Big Data Strategy

Yelp review

0503 Team G: Yifan Dang, Zongdi Zhou, Dixi Lin, Qihan Hu

Data Information and Analysis Tool

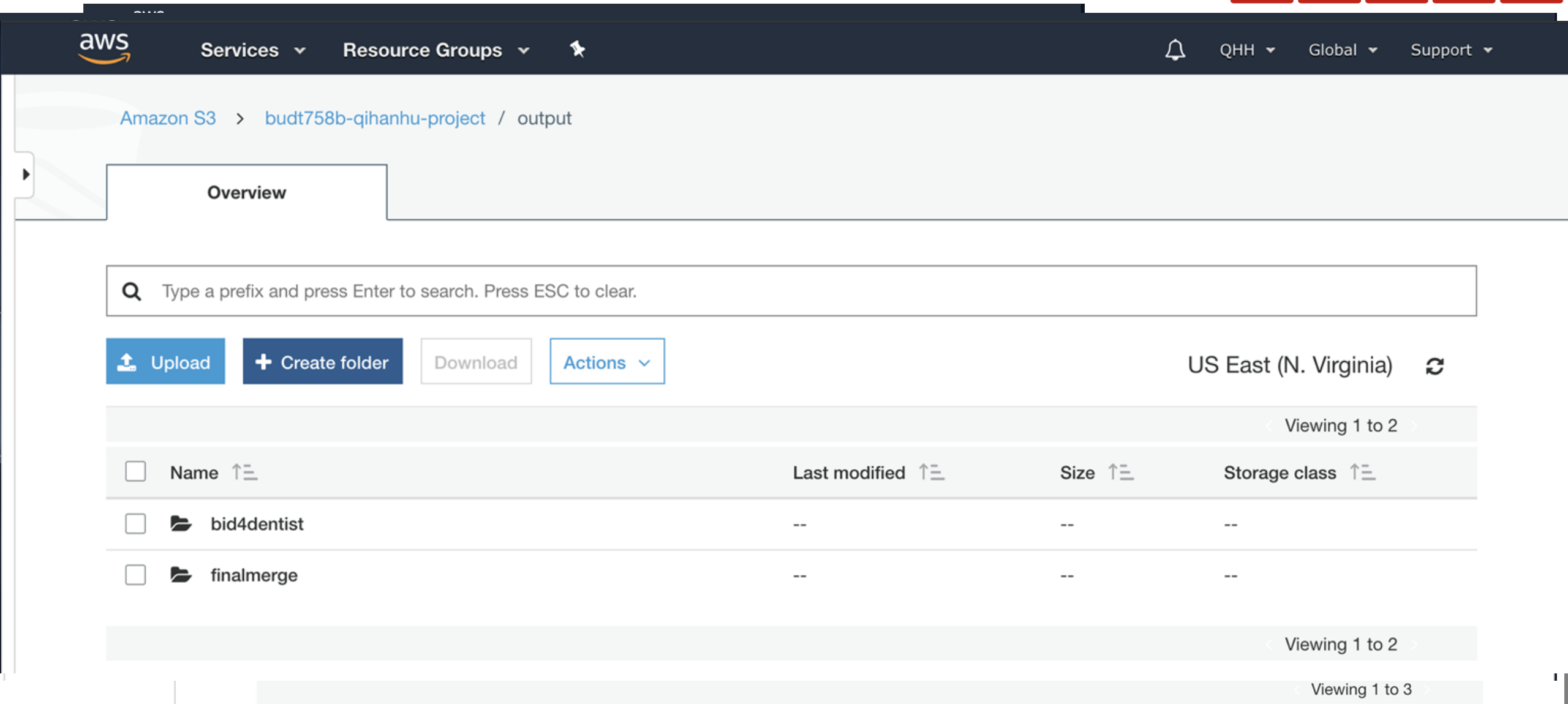
Data Set: yelp_academic_dataset_review.csv (4GB)
(rows:5996996, columns:9)
yelp_academic_dataset_business.csv(163MB)
(row:188593)

Information includes: “business_id”, “cool”, “date”, “funny”, “review_id”, “starts”, “text”, “useful”, “user_id” and “categories”

