**Business Analysis**

Name: Lintong Yang

NUID: 001830082

*Abstract*

As it’s the era of market economy today, when launching a new product or proving a new service, it’s important to get feedback from customers. Even though new products are sold well and occupy a large market share at last, the reviews from customers provide instructions and destinations for business to make progress. What’s more, user order records and ratings are precious data sources for some service provider platform like UberEat, foodeasygo or yelp. As this reason, I choose data set from yelp, and do data analysis on three table of them, they are business, user and review json file.

In this project, I use the following technologies: Hadoop, Hbase, Pig, and Hive. As general, I use Hadoop to store output or input file, use Hbase to store datas , use Pig to do some calculation, and use Hive to analysis positive or negative sentiments of reviews from users. Next, I will describe them individually in detail.

*About Database*

The database includes 6 json files, which is downloaded from yelp database open source website. In this research, I mainly use 3 of them, they’re business, review and user. The size is about 3.1GB.

*Hbase*

I create three tables in hbase and writes data into it by java, they are “businesstable”, ”reviewtable” and “ usertable ”. And in each table, I set several families in order to cluster columns and attributes.

|  |  |
| --- | --- |
| Business Table | |
| Families | Colomn |
| Row Key | Business id |
| binfo | Address |
| State |
| Latitude |
| Longitude |
| name |
| Neighborhood |
| City |
| Ratings | Review\_count |
| Stars |
| Categories | Categories |
| Attributes | Attribute1 |
| Attribute2 |
| … |

|  |  |
| --- | --- |
| User Table | |
| Families | Columns |
| Row Key | User id |
| Votes | Funny |
| Useful |
| Cool |
| Uinfo | Reviewcount |
| Name |
| Fans |
| Average Stars |
| Yelp Since |
| Compliments | Compliment1 |
| Compliment2 |
| … |

|  |  |
| --- | --- |
| Review Table | |
| Row Key |  |
| rinfo | User id |
| Text |
| Business id |
| Date |
| Stars |
| Votes | Cool |
| Funny |
| Useful |

*Pig*

Analysis

1. Business

After process of data storage, begin with data analysis. From business, in different cities and states, as general, what we want to learn is which businesses are popular in local. This can be concluded from customer rating and rating times. After parallel comparison, vertical comparison is also important. In this research, business in a state that rated above average rating are selected, and overall data is group by state. This could show economic level of different state in US to some extent.

Based on consideration above, I try to figure out following questions:

* Find top business in a city with respect to rating
* Find business number in a state that rating above the average
* The number of business open from Monday to Sunday
* The “hottest business”: businesses that are reviewed for mostly

1. User

The users who are active on platform are the ones who are strong potential customers who will pay the bill. In this research, I select user from user table who has most number of fans. In fact, the user who has most number of review\_count could also be seen as active users.

* Top 10 users with most numbers of fans

1. Review

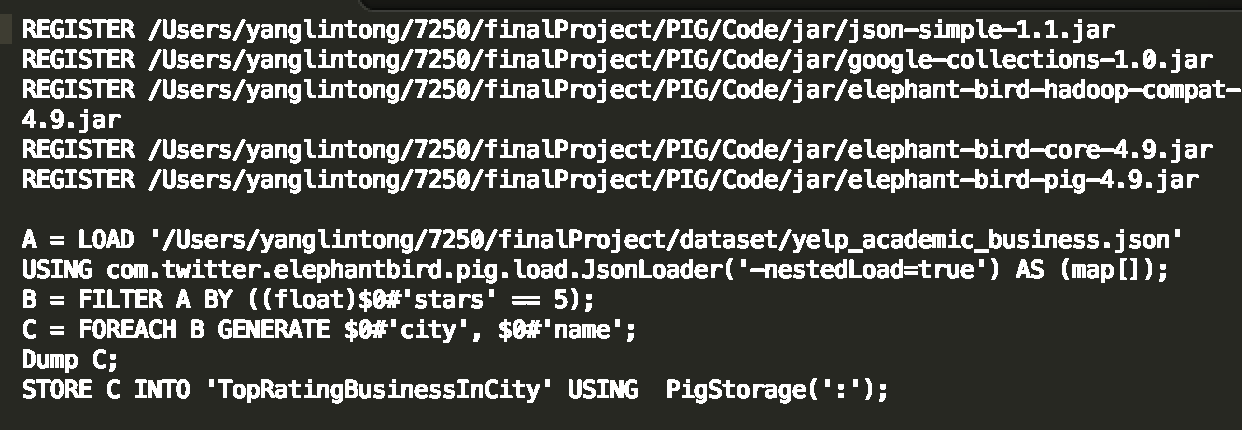
As for review table, besides review number, what we should pay attention to is percentage of each rating. After analysis that, we could position a level of service provided by a business roughly. But there exists one problem, rating of business would be influenced by users personal interests. We should also take notice of votes of “cool”, ”funny”or useful votes for each user.

