CS 6220
Data Mining

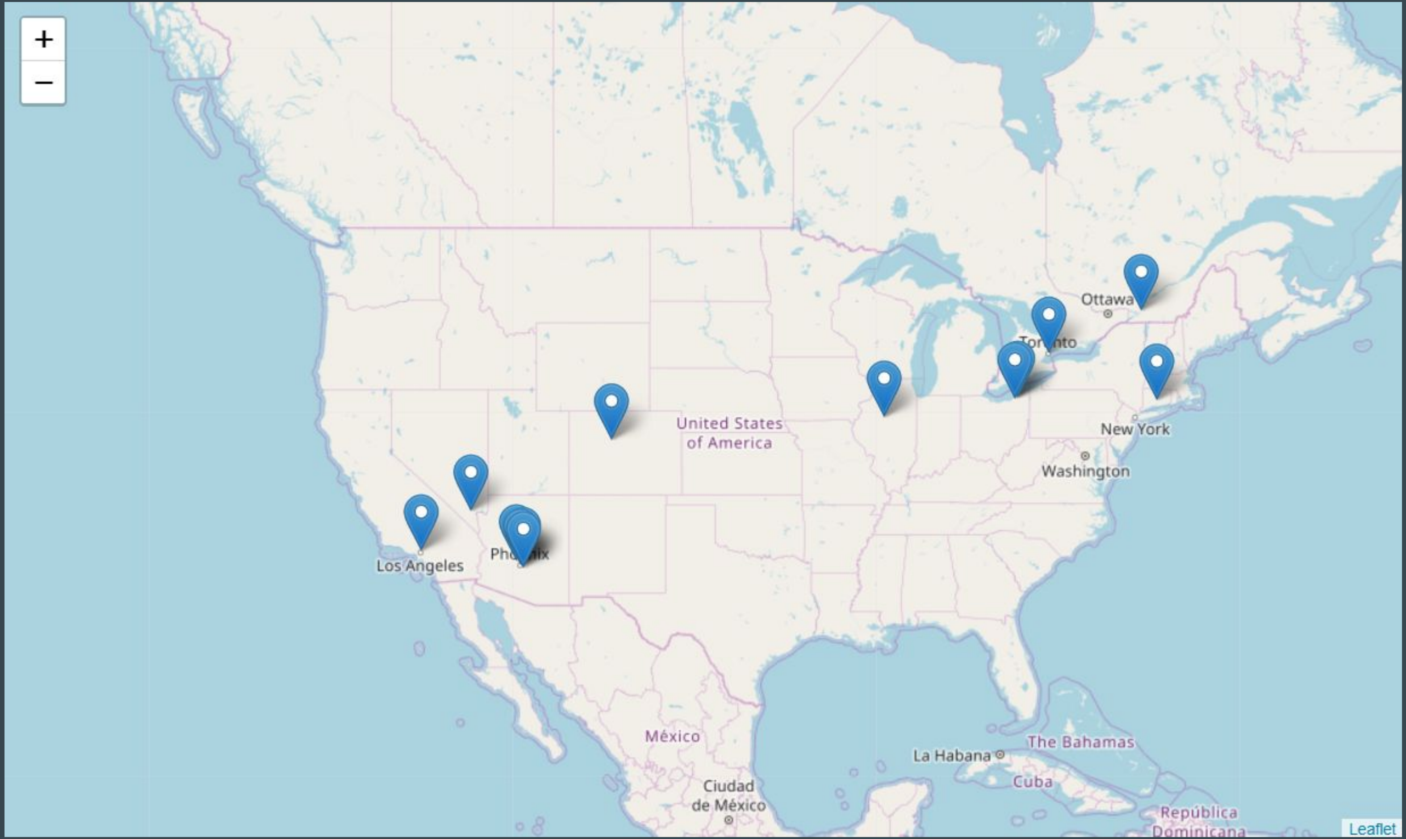
•••

Final Project
December 2017

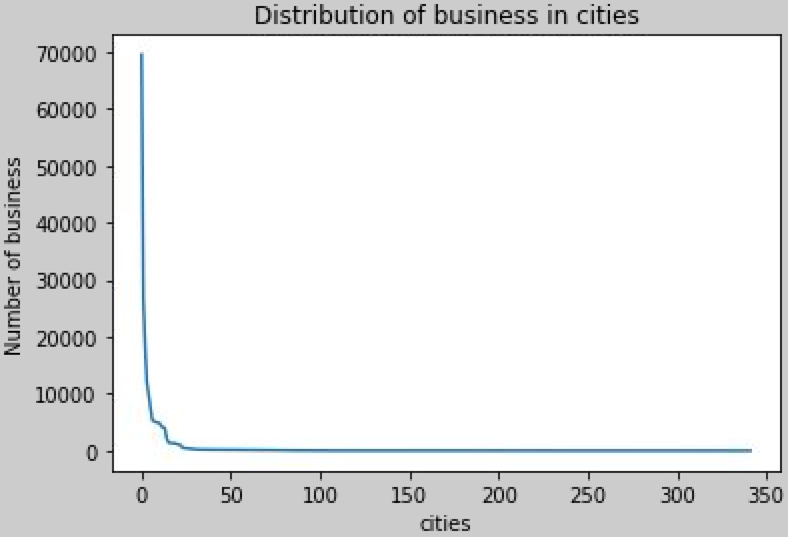
Anupam Sapre Raghav Sairam Praveen Singh Kuai Hu

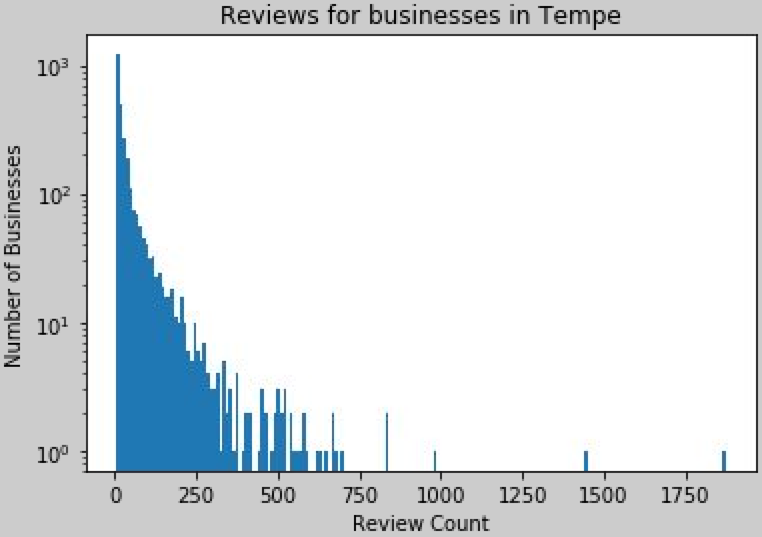
yelp
DATASET
CHALLENGE

Exploratory Analysis



Exploratory Analysis





Suggest additional services for businesses to increase customer attraction.

Each business in the yelp dataset has a list of categories associated to it that tells us about the types of services provided by the business.

We have tried to come up with association rules between the category of services, based on user reviews on businesses.

How it’s done ?

We have applied the FP-Growth algorithm to find the association between the categories.

The itemset provided to the algorithm to find the frequent item patterns and to find the rules has been prepared as follows:

1. Find all reviews above the given threshold stars (ex. 3).
2. For each review find the business.
3. Add the category list of the business to our list of itemsets .

With this list of categories as itemset we find the association rules between the categories and this can be used to suggest services to businesses to attract more customers.

