Find the Top 50 Most Popular Coffee Shops in San Francisco

Yaonan Zhong, Aug 10 2015

My approach to this problem is divide-and-conquer. The following are some questions we need to answer in this project.

- How do we define the popularity of a coffee shop?
- How do we design the ranking algorithm?
- What Yelp data do we need for such ranking algorithm?
- How can we get the data?

1. Ranking algorithm

So how do we decide the popularity of a coffee shop? Here are some features we can consider:

- Yelp coffee shop rating
- Review count
- Average rating of all reviews
- The percentage of reviews with rating above 3.5
- Other statistic properties of rating distribution, such as variance, time interval
- Location
- …...

For example, the most simplest way is ranking the coffee shops by their Yelp rating. However, it may lead to some problems. First, the Yelp rating algorithm is like a black-box. We don't know the exact details. Second, the rating scale is too narrow. It is very likely that more than 50 coffee shops has the same rating such as 4.5, so we could not distinguish them. Do we just randomly pick up them? It is possible that one coffee shop has more than 1000 reviews while another only has one review. So will you agree that the latter one is one of the most popular one? I guess your answer is no. That's why I consider the second feature: review count of a coffee shop. More reviews means more customers, and more popular. But can we confidently say that "the more reviews, the more popular"? It is a little bit tricky. What if a coffee shop has over 2000 reviews while more than half of the ratings are below 3? This seems to be a contradiction. Why would people choose a coffee shop if it has a lot of negative reviews? I think we are making an assumption here: "more positive reviews = more popular". Are we running into the sentimental analysis.... Let's talk about some other features. What about the average rating of all reviews for a coffee shop? Actually this approach has the same problems as those of the first feature. Another problem is it assumes that each review contributes the same weight to the final rating. How about considering the percentage of reviews above a threshold? It gives us more information about the distribution of review ratings for a coffee shop. We can also consider sorting coffee shops by neighborhood.

Here I propose a simple linear regression model:

$$popularity = weight\ of\ review\ count * review\ count + weight\ of\ Yelp\ rating * Yelp\ rating + weight\ of\ positive\ reviews * percentage\ of\ positive\ reviews$$

The choice of weightings for random variables are very heuristic. In the following parts, I will discuss some modification to this model.

2. Getting the data

OK, at least we need the following dataset:

- Basic info of all coffee shops in SF(name, address, review count)
- All the Yelp review rating values for each coffee shop
- Yelp business rating value for each coffee shop

Yelp provides open-source API for searching business by filters like location, keyword and category. But there are some restrictions:

- Maximum number of business records returned in each request is 20
- Maximum number of accessible business records is 1000
- Only one review associated with business is returned

So we can get up to 1000 coffee shop records. The returned result can also be sorted by best matched, distance and highest rated. If we use best-matched filter, there is a strong possibility that the top 50 most popular coffee shops will be included in the first 1000 records. Since only one review is returned for each business, we can't collect all the review rating values for a coffee shop. Yelp has an open dataset including businesses, reviews and users in 10 cities across 4 countries. Unfortunately, San Francisco is not in that list. One solution to this I think is to scrape the data off of their public site. But this violates the Yelp Terms of Service. OK, let's simplify our model by removing the last variable "percentage of reviews with rating above a threshold".

After preprocessing the data obtained by Yelp search API, it has the following format:

- name
- address
- rating
- review count

3. Computing the popularity

The modified model is:

$$popularity = weight\ of\ review\ count * normalized\ review\ count + weight\ of\ Yelp\ rating * normalized\ Yelp\ rating$$

The review count and Yelp rating are normalized by their maximums. The popularity is in the range of [0, 1].

Here are the top 50 most popular coffee shops with 0.5/0.5 weights. The center of the search area is Civic Center Plaza and the radius is 8000 meters (about 5 miles).

Name	Popularity	Address			
Blue Bottle Coffee	0.9414582253	315 Linden St	Hayes Valley	San Francisco	CA 94102
Blue Bottle Coffee Co	0.8999999762	66 Mint St	SoMa	San Francisco	CA 94103
Ritual Coffee Roasters	0.8954240084	1026 Valencia St	Mission	San Francisco	CA 94110
Philz Coffee	0.8935631514	3101 24th St	Mission	San Francisco	CA 94110
Dynamo Donut + Coffee	0.8835265636	2760 24th St	Mission	San Francisco	CA 94110
Four Barrel Coffee	0.8786455393	375 Valencia St	Mission	San Francisco	CA 94103
Philz Coffee	0.8496339321	748 Van Ness Ave	San Francisco	CA 94102	
Philz Coffee	0.82614398	201 Berry St	SoMa	San Francisco	CA 94158
Blue Bottle Coffee	0.7691274881	1 Ferry Bldg	Ste 7	Embarcadero	San Francisco
Pork Store Cafe	0.7453325391	1451 Haight St	The Haight	San Francisco	CA 94117
Sightglass Coffee	0.7358755469	270 7th St	SoMa	San Francisco	CA 94103
b. Patisserie	0.7190665007	2821 California St	Pacific Heights	San Francisco	CA 94115
Philz Coffee	0.7184563875	4023 18th St	Castro	San Francisco	CA 94114
Piccino	0.6998779774	1001 Minnesota St	Dogpatch	San Francisco	CA 94107
farm:table	0.6696766615	754 Post St	Lower Nob Hill	San Francisco	CA 94109
Golden Bear Trading Company	0.6623245478	1401 6th Ave	Inner Sunset	San Francisco	CA 94122