Tools: Hadoop(virtual Machine, AWS), R, Python, Pig, Spark, Tableau

Data storage strategy

Cluster configuration, location, use of a database or text files



DATA queries

find_dentist_businessid.pig

```
REGISTER 's3://budt758b-qihanhu-project/loudacre/piggybank.jar';

DEFINE CSVLoader org.apache.pig.piggybank.storage.CSVLoader();

business = LOAD 's3://budt758b-qihanhu-project/loudacre/bid_categ.csv' USING CSVLoader AS
(categories:chararray, business_id:chararray);

a = FILTER business BY (categories MATCHES '.*Dentist.*') OR (categories MATCHES
'.*Dental.*') OR (categories MATCHES '.*dentist.*') OR (categories MATCHES '.*dental.*');

b = DISTINCT a;


STORE b INTO 's3://budt758b-qihanhu-project/output/bid4dentist';
```

Pig Script that pulls out all the dentists' "*business_id*" that matches the "*categories*" being "***dentist***" and "***dental***".

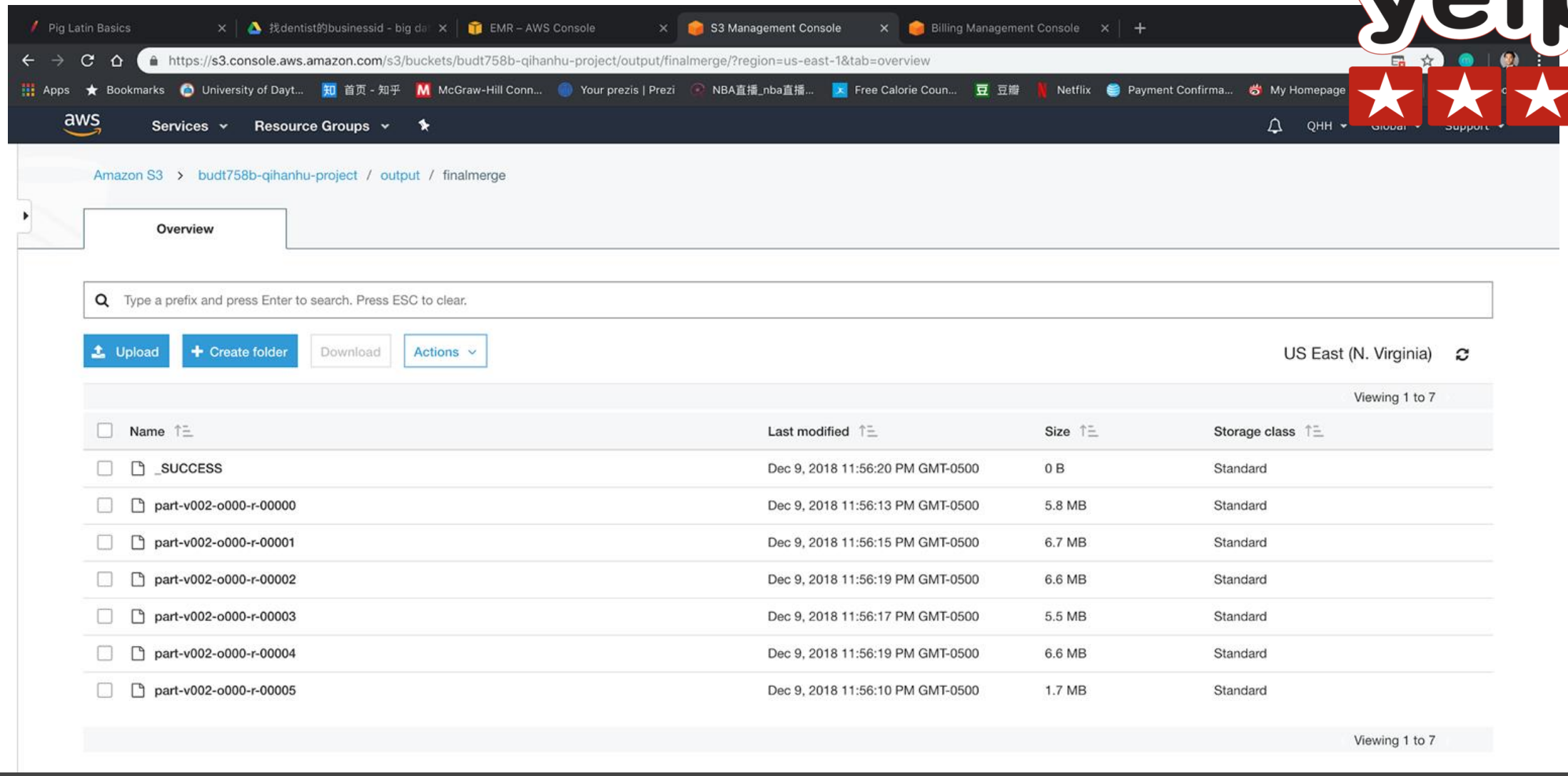
DATA queries Cont.

