YELP COMMENT ANALYSIS AND RECOMMENDATION

Li Qi
lq2156
Project Overview

- Yelp is a mobile application which publishes crowd-sourced reviews about local business, as well as the online reservation service and online food-delivery service.
- People mainly use Yelp to choose nearby restaurant.
- Problem: Overwhelming Recommendations
- Main idea: Reduce the recommendation list.
- Using filter to present the top 5 most relevant results.
- Related course: Text-mining skills introduced by Professor Oded Netzer
- Future Challenge: Combining Twitter data to get a complete view of certain restaurant.
<table>
<thead>
<tr>
<th>Rank</th>
<th>Location</th>
<th>Type</th>
<th>Distance</th>
<th>Reviews</th>
<th>Price</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Koko Wings</td>
<td>Chicken Wings, Korean</td>
<td>0.6 mi</td>
<td>140</td>
<td>$$</td>
<td>248 W 106th St, Manhattan Valley</td>
</tr>
<tr>
<td>2</td>
<td>Dinosaur Bar-B-Que</td>
<td>Barbeque, Music Venues</td>
<td>0.6 mi</td>
<td>2165</td>
<td>$$</td>
<td>700 W 125th St, Harlem</td>
</tr>
<tr>
<td>3</td>
<td>Atomic Wings</td>
<td>Chicken Wings</td>
<td>0.5 mi</td>
<td>82</td>
<td>$$</td>
<td>2090 Frederick Douglass Blvd, Harlem</td>
</tr>
<tr>
<td>4</td>
<td>Mel's Burger Bar</td>
<td>Burgers</td>
<td>0.4 mi</td>
<td>505</td>
<td>$$</td>
<td>2850 Broadway, Morningside Heights</td>
</tr>
<tr>
<td>5</td>
<td>Amy Ruth's</td>
<td>Southern, Soul Food</td>
<td>0.7 mi</td>
<td>1634</td>
<td>$$</td>
<td>113 W 116th St, Harlem</td>
</tr>
<tr>
<td>6</td>
<td>LoLo's Seafood Shack</td>
<td>Seafood, Sandwiches</td>
<td>0.4 mi</td>
<td>215</td>
<td>$$</td>
<td>303 W 116th St, Harlem</td>
</tr>
</tbody>
</table>

**Information Overwhelming!**
Evaluation Metrics

- Here I choose *root mean square error (RMSE)* and *mean absolute error (MAE)* as metrics through K-fold cross validation.
- Reason: As the Microsoft Research paper* on evaluating recommendation system argues, both RMSE and MAE are highly-accurate metrics.

*Gunawardana A., Shani G, Evaluating Recommendation Systems*
Flowchart

- Yelp/Twitter Data (JSON)
- Dataset (CSV)
- Latent-factor Recommendation system
- Text Processing
- Word count (MapReduce)
Dataset

- Yelp data can be downloaded from Yelp Dataset Challenge
- There are five files contained inside the folder and we choose:
  - yelp_academic_dataset_bussiness.json
  - yelp_academic_dataset_checkin.json
  - yelp_academic_dataset_review.json
- We concentrate on “Bussiness ID”, “name”, “City”, “State”, “Longitude”, “Latitude” features in bussiness.json.
- “User ID” “Star” “text” in review.json
"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "PUFPaY9KxDAcGqfsorJp3Q", "review_id": "Ya85v4eQdd6k90G8H8BqyA", "stars": 4, "date": "2012-08-01", "text": "Mr. Hoagie is an institution. Walking in, it does seem like a throwback to 30 years ago, old-fashioned menu board, booths out of the 70s, and a large selection of food. Their specialty is the Italian Hoagie, and it is voted the best in the area year after year. I usually order the burger, while the patties are obviously cooked from frozen, all of the other ingredients are very fresh. Overall, it's a good alternative to Subway, which is down the road."}, "type": "review", "business_id": "5UmKjUEUNDywQAnHgckJw"}
"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "Tu6AxDbYGR4A0wspR9BYHA", "review_id": "KPv5J21_4wbYNctr0WwOQ", "stars": 5, "date": "2014-02-13", "text": "Excellent food. Superb customer service. I miss the marlin machines they used to have, but it's still a great place steeped in tradition."}, "type": "review", "business_id": "5UmKjUEUNDywQAnHgckJw"}
"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "auESfwWwW2h6a1XgFxAQQ", "review_id": "fF5oGd46Yxwuwr3fHNuZig", "stars": 5, "date": "2015-10-31", "text": "Yes this place is a little out dated and not opened on the weekend. But other than that the staff is always pleasant and fast to make your order. Which is always spot on fresh veggies on their hoggies and other food. They also have daily specials and ice cream which is really good. I had a banana split they piled the toppings on. They win pennysaver awards ever years i see why."}, "type": "review", "business_id": "5UmKjUEUNDywQAnHgckJw"}
"votes": {"funny": 0, "useful": 0, "cool": 0}, "user_id": "uK8tzra0p4M5u3uYrqIBXg", "review_id": "Di3exaUCfnv1V4kSNW5pqA", "stars": 5, "date": "2013-11-08", "text": "All the food is great here. But the best thing they have is their wings. Their wings are simply fantastic!! The "Wet Cajun" are the best & most popular. I also like the seasoned salt wings. Wing Night is Monday & Wednesday night, $0.75 whole wings! 

