Title: Recommending Stores to Shopping Malls

Team:
John Wu
Ying Zhang
Purnima Padmanabhan

Abstract

- Recommend stores to shopping malls from demographic records and yelp records
- Run Hadoop technologies to filter data and to evaluated if a store is appropriate for a mall with Logistic Regression
- Recommended stores to shopping centers with various recommendation techniques

May 5, 2015

Background

- Train a logistic regression classifier to predict whether a store i is in mall j.
- $\forall s \in S$
  - X: Demographic Information of the Malls
  - $y_j = \begin{cases} 1 \text{ if } s \in m_j \\ 0 \text{ otherwise} \end{cases}$
  - Regression on y_j from X for S
  - Recommend s to m_j if $y_j = 1$

Background

- A mall having a store can be modeled as binary matrix where the $X_{ij} = 1$ if the *jth* mall has the it store
- 0s can represent if a mall greatly dislikes a store or has not yet discovered the stores.
  - Positive examples are 1s
  - Traditional recommendation techniques do not necessarily apply to these problems
- Implemented Algorithms
  - Item-Based Collaborative Filtering
  - User-Based Collaborative Filtering
  - Top-K items
  - Content Based Recommendations
  - One Class Collaborative Filtering
  - Ensembler - Linear Regression

Motivation

Who are the users of the analytic?

- Mall owners
- Customers who would like to discover new stores
- Search Engine Services such as yelp

Who will benefit from this analytic?

- Businessmen seeking to maximize profit
- Shoppers in areas who are attracted to by different shops

Why is this analytic important?

- Use a statistical and mathematical approach to discover new stores for users that are important to them

Data Sources

Name: Mall Data
Description: Web-scraped basic information of shopping malls from MallsInfo.com

- Location (Latitude and Longitudinal Coordinates)
- Name of Stores that each mall has
- Count of Store Categories

Name: Demographic Information of Malls (census.org/bls.gov)
Description:

- 29 Datasets that describe Demographic Information

Name: Industry Information (bls.gov)
Description:

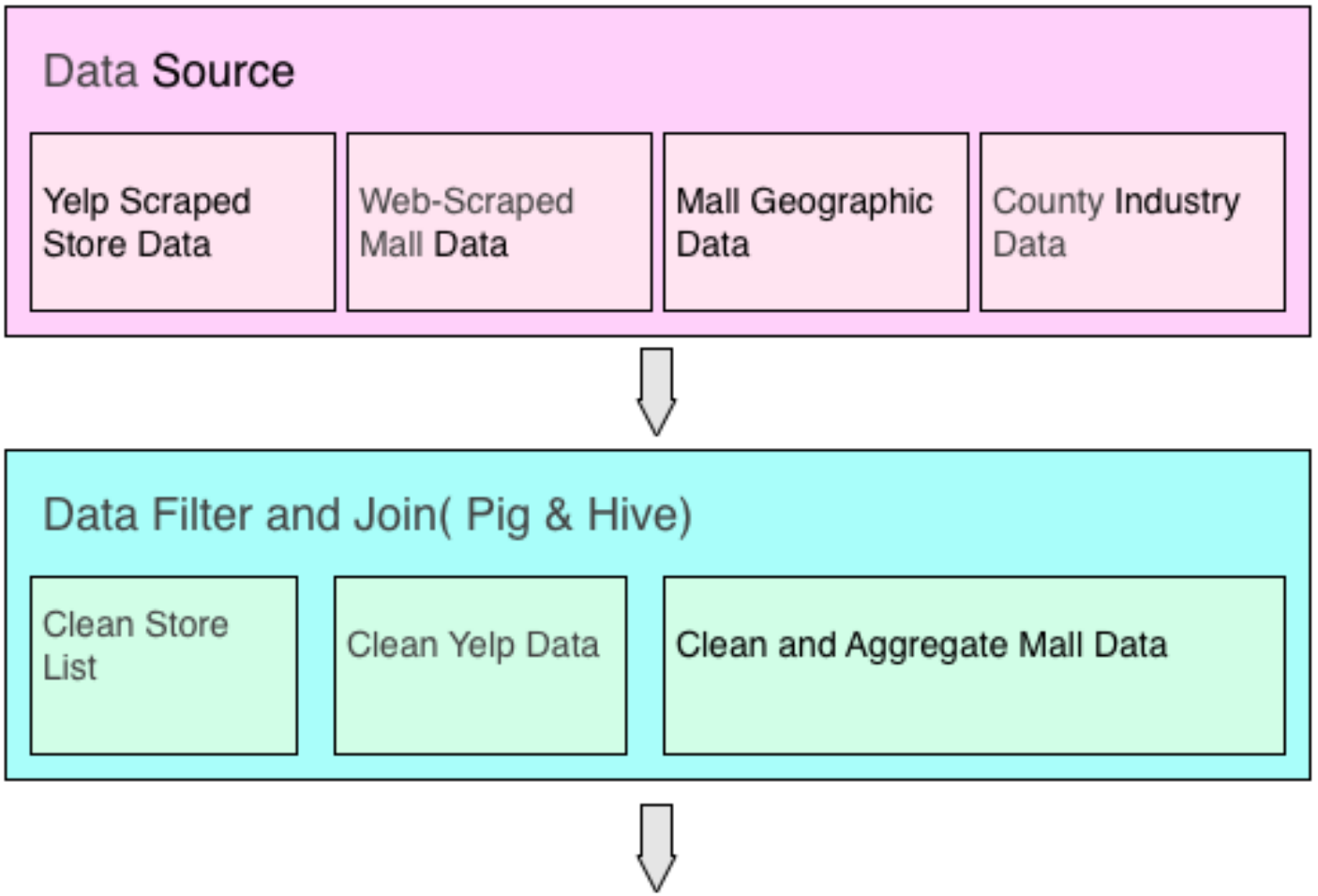
- Average Biweekly Earnings of Industries