mergesample.pig

```
REGISTER 's3://budt758b-qihanhu-project/loudacre/piggybank.jar';

DEFINE CSVLoader org.apache.pig.piggybank.storage.CSVLoader();

dentist = LOAD 's3://budt758b-qihanhu-project/output/bid4dentist' AS
(categories:chararray, business_id:chararray);

reviewall = LOAD 's3://budt758b-qihanhu-project/loudacre/final_review.csv' USING CSVLoader
AS (
business_id: chararray,
cool: int,
date: chararray,
funny: int,
review_id: chararray,
stars:int,
text: chararray,
useful: int,
user_id: chararray);

dentistreview = JOIN dentist BY business_id, reviewall BY business_id;


STORE dentistreview INTO 's3://budt758b-qihanhu-project/output/finalmerge';
```

Pig Script that uses all dentists' "*business_id*" that pulled from previous step, then matches against "*final_review.csv*" where contains all the review information including "*business_id*", "*cool*", "*date*", "*funny*", "*review_id*", "*starts*", "*text*", "*useful*" and "*user_id*".

DATA queries Cont.



MapReduce results in AWS.

Descriptive Queries

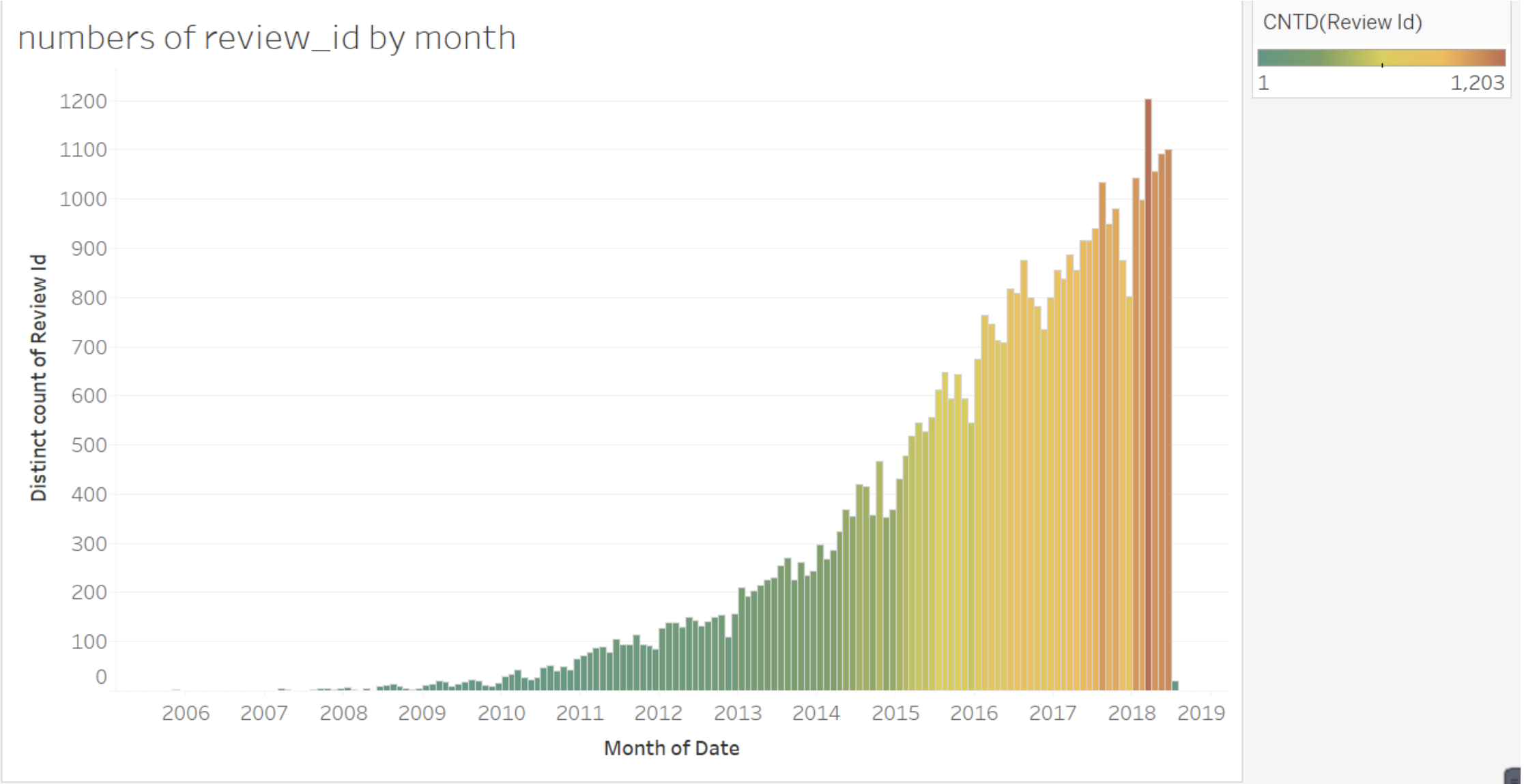
count.pig

```
dentist = LOAD '/Users/zhou/Desktop/bigdata/dentist_review' AS (categories:chararray,
business_id: chararray,
business_idaa: chararray,
cool: int,
date: chararray,
funny: int,
review_id: chararray,
stars:int,
text: chararray,
useful: int,
user_id: chararray);

/* #####total records################################## */

dent_details = GROUP dentist ALL;
dent_total_num = FOREACH dent_details GENERATE COUNT(dentist.review_id);
DUMP dent_total_num;

/* #####stars########################################## */

dentstars = GROUP dentist BY stars;
starnum = FOREACH dentstars GENERATE COUNT(dentist.review_id);
DUMP starnum;

/* #####positive vs negative########################### */

dent_positive = FILTER dentist BY (stars >= 3);
dent_pos_details = GROUP dent_positive ALL;
dent_pos_num = FOREACH dent_pos_details GENERATE COUNT(dent_positive.review_id);
DUMP dent_pos_num;

dent_negative = FILTER dentist BY (stars < 3);
dent_neg_details = GROUP dent_negative ALL;
dent_neg_num = FOREACH dent_neg_details GENERATE COUNT(dent_negative.review_id);
DUMP dent_neg_num;
```