Results

('Books', 'Bookstores', 'Music & Video', 'Shopping') ---> ('Mags')

('Beauty & Spas', 'Cosmetics & Beauty Supply', 'Eyelash Service', 'Hair Removal', 'Nail Salons', 'Waxing') ---> ('Day Spas', 'Skin Care')

('Fashion', "Men's Clothing", 'Shopping', 'Sporting Goods', 'Sports Wear', "Women's Clothing") ---> ('Accessories')

('Home Services', 'Local Services', 'Self Storage', 'Shopping') ---> ('Movers')

('Active Life', 'Arts & Entertainment', 'Dance Clubs', 'Dance Studios', 'Event Planning & Services', 'Fitness & Instruction', 'Nightlife', 'Party & Event Planning', 'Venues & Event Spaces') ---> ('Performing Arts',)

User Location Inference

Procedure:

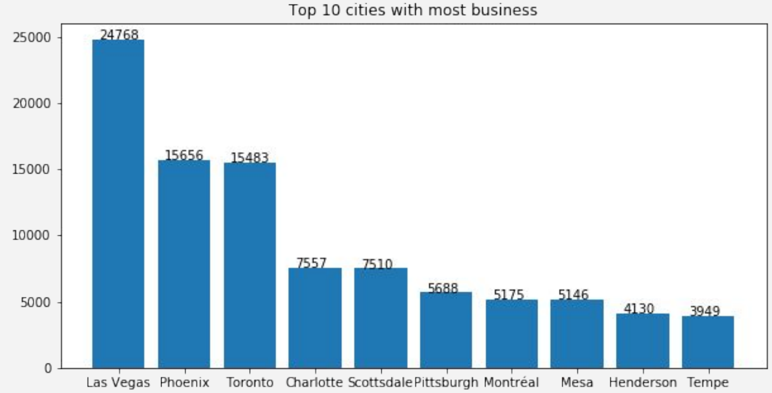
1. Pre-process, such as removing user with few reviews, removing business without city/geo-location, etc.
2. Connect users with business by their reviews respectively
3. Add business cities/geo-locations to users respectively
4. Represent user cities/geo-locations by most common city/geo-location in their consumption records.

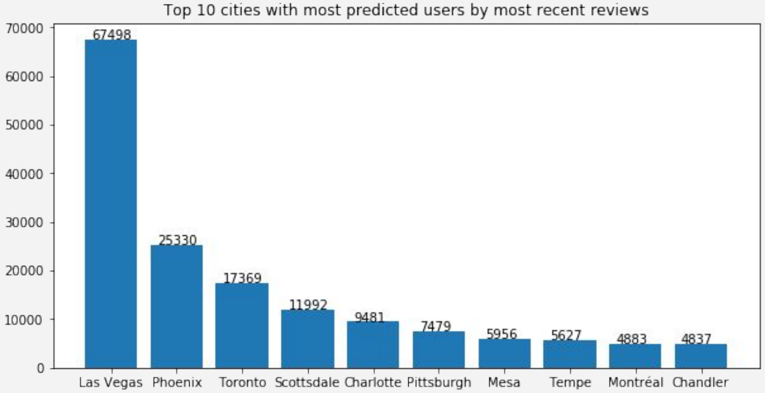
Optimization

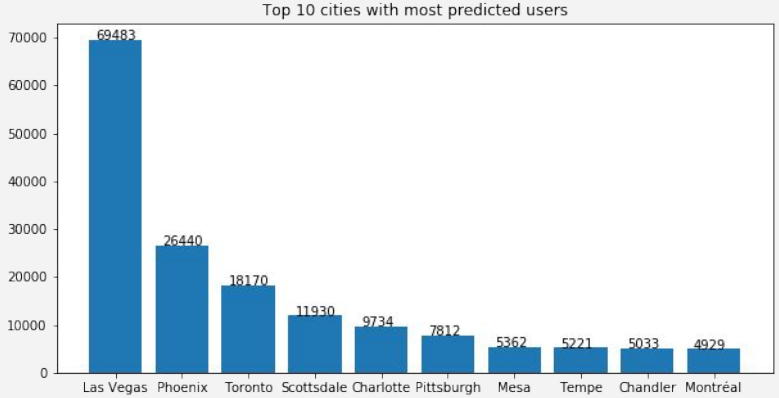
1. Extract most recent reviews
    a. avoid case that users possible relocation
    b. avoid case that users as travelers