Data Sources

Name: Store Information (Reviews)
Description: Web-scraped store information (yelp.com)

- Average rating of stores (out of 5 stars)
- Number of Ratings
- Indicator of expensiveness (out of 4 dollar signs)
- Category of store

Design Diagram



Design Diagram

A Data Filtering Problems

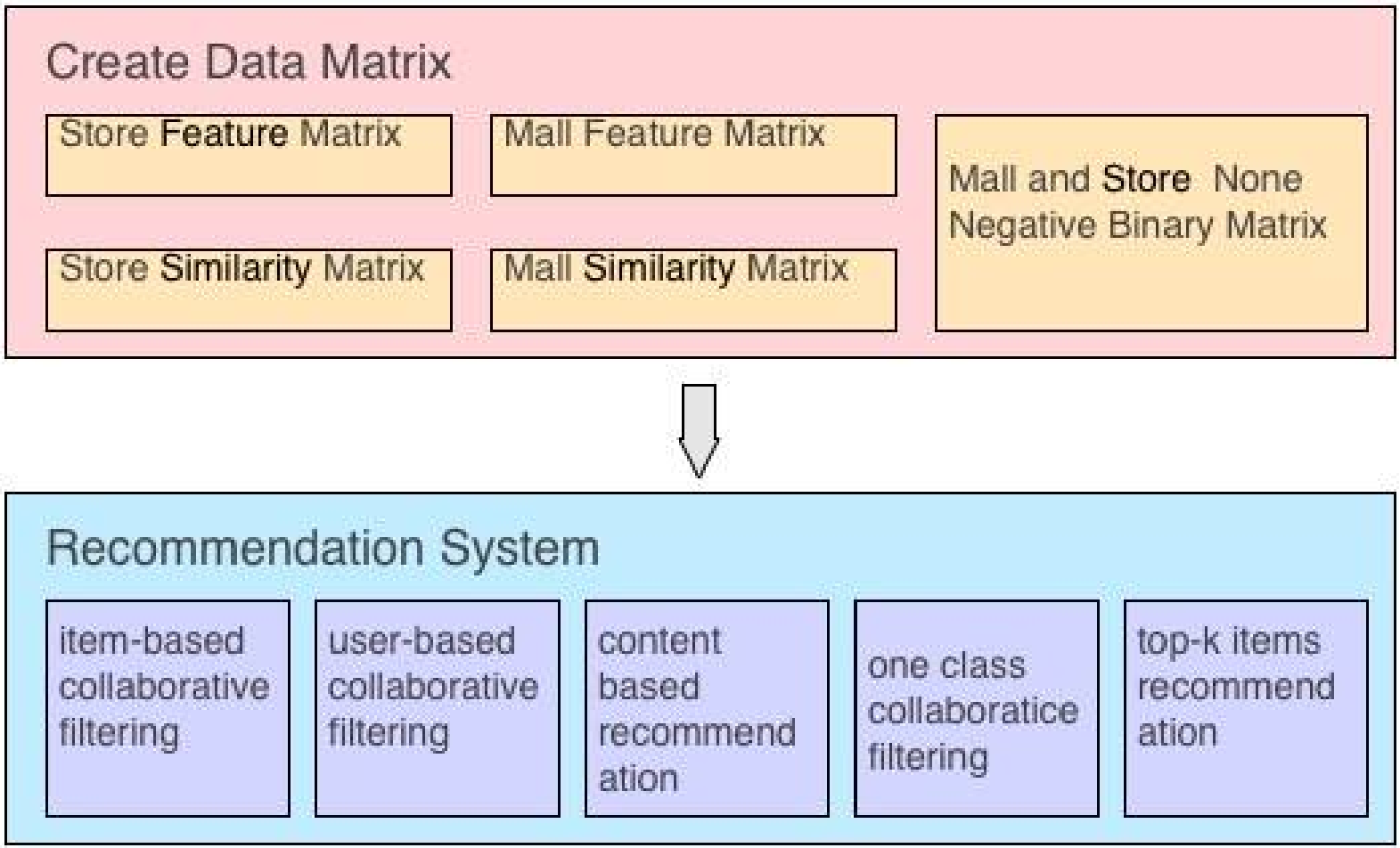
Problem

- Stores listed in the mall dataset are not store. These items have to be deleted
- Most stores have different names that need to be aligned the name the same
- Lots of stores have special characters and misspellings
- Yelp Crawler have some empty columns that need to be removed