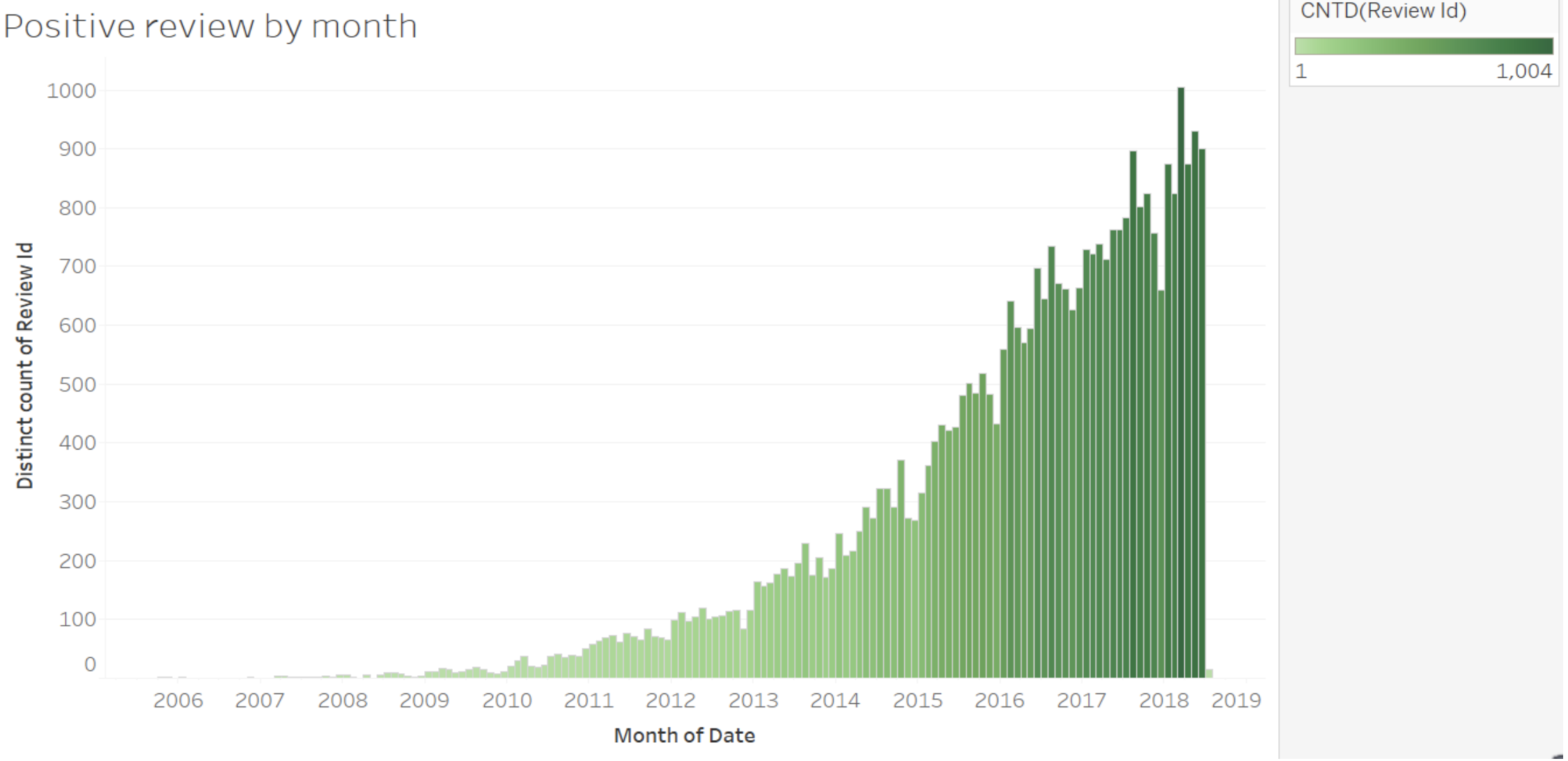
Totally data set after join is 32.8 MB, and we have 44366 dentist reviews, of which 7036 reviews give one star, 1165 reviews give two stars, 651 reviews give three stars, 1658 reviews give four stars and 33856 reviews give five stars.