* find the total number of cool, funny, useful votes on the reviews of a particular user.
* The number of reviews over the years
* Percentage of 1-star to 5 - star ratings

Implementation

In order to implement analysis tasks, I use pig shell, run pig script in it.

This is a screenshot of a script that I try to implement counting top businesses in a city.



The whole implementation and code description could be seen in appendix files included in Code.

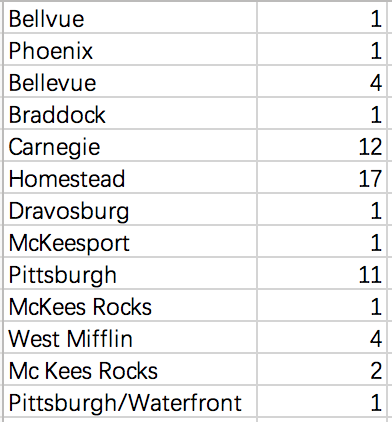
Output

1. Business

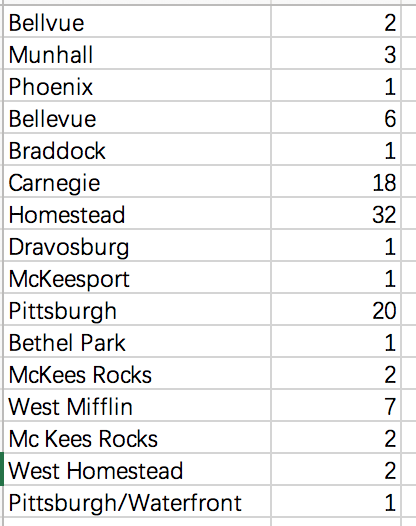
Top business in a city



Number of business in a state above average rating



Business reviewed most



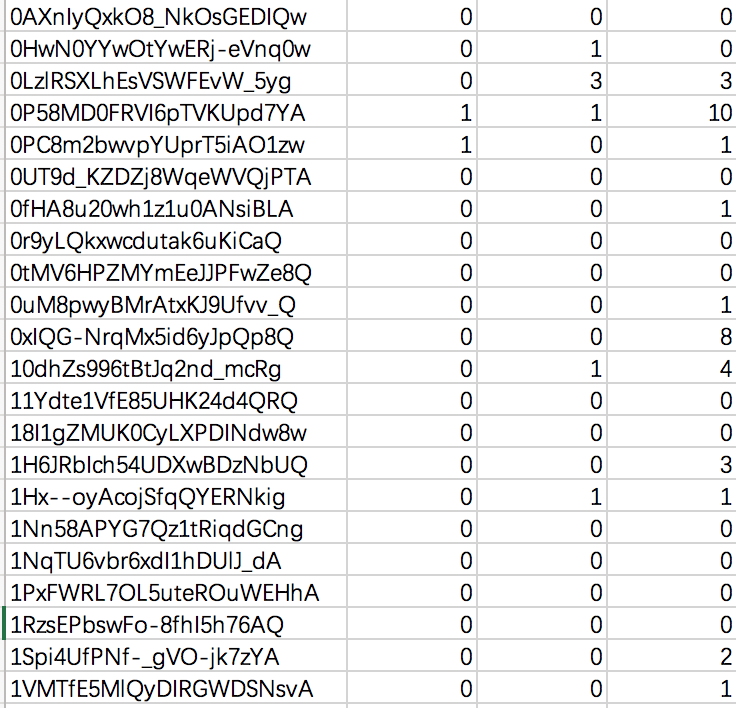
1. User

Top ten users with most fans

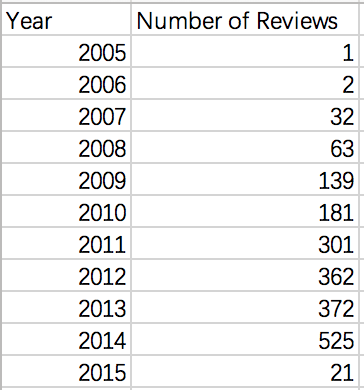


1. Reviews

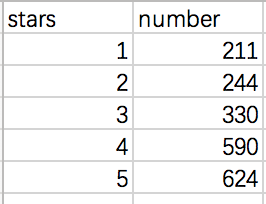
total number of cool, funny, useful votes on the reviews of a particular user



The number of reviews over the years.