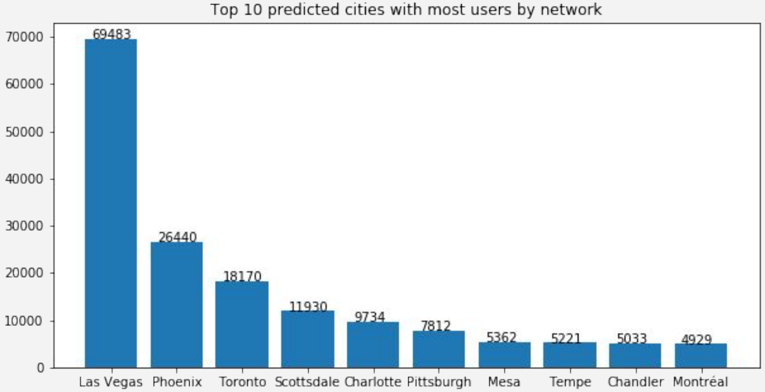
2. Optimize prediction based on users network
    a. evaluate prediction performance
    b. improve prediction performance

Results


Top 10 cities with most business


Top 10 cities with most predicted users by most recent reviews


Top 10 cities with most predicted users


Top 10 predicted cities with most users by network

Suggest Businesses based on user review patterns

Approach: Content Based Recommendation System

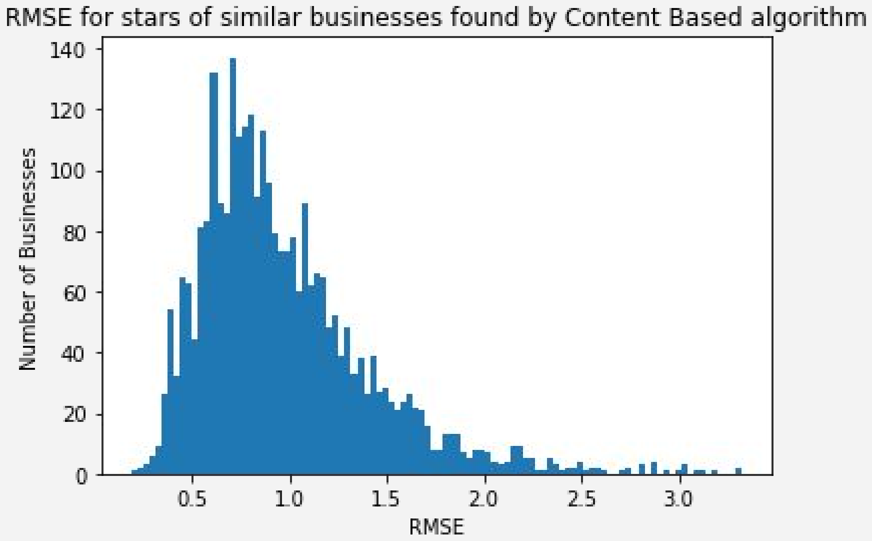
Procedure: 1) Find a business for a particular city with review counts within limits
2) Obtain reviews and users for these business
3) Use sklearn tfidf vectorizer to generate TF IDF matrix
4) Use SVD to reduce dimensionality of TF IDF matrix
5) Run cosine similarity to find distances between users and business.

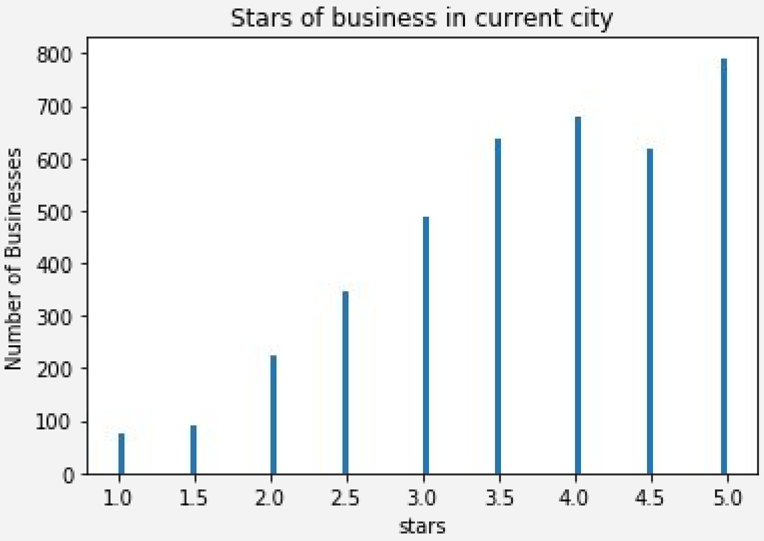
Suggested Businesses Example

Primary Business : Cobblestone Veterinary Care ['Veterinarians', 'Pets', 'Health & Medical', 'Acupuncture']