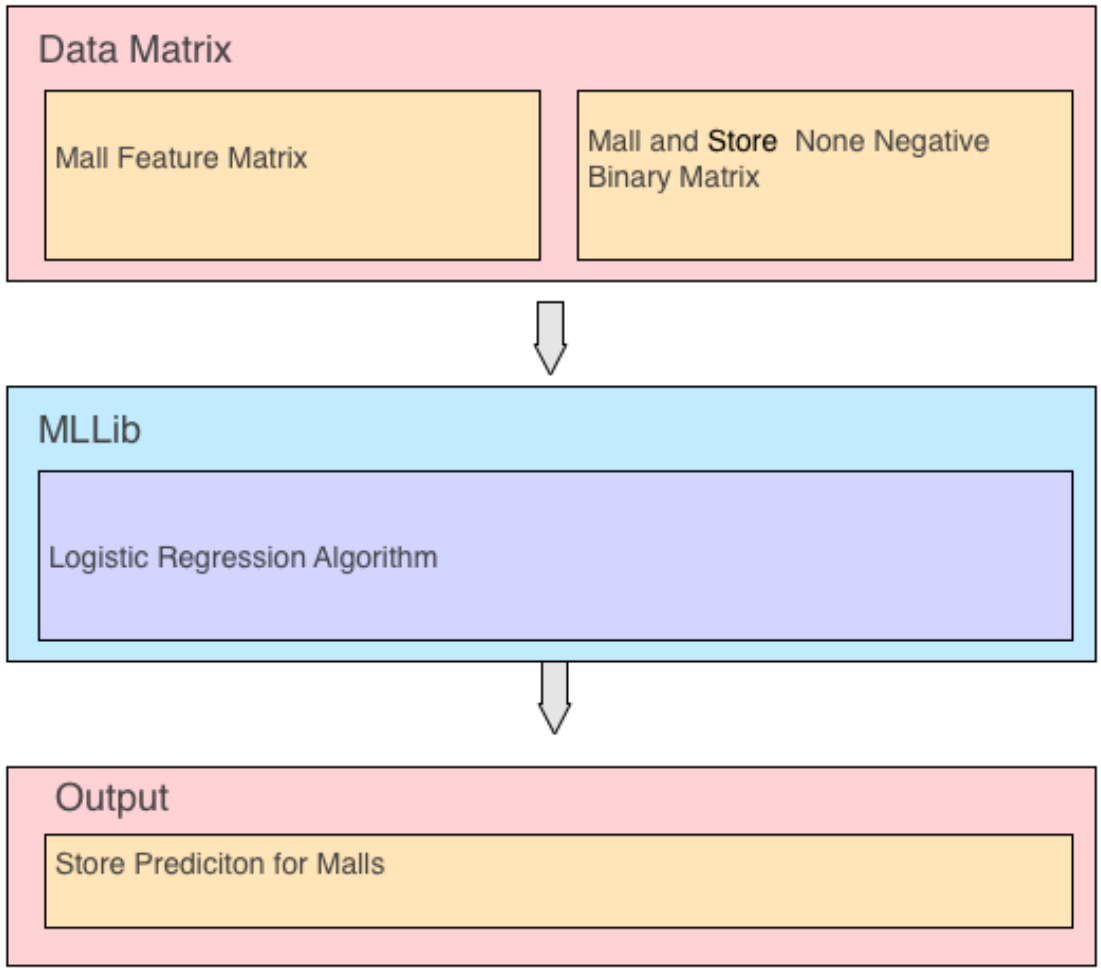
Solution

- Created three different regular expression rules to filter our data and create unique id for every store
- Write Pig script to load data first, then use a pig user-define function to filter these data
- Write Hive script to remove empty columns