To sum up, there are 36165 positive reviews and 8201 negative reviews.

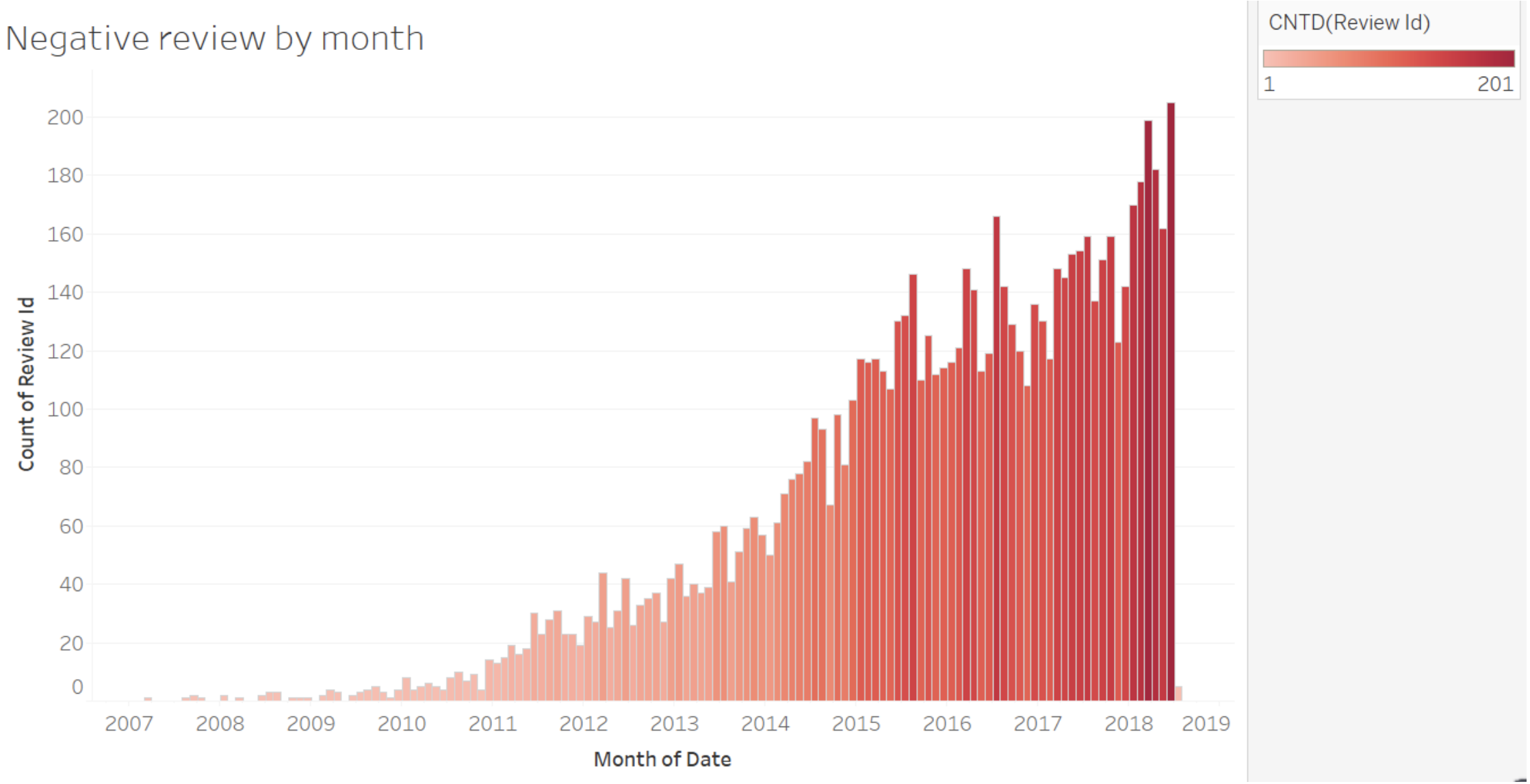
Reviews Trending



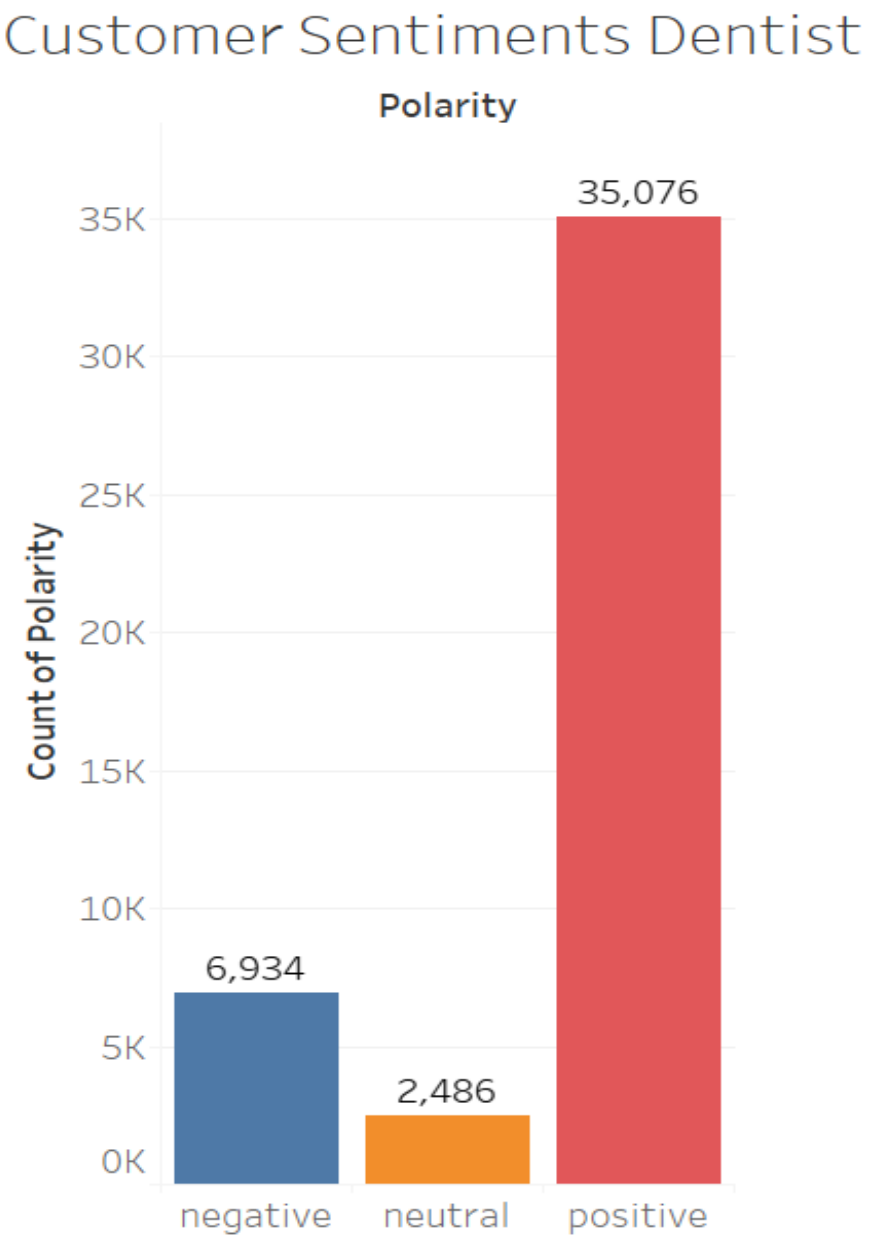
Positive Reviews Trending

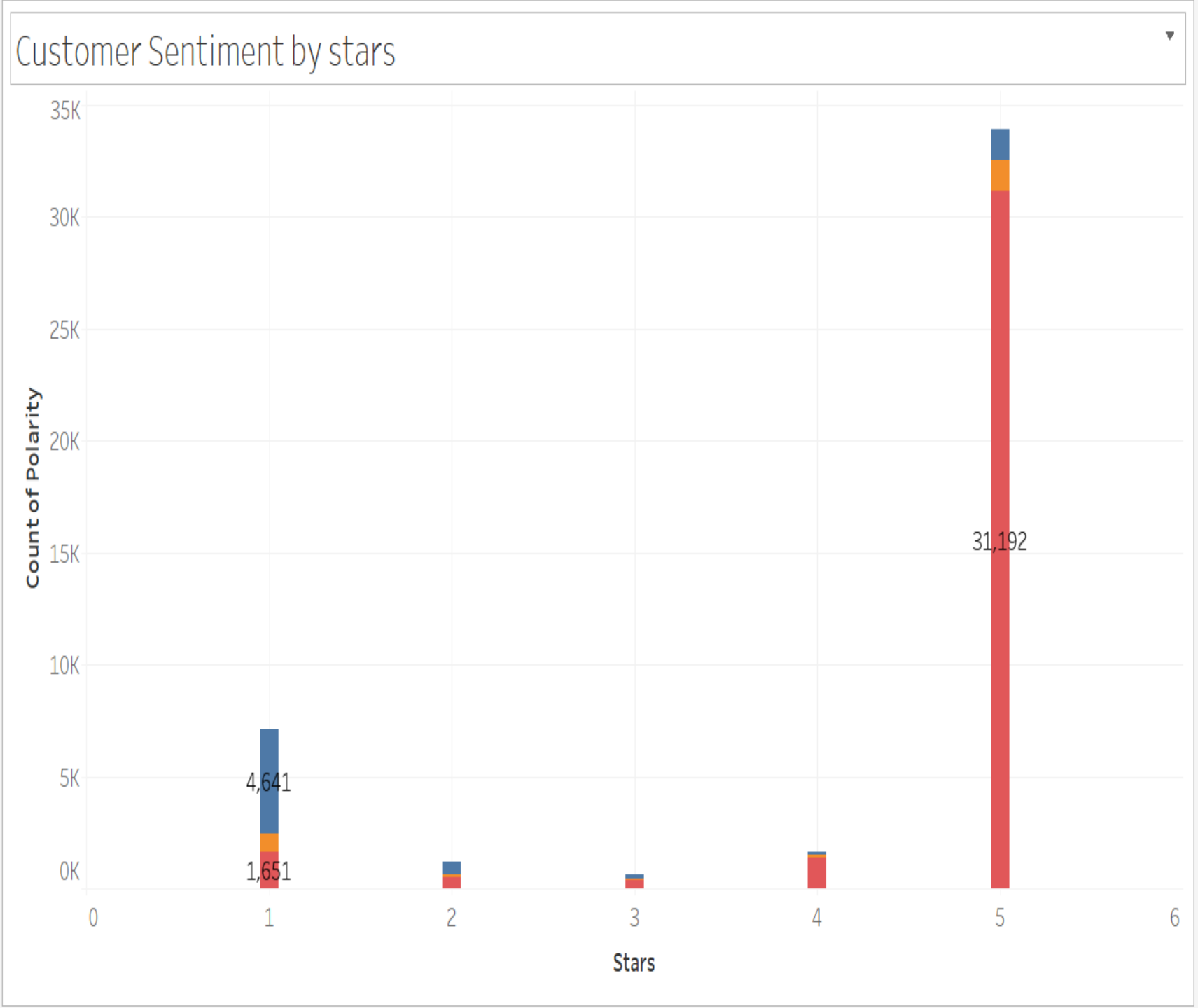


Negative Reviews Trending