Number of 1 - star to 5 - star ratings



*Hive*

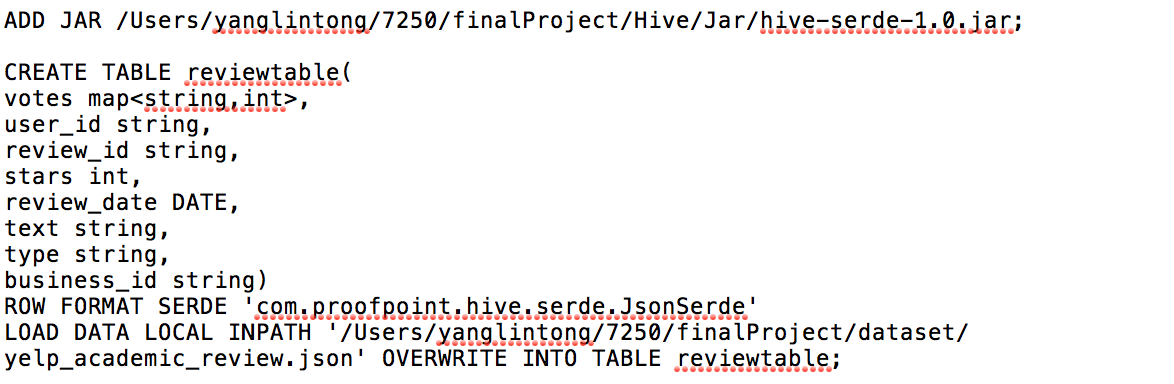
Analysis

There are some logical and useful information we could find when do analysis uses’ reviews. Like what are most used words in customer reviews, what’s the relationship between review word numbers and business rating. Next, I plan to implement sentiment analysis according to count words polarity in review text.

Implementation

Firstly, create three tables in hive. In order to implement this, I need to connect hive, Hadoop and Mysql at start. Then write mysql script to create table and insert value in it.

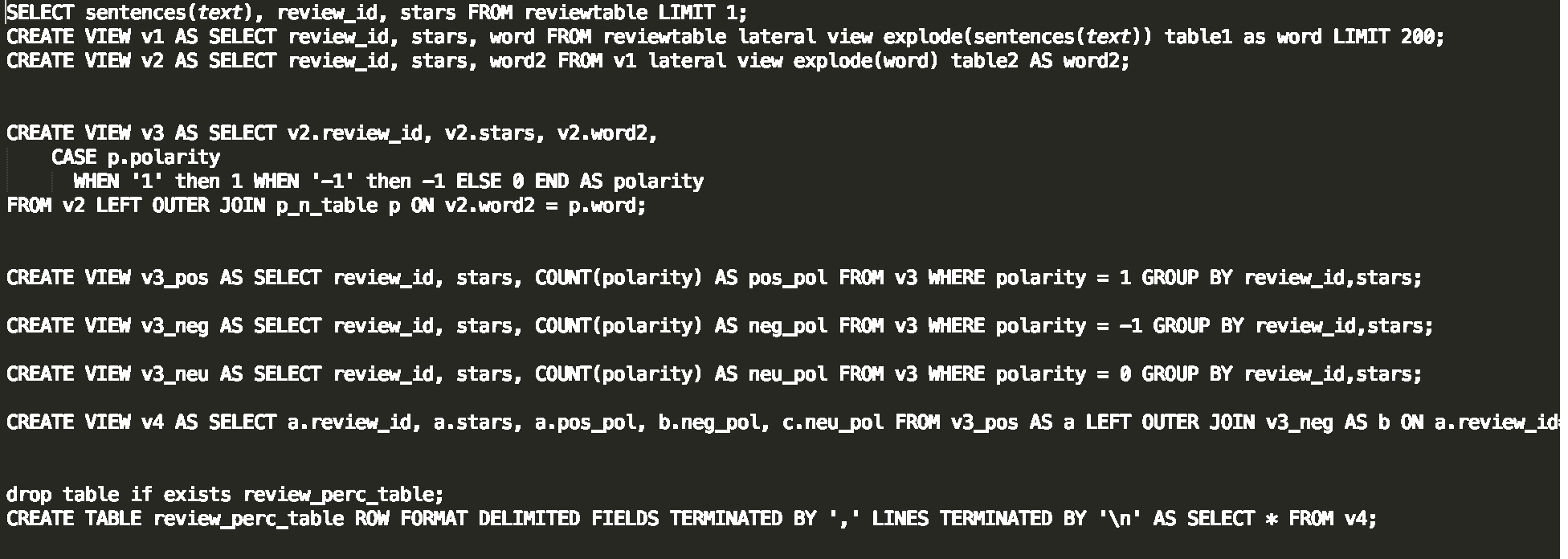
A part of screenshot in creating table sql script.