The dining area is nice. Very family friendly! The bar is very nice as well. This place is truly a Yinzer's dream!! "Pittsburgh Dad" would love this place n'at!!"}, "type": "review", "business_id": "UsFtgoBl7naz8AVUBZMjQQ"}
Algorithm

- Cascaded Clustered multi-step weighted bipartite graph projection algorithm: 1. use K-means clustering algorithm to partition the users. 2. construct a compressed graph. 3. run the multi-step random walk based algorithm on the compressed map to make prediction.

- A combination of Clustered Weighted Bipartite graph projection algorithm and multi-step random walk weighted bipartite graph projection algorithm

- Have good performance, have the least score of MAE and RMSE, compared with conventional KNN clustering algorithms.
## Compared Result

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Baseline</td>
<td>2.05788</td>
<td>1.27005</td>
</tr>
<tr>
<td>Singular Value Decomposition</td>
<td>1.56572</td>
<td>1.03162</td>
</tr>
<tr>
<td>Hybrid Cascade of KNN Clustering</td>
<td>1.52754</td>
<td>0.97114</td>
</tr>
<tr>
<td>Weighted Bi-Partite Graph Projection</td>
<td>1.48647</td>
<td>0.95232</td>
</tr>
<tr>
<td>Clustered Weighted Bi-Partite Graph Projection</td>
<td>1.14625</td>
<td>0.71332</td>
</tr>
<tr>
<td>Multi-Step Random Walk Weighted BiPartite Graph Projection</td>
<td>1.23862</td>
<td>0.79792</td>
</tr>
<tr>
<td>Cascaded Clustered Multi-Step Weighted BiPartite Graph Projection</td>
<td>1.09263</td>
<td>0.67548</td>
</tr>
</tbody>
</table>

Sumedh Sawant, Gina Pai, Yelp Food Recommendation System.

Latent-factor Recommendation System

- Latent-factor recommendation model trains a model capable of predicting a score for each possible combination of users and items. The internal coefficients of the model are learnt from known rating of users and items.
- List out top 5 highest results for consumers.
- “Cold-start” problem: for new users, let them doing a list of questions concerning their taste, their favorite food style and so on to do the recommendation.
- This system achieved RMSE of rating prediction as 0.9422*.

*Bee-Chung Chen, *Latent Factor Models for Web Recommender Systems*
http://www.ideal.ece.utexas.edu/seminar/LatentFactorModels.pdf
Algorithm

- Here we use LDA for topic modeling, MapReduce for word count and use WordNet, a large lexical database of English developed by Princeton University, to expand keyword matching.
- LDA is used as a topic model to discover the underlying topics from text documents.
- Sentiment Analysis: Combining the key words in the review with ratings, we can get a fully understanding of the user’s attitude towards the certain restaurant.
This consumer gave 5 stars to the restaurant and mentioned best, nice and great several times (Word count). Also mentioned wings with fantastic.

From the perspective of restaurant, wings is popular and will be recommended to other consumers who like wings. From the perspective of consumer, since wings is a traditional American dish, the system will recommend other high rating American dishes to the consumer.
Future Challenge

- Adding Twitter dataset to get more information about certain restaurant
- Combining Twitter data with Yelp
- Key Technique:
  - Crawling Twitter Data: Twitter API (REST API & Streaming API)
  - Storing Twitter Data: MongoDB (NoSQL)
  - Data Analysis: Finding topic & Sentiment analysis
Thank you!