Sentiment Analysis



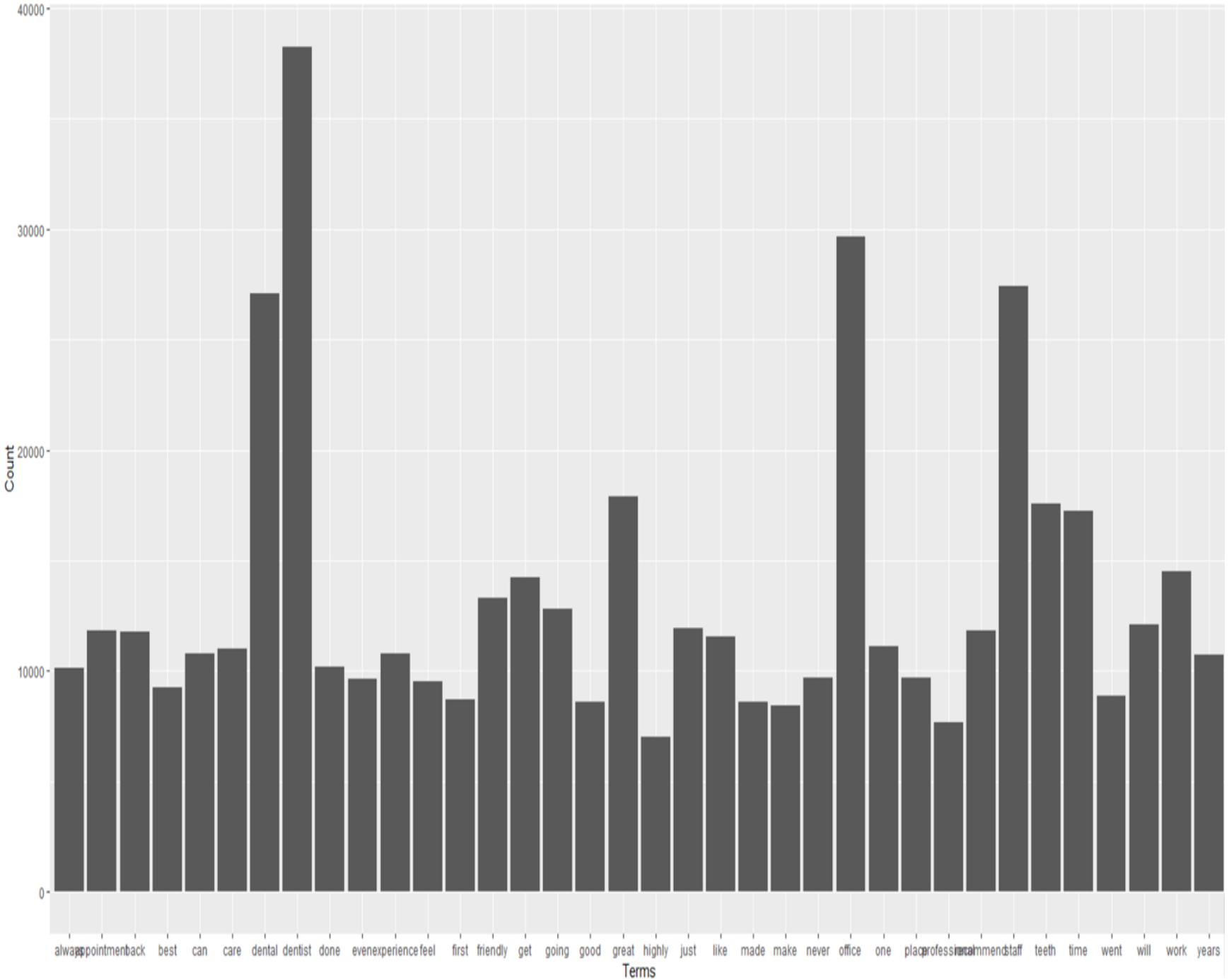


Star 1 and Star 2 has relative higher negative review

Star 3,4 has relative more positive review

Star 5 has a significant positive reviews

Generally, the higher the rank, the more the positive review.



- Convert all text contents to lowercase
- Remove Sparse Terms(in fewer than 10% of documents)
- Remove Stopwords(e.g. “the”, “of”, “and”, “to”...)
- RemoveNumbers
- Stem the words (e.g. users--user, acting--act)
- Strip Whitespace
- weight the term-document matrix by term frequency.

Prediction Model

Dang - VMware Workstation 14 Player (Non-commercial use only)

Player | Applications Places System

Browse and run installed applications | *logistic regression.py (~)

File Edit View Search Tools Documents Help

