Data Processing

1. Data Collect
    1. Mall data: use web crawler to get data from a mall website
    2. Stroe data: use webcraler to get data from yelp.com
    3. geographic data: find the data from: census.org /bls.gov
    4. industry data: find the data from: census.org / bls.gov

Data Filter:

1. Store Data Filter

Problem:

1. Lots of Stores are not store, for example: atm, vending machines , community service etc. we need to delete these items.
2. Lots of same Stores have same names, for example: Starbucks and Starbucks Coffee, we need to keep the name the same
3. Lots of stores have special characters and misspelling, we need to correct these.

Solve:
In order to get the best predict result, we create three different regular expression rules to filter our data and create unique id for every store:

1. Detail rules:(https://github.com/lily-zhangying/find_best_mall/tree/master/filter_store_data/filter_rules)
2. Write Pig Script to load data first, then use a pig use-define function to filter these data.

2. Yelp Store Data
After we get the clean store list, we use a yelp crawler to get all there store data.
Data Format:
    1. store name
    2. category
    3. level of expense
    4. popularity rating

3. Industry Data Filter

Problem:

1. Every County has different industry category, (the category is influenced by the owner, industry and time), some type of industry is very rare, just existing n some special places. These items will influence the accuracy of the prediction

Solve:
Find the intersection of all counties industry, and just keep the common industry in the data.

Data Join:

1. After filter all mall data, we need to join all geographic data , industry data and mall data together, so that in machine earning part ,we can create mall fearture matrix from these data.

2. Join Store list with yelp data as the store data.

Machine learning:

1. Create Matrix:
    1. Because we are using several different recommendation system algorithm, we will use our data to create mall feature matrix, store feature matrix and mall store binary non negative matrix.
    2. Then we use some machine learning method to generate mall similarity matrix from mall feature matrix, and generate store similarity matrix from store feature matrix.

2. using different machine learning algorithm