Similar Businesses: South Point Animal Clinic ['Pets', 'Veterinarians']
Pecan Grove Veterinary Hospital ['Pets', 'Veterinarians']
Topaz Veterinary Clinic ['Veterinarians', 'Pets']
Banfield Pet Hospital ['Pets', 'Veterinarians']
Tempe Veterinary Hospital ['Pets', 'Veterinarians', 'Pet Boarding/Pet Sitting', 'Pet Groomers', 'Pet Services']

Results


RMSE for stars of similar businesses found by Content Based algorithm
Number of Businesses
0
20
40
60
80
100
120
140
RMSE
0.5
1.0
1.5
2.0
2.5
3.0


Stars of business in current city
Number of Businesses
0
100
200
300
400
500
600
700
800
stars
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

Estimation of user ratings for unvisited businesses

Introduction:

This problem is aimed at serving a proof of concept for the implementation of a real time recommender system, should Yelp release a public API that allows querying of this data.

We currently use the Yelp academic dataset for analysis and performance of the algorithm applied.

Collaborative filtering

Collaborative filtering is a method of making automatic predictions about the interests of a user by collecting preferences or taste information from many users.

The underlying assumption for this method is that, people with similar opinions of particular topics would share this similarity for other topics. Such an assumption gives reason to implement collaborative filtering.

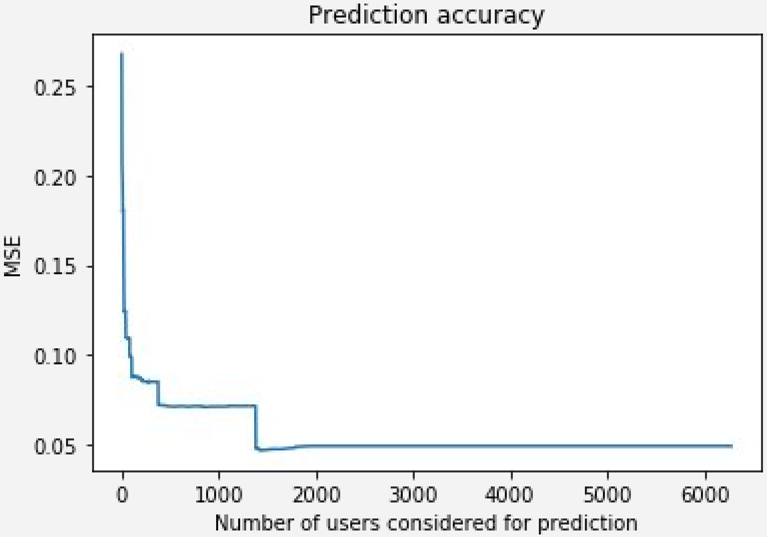
Implementation

Prediction for active user:

- Calculate similarity of active user with other users.
- The assumption behind this method suggests that users who are more similar to the current user are more useful for prediction.
- The similarity between users behaves as a metric which decides the influence of opinions by a random user on the active user.
- The prediction is a weighted mean of the most similar users (similarity weights).

prediction for $a_j = \Sigma\ (b_j * \mathrm{sim}(a,b)\ )\ /\ \ \Sigma\ \mathrm{sim}(a,b)$ [where $b_j \neq 0$]

Results



Debug run:

Suggests increase in accuracy with number of similar users considered.

Change in accuracy dwindles since the similarity between users would have little to no impact on prediction.

Future Work

- Collaborative filtering highly applicable to a variety of “user interest” problems.
- Develop methods to better tune hyperparameters (thresholds, learning rate etc).
- Use more reactive distance metrics to better suit different situations.