Design Diagram



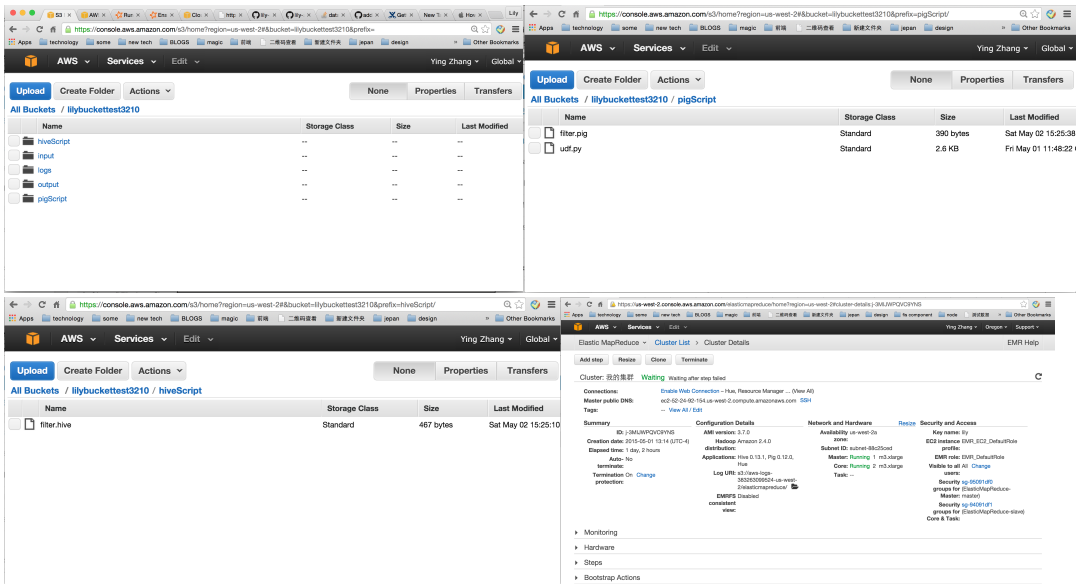
Design Diagram



Recommending Stores to Shopping Malls

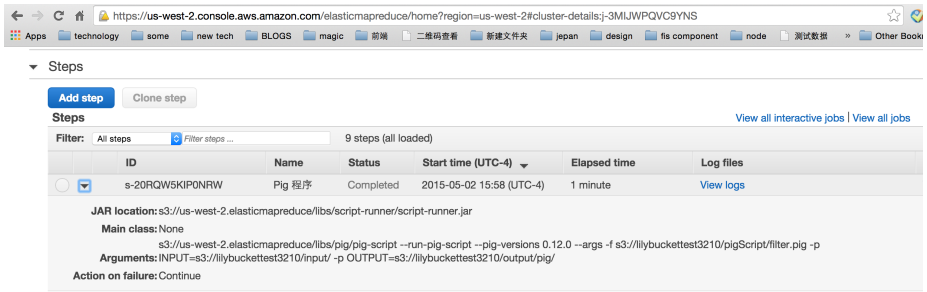
Platform(s) on which the analytic ran

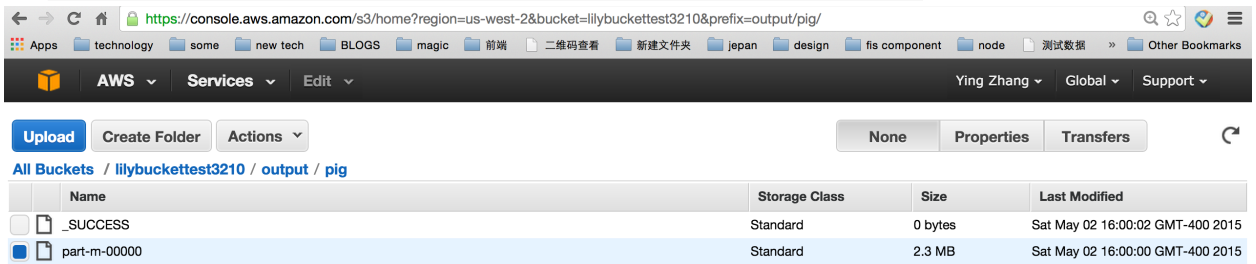
- MLlib was run on Cloudera
- Pig and Hive Scripts were ran on Amazon AWS

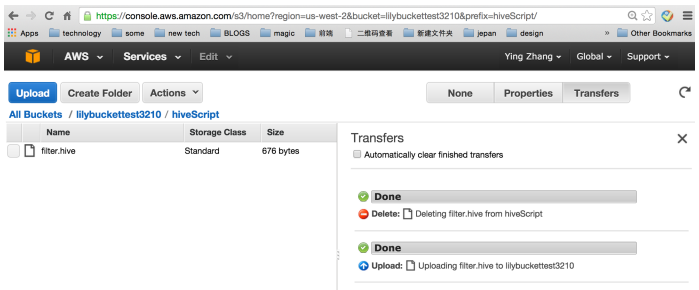


Results

1. Pig and Hive Scripts Completed their Tasks







Results

2. Logistic Regression

100 Iterations Logistic Regression Total Error: 0.4395	
Stores	RMSE
Best Buy	0.7139
Daphnes Greek Cafe	0.0103
Bed Bath and Beyond	0.1030
Nordstrom	.8891
Att	0.4201
Panera	0.1391
Ann Taylor	0.8247
Chicos	0.2242
Gap	0.6314