Stella Pastry & Cafe	0.6611348391	446 Columbus Ave	Russian Hill	San Francisco	CA 94133
Coffee Bar	0.6535082459	1890 Bryant St	Mission	San Francisco	CA 94110
The Plant Café Organic	0.6373398304	3352 Steiner St	Marina/Cow Hollow	San Francisco	CA 94123
Caffe Greco	0.6257474422	423 Columbus Ave	Russian Hill	San Francisco	CA 94133
Plentea	0.6227272749	341 Kearny St	Chinatown	San Francisco	CA 94108
Crossroads Cafe	0.6141549945	699 Delancey St	SoMa	San Francisco	CA 94107
Bean Bag Cafe	0.6077486277	601 Divisadero St	NoPa	San Francisco	CA 94117
Blue Barn Gourmet	0.6043929458	2105 Chestnut St	Marina/Cow Hollow	San Francisco	CA 94123
Fraîche	0.6028676033	1910 Fillmore St	Lower Pacific Heights	San Francisco	CA 94115
La Boulange de Hayes	0.6001220346	500 Hayes St	Hayes Valley	San Francisco	CA 94102
Arlequin Cafe & Food To Go	0.5873093605	384 Hayes St	Hayes Valley	San Francisco	CA 94102
Cafe du Soleil	0.5845637321	200 Fillmore St	Lower Haight	San Francisco	CA 94102
Blue Front Cafe	0.5766320825	1430 Haight St	The Haight	San Francisco	CA 94117
Velo Rouge Cafe	0.5662599206	798 Arguello Blvd	Inner Richmond	San Francisco	CA 94118
Blue Bottle Coffee Stand	0.5656498075	Ferry Plaza Farmers Market	1 Ferry Bldg	Embarcadero	San Francisco
Toy Boat Dessert Café	0.5650396347	401 Clement St	Inner Richmond	San Francisco	CA 94118
Wooly Pig Cafe	0.5622635484	205 Hugo St	Inner Sunset	San Francisco	CA 94122
Jane on Fillmore	0.5601586103	2123 Fillmore St	Pacific Heights	San Francisco	CA 94115
The Mill	0.5592434406	736 Divisadero St	Alamo Square	San Francisco	CA 94117
CoffeeShop	0.5527760983	3139 Mission St	Bernal Heights	San Francisco	CA 94110
Contraband Coffee Bar	0.5513117909	1415 Larkin St	Nob Hill	San Francisco	CA 94109
La Boulange de Cole	0.5500915051	1000 Cole St	Cole Valley	San Francisco	CA 94117

YakiniQ Cafe	0.5500915051	1640 Post St	2nd Fl	Japantown	San Francisco
Beanstalk Cafe	0.5500609875	724 Bush St	Nob Hill	San Francisco	CA 94108
Caffe Trieste	0.5464307666	609 Vallejo St	North Beach/Telegraph Hill	San Francisco	CA 94133
La Boulange de Fillmore	0.5446003675	2043 Fillmore St	Lower Pacific Heights	San Francisco	CA 94115
Mojo Bicycle Café	0.5446003675	639 Divisadero St	NoPa	San Francisco	CA 94117
Higher Grounds Coffee House	0.5387736559	691 Chenery St	Glen Park	San Francisco	CA 94131
The Revolution Cafe	0.5384991169	3248 22nd St	Mission	San Francisco	CA 94110
Philz Coffee	0.5363331437	399 Golden Gate Ave	Civic Center	San Francisco	CA 94102
Boba Guys	0.5360890627	3491 19th St	Mission	San Francisco	CA 94110
Bereka Coffee	0.5344722271	2320 Lombard St	Marina/Cow Hollow	San Francisco	CA 94123
Cafe Algiers	0.5320927501	50 Beale St	Financial District	San Francisco	CA 94105
Epicenter Cafe	0.5308724642	764 Harrison St	SoMa	San Francisco	CA 94107

For the complete list with 736 coffee shops: https://docs.google.com/spreadsheets/d/1H5_-6mGb3KYlnESodiEeMZwkdOJaEAYnbM99TXLnwTQ/edit?usp=sharing

4. Summary (to make it more like a homework)

In this project, I designed a popularity ranking model for the coffee shops in San Francisco. (Actually it can be applied to any business area.) The model is based on linear regression with two random variables: review count and rating. I also discussed the pros and cons for each possible ranking feature. I used Yelp search API to query for the coffee shops information. The top 50 most popular coffee shops from my ranking model are shown above. And there are still a lot more interesting things we can discuss.