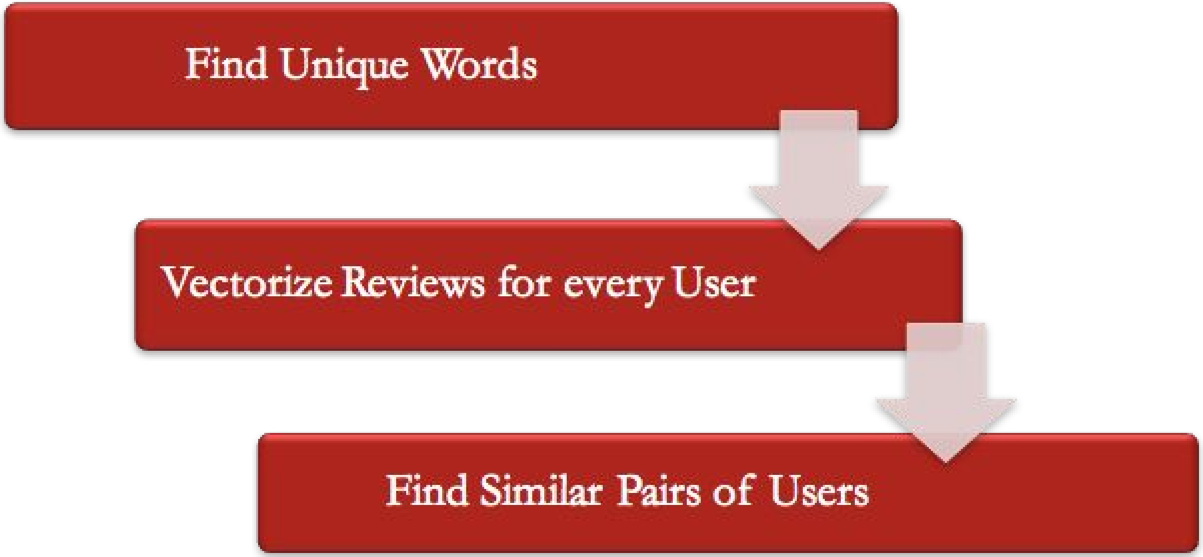
Restaurant Recommendation

CMSC 12300
Y Cube

Dataset	Size	Variables
Yelp _user	1M users	**User_id**: “encrypted user id” **Name**: “first name”
Yelp_review	2.93GB	**Review_id**: “encrypted review id” **User_id**: “encrypted user id” **Business_id**: “encrypted business id” **Text**: “review text”
Yelp_ business	144,070 Restaurants	**Business_id**: “encrypted business id” **Name**: “business name” **Address:** “full address” **Stars**: star rating **City**: “city”

Algorithm



Find Unique Words

Vectorize Reviews for every User

Find Similar Pairs of Users

Find Unique Words

- MRJob
- Key as none, Value is a list of unique words
- Create a LARGE vector, with unique words as elements
- Export to csv file

Vectorize Reviews For Every User

- MRJob
- Key is user_ID; Value is a vector that maps to the LARGE vector
- Word frequency += 1

Find Similar Pair of Users

- MRjob
- Input: pairwise user_ID
- Mapper: Key: user 1; Value: a list of 3 elements
    - User 2, user 1 vector, user 2 vector
- Combiner: Key: user 1; Value: a list of 2 elements
    - User 2, cosine similarity of User 1 & 2
- Reducer: Key: user 1; Value: a list of 2 elements
    - The most similar pair of user 1: user X, cosine similarity of 1&X

Results

- Small test file
  - 500 out of 1M users

- "**-1Eu-fym0JHDzU8dVYPUuw"** ["**LjDSVQGLLiOO7NCfvmV_MQ**", **0.4233944627**]
  "-7UURB-qhCeST2DGjjRyeQ" ["RBZ_kMjowV0t6_nv2UKaDQ", 0.4529234573]

- Link Back to visited_restaurant database
  - **-1Eu-fym0JHDzU8dVYPUuw**| **McDonalds**
  - **LjDSVQGLLiOO7NCfvmV_MQ**| **Cana Latin Kitchen & Bar**

Challenges

- SUPER Long run time
    - 500 users > 9 hours
    - What about 1M users?