How Positive or Negative a review is? Add polarity value(positive:1 negative:2, neutral:3) of each word in review and compare sum with the rating to see if it is coinciding. if it tends more to the positive side the rating must be above average and if it towards the negative side it should be below average.

We need to load a word polarity table as the dictionary.

I start with group word by polarity.

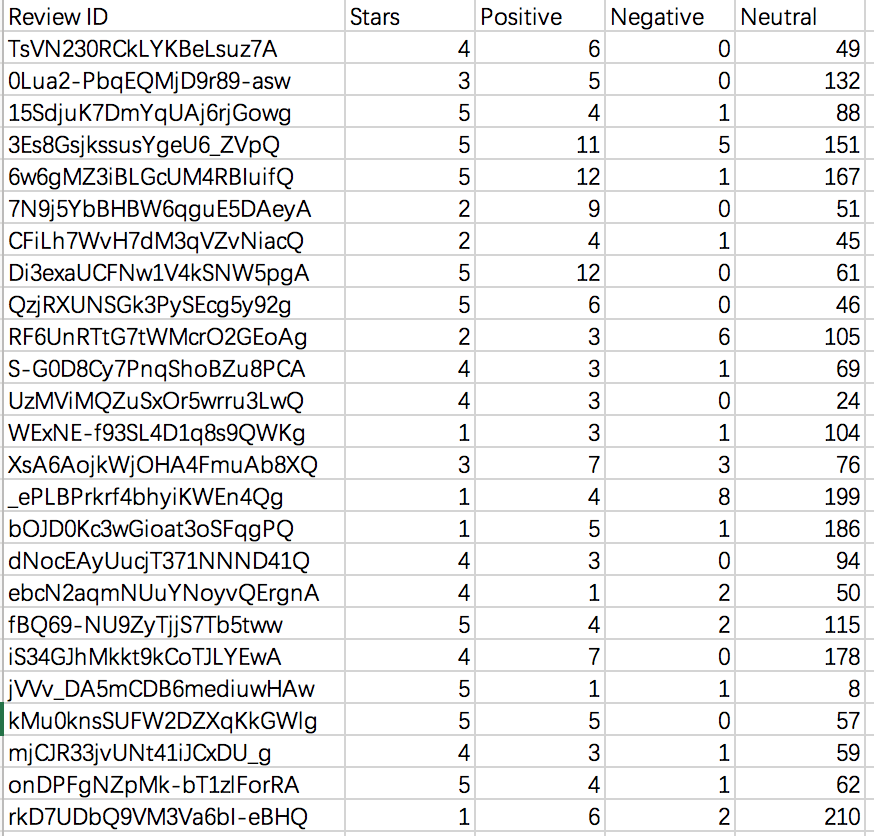


And copy from Hadoop to local.

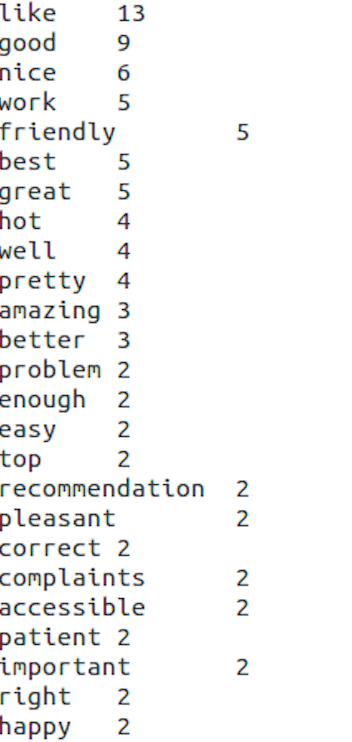
bin/Hadoop fs -cat /user/hive/warehouse/review\_perc\_table/\* > /Users/yanglintong/7250/finalProject/Hive/output/review\_percent.csv;

Output

Stars and Sentiment statistical



The top 25 most used words in reviews



*Summary*

In research this time, I try all kinds of tools connected with Hadoop, and get to know the convenience and high efficiency of them. Especially for Hive, I’m be amazed at its speed of Map Reduce task even though data is of huge volume. It could also connect with Hadoop, mysql, what’ means we could use it to implement data storage and mining easily.

*Appendix*

Dataset Link: https://www.yelp.com/dataset/download