Project Report

Guide : Soham Saha
Nilesh Patil (201405550)
Pankaj Shipte (201405614)
Akshay Joshi (201406553)

Problem : *Restaurant Recommendation System*

Introduction :

There are many recommendation systems available for problems like shopping, online video entertainment, games etc. Restaurants & Dining is one area where there is a big opportunity to recommend dining options to users based on their preferences as well as historical data. Yelp is a very good source of such data with not only restaurant reviews, but also userlevel information on their preferred restaurants. This report describes the work to learn to predict whether a given yelp user visiting a restaurant will like it or not. I explore the use of different machine learning techniques and also engineer features that perform well on this classification.

Motivation :

Recommendation systems help users find and select items (e.g., books, movies, restaurants) from the huge number available on the web or in other electronic information sources. Given a large set of items and a description of the user's needs, they present to the user a small set of the items that are well suited to the description. Recent work in recommendation systems includes intelligent aides for filtering and choosing web sites, news stories and other information. The users of such systems often have diverse, conflicting needs. Differences in personal preferences, social and educational backgrounds, and private or professional interests are pervasive. As a result, it seems desirable to have *personalized* intelligent systems that process, filter, and display available information in a manner that suits each individual using them. The need for personalization has led to the development of systems that adapt themselves by changing their behavior based on the inferred characteristics of the user interacting with them. The ability of computers to converse with users in natural language would arguably increase their usefulness and flexibility even further. Research in practical dialogue systems, while still in its infancy, has matured tremendously in recent years.

Techniques :

Dataset :

The data that we are planning to use in this project is obtained from the Yelp Dataset challenge. The dataset contains five different tables: User, Business, Review, Check-In and Tips. The data has 14092 restaurants, 252395 users, 402707 tips and 1124955 reviews. The reviews span over 10 years of data.

Restaurant Objects

Restaurant objects contain basic information about local restaurant. The 'business_id' field can be used with the Yelp API to fetch even more information for visualizations, but note that you'll still need to comply with the API TOS. The fields are as follows:

```
{
   'type': 'business',
   'business_id': (a unique identifier for this
business),
   'name': (the full business name),
   'neighborhoods': (a list of neighborhood
names, might be empty),
   'full_address': (localized address),
   'city': (city),
   'state': (state),
   'latitude': (latitude),
   'longitude': (longitude),
   'stars': (star rating, rounded to
half-stars),
   'review_count': (review count),
   'photo_url': (photo url),
   'categories': [(localized category names)]
   'open': (is the business still open for
business?),
   'schools': (nearby universities),
   'url': (yelp url)
}
```

Review Objects

Review objects contain the review text, the star rating, and information on votes Yelp users have cast on the review. Use user_id to associate this review with others by the same user. Use business_id to associate this review with others of the same business.

```
{
   'type': 'review',
   'business_id': (the identifier of the
reviewed business),
   'user_id': (the identifier of the authoring
user),
   'stars': (star rating, integer 1-5),
   'text': (review text),
   'date': (date, formatted like '2011-04-19'),
   'votes': {
       'useful': (count of useful votes),
       'funny': (count of funny votes),
       'cool': (count of cool votes)
   }
}
```

User Objects

User objects contain aggregate information about a single user across all of Yelp (including businesses and reviews not in this dataset).

```
{
  'type': 'user',
  'user_id': (unique user identifier),
  'name': (first name, last initial, like
'Matt J.'),
  'review_count': (review count),
  'average_stars': (floating point average,
like 4.31),
  'votes': {
      'useful': (count of useful votes across
all reviews),
      'funny': (count of funny votes across
all reviews),
      'cool': (count of cool votes across all
reviews)
  }
}
```

Techniques :

We have used the place frequency range as a parameter and tested 8 binary-class classifiers (KNN, K*, Decision Table, Random Forest, SMO, SVM, Logistic Regression, Naive Bayes).

Features :

a) **User-level features :**

User level features are such as number of days in yelp, number of fans, number of votes etc.

b) **Restaurant-level features :**

Restaurant level features are such as binary features for attributes (parking, take out, etc) and categories (cuisine)

c) **User-Restaurant features :**

This set of attributes described the relations between a user and a restaurant.These features capturing similarity between a user and a restaurant. We will consider the factors like the location similarity, the background similarity.

d) **Derived features :**

1. **friends_avg_rating_business** : Average ratings given by user's friends for that business if user didn't have rated that business in training set. If none of his friends also haven't rated that restaurant, then average stars of that restaurant are assigned. If that business has not been rated by any user then average_stars of that user are assigned as default.
2. **avg_user_rating** : Average of all the ratings given by a user to all restaurants in dataset. If user haven't submitted any reviews to any restaurant then this value will be defaulted to average stars of corresponding business.
3. **avg_business_rating** : Average of all the ratings given by all users to particular restaurants in dataset. If restaurant haven't got any reviews then this value will be defaulted to average stars of corresponding user.
4. **Avg_category_rating** : Average ratings given by each user to particular category calculated over all categories for all users.
5. **Min_category_rating** : Minimum ratings given by each user to particular category calculated over all categories for all users.
6. **Max_category_rating** : Minimum ratings given by each user to particular category calculated over all categories for all users.
7. **Review_votes** : Total count of all likes in terms of Useful,funny and cool given for particular review.

- Support Vector Machine :

The SVM performs pattern recognition for two-class problems by determining the separating hyperplane with maximum distance to the closest points of the training set. These points are called support vectors. In its simplest linear form, an SVM is a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin.

- Logistic regression :

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems asprobit regression using similar techniques, with the latter using a cumulative normal distribution curve instead. Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors.

Results :

Accuracy : 77.64 %

Conclusion :

1. We have performed various experiments on Yelp dataset to improve accuracy of this recommendation system. We found SVM performs better than other classifiers in terms of accuracy.
2. We see a significant improvement from derived features, specifically from using the following:

    friends_avg_rating_business
    Avg_user_rating
    Avg_business_rating

3. The main boolean features which had impact on accuracy were

    business_attributes.Parking.garage,
    business_categories_Buffets,
    business_categories_FastFood.

References :

[1] Yelp dataset: https://www.yelp.com/academic_dataset

[2] Restaurant Recommendation System By Ashish Gandhe