number of check-ins are considered more successful and popular. The neighborhoods in 3 were used for evaluating the goodness of recommendations:

Table 3: Neighborhoods used from Yelp dataset to evaluate the proposed method.

Location	**Users**	**Businesses**
NC, Charlotte, Derita	550	200
PA, Pittsburgh, Strip District	1794	155
WI, Madison, Capitol	3464	451
NV, Henderson, Southeast	4666	1183

For each of these neighborhoods, we considered first k businesses ($1 \leq k \leq 10$) and compared the set of suggestions to our baseline, which is based on check-ins. Figure 2 shows these results for these neighborhoods.

5.1 Consideration for all experiments

1. We consider the users that are present in a neighborhood.
2. Friends of local users are restricted to those within the state and city since the behavior is expected to be much closer than a long distance friend.
3. Businesses are considered as long as they belong to the same state / city but doesn't belong exclusively to the user neighborhood. This is done to simplify the businesses that are visited outside the same city.
4. Tips are considered for businesses that match the condition above and the tip sentiment is non-negative. Negative tips produced lower precision than the rest of them. So, they were ignored.

We experimented with different combinations of the data in Yelp dataset. The different approaches are as follows:

1. Local users (customers within the neighborhood) and reviews,
2. Friends of local users and reviews,

Figure 2: Precision values for 4 different neighborhoods with different set of input data.

3. Local users, their friends and reviews,
4. Local users, non-negative tips and reviews
5. Friends of local users, non-negative tips and reviews,
6. Local users, their friends, their non-negative tips, reviews.

Table 4: Top baseline categories and the ones recommended by the best approach (the last one) within WI, Madison, Capitol.

Baseline	**Recommended**
Restaurants	Restaurants
Food	Food
Nightlife	Nightlife
Bars	Bars
American (New)	American (New)
American (Traditional)	Breakfast & Brunch
Coffee & Tea	American (Traditional)
Shopping	Coffee & Tea
Breakfast & Brunch	Bakeries
Hotels & Travel	Shopping
Grocery	Mexican

5.2 Observations

1. The combination of local users, their friends, non-negative tips and reviews produces better results than the rest of the algorithms all the time for smaller values of precision.
2. Tip data does not add too much weight to the precision. More on this in the limitations.
3. Friends have a good influence (homophily) in determining the category of businesses to start.

5.3 Limitations

Let's consider the top business categories of WI, Madison, Moorland-Rimrock. The most popular business category is related to Automobile Repairs. Let's assume that we took this automobile shop. Now, based on our system, we'd end up recommending a category like Food. This is because, the relative frequency of visits among local folks and their in-state friends would be related to food. So,it would've been great if we had user check-in data. Through this, we could have used A priori approach to infer that they ate when they visited a repair shop. That way, a strong feature inference can been made.

Tip data is sparse and we're unable to make better observations in terms of using sentiment data. Most of the tips were neutral. So, the sentiment part was useful in excluding negative tips. Again, negative tips could be used as a strong signal for recommending a bad business. Shows a strong signal for a customer need.

Evaluation is limited to data points where there is a good number of check-ins, users and businesses, but a lot of neighborhoods are sparse.

Our consideration of reviews and tips is a non-set model. So, we could have incremented the category counts based on a few users who could have reviewed businesses when they might have improved. This happens in the case of elite users.

6 Conclusion and Future Works

In this work, we focused on suggesting businesses that have a demand from people of a specific region, but are located outside of that region. As we understood, extracting neighborhoods as we defined in this project is a hard (if not impossible) task and requires more attention and data. Neighborhoods cannot be seen as separate sets that partition a city, but they should be seen as overlapping sets.

Access to the actual check-in data for users can be a lot helpful and some of the benefits are discussed in the previous sections. Access to such data will help us to understand behaviors of the customers that seem complex at first.

Having the start date of a business can be really helpful. We can set our Yelp world to a set point in time. Reviews and tips have timestamps associated with them. If check-ins have timestamps and user data, it would be a really nice way to evaluate our models.

Last, but not least, it is also interesting to look at this problem from the side of machine learning, instead of social network analysis.