Results

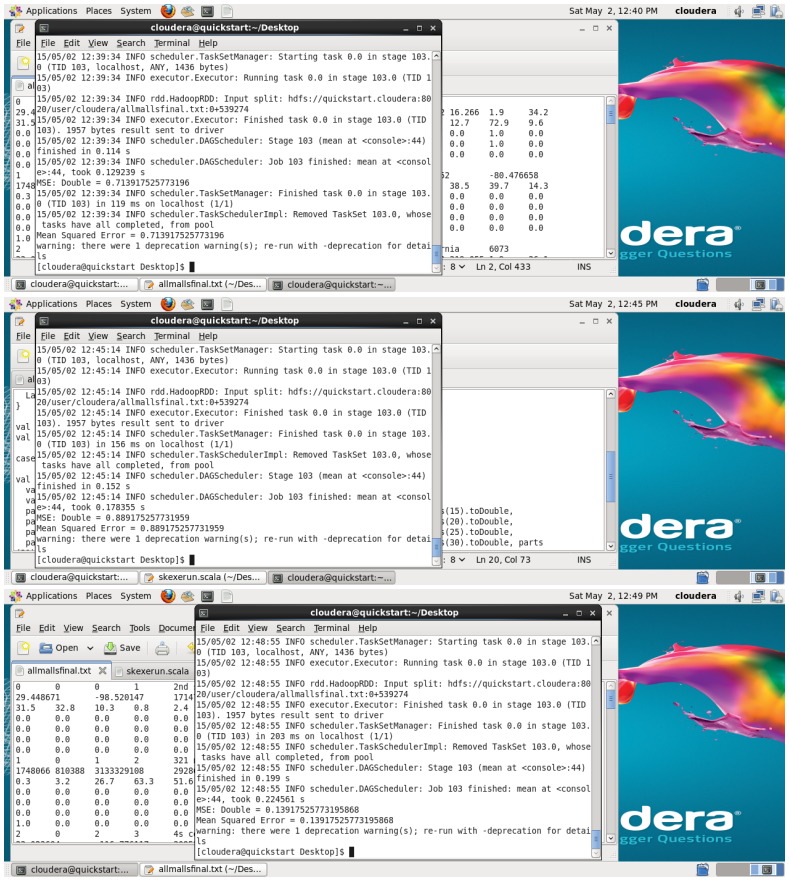
2. Logistic Regression

10 Iterations Logistic Regression	
Stores	RMSE
Gap	0.3685
Ann Taylor	0.1752

When number of iterations was reduced to 10 from the default value of 100, the result of predictions/categorization showed improvement for some malls. This may be due to overfitting at the value of 100.

Recommending Stores to Shopping Malls

Results 2.



Results 3.

id	Algorithms	MAP	RMSE
0	PopRec	0.389	0.389
1	UBCF with NMF-Category	0.447	0.072
2	UBCF with Demographic Percentage	0.379	0.0765
3	UBCF with Geographic Locations	0.291	0.079
4	CBR with Demographic Averages	0.0338	0.94
5	CBR with Geographic Location Averages	0.006	3261
6	WLAS with User Weights	0.378	0.083
7	Ensembler: 0+1+2+3	0.412	0.071
8	IBF with Yelp Data	TBA	TBA

MAP: Mean Average Precision, **RMSE**: Root Mean Square Error
UBCF: User Based Collaborative Filter, **CBR**: Content Based Recommendations, **WLAS**: Weighted Least Alternating Squares, **IBCF**: Item Based Collaborative Filter

Obstacles

1. Difficult to run MLlib on a cluster and on
2. Better Ensemble Method (Gradient Boosting) could have been implemented

Conclusion

Acknowledgments

- Roy Lowrance, Dennis Shasha,Joe Jean
- Suzanne McIntosh
- NYU HPC Team
- Amazon Web Services
- NYU HPC

Conclusion

- Implemented Logistic Regression to determine if a store would be a good fit for a mall
- Created Recommendation Systems to recommend stores to malls

Acknowledgements

- Roy Lowrance, Joe Jean Dennis Shsha
- Suzanne McIntosh
- Amazon Web Services
- NYU HPC

References


Bureau of Labor and Statistics (bls.gov)(2013)

Quarterly Census of Employment and Wages 2013.


J. Jean, R. Lowrance, and D. Shasha, (http://ml2014.herokuapp.com/) (2014)

Store-Mall Recommendation


A. Rajaraman and J. D. Ullman (2011)

Mining of Massive Datasets.


R. Pan, Y. Zhou, B. Cao, N. N. Liu, R. Lukose, M. Scholz, and Q. Yang (2008)

One-Class Collaborative Filtering.


Yelp (yelp.com) (2015)

Yelp Data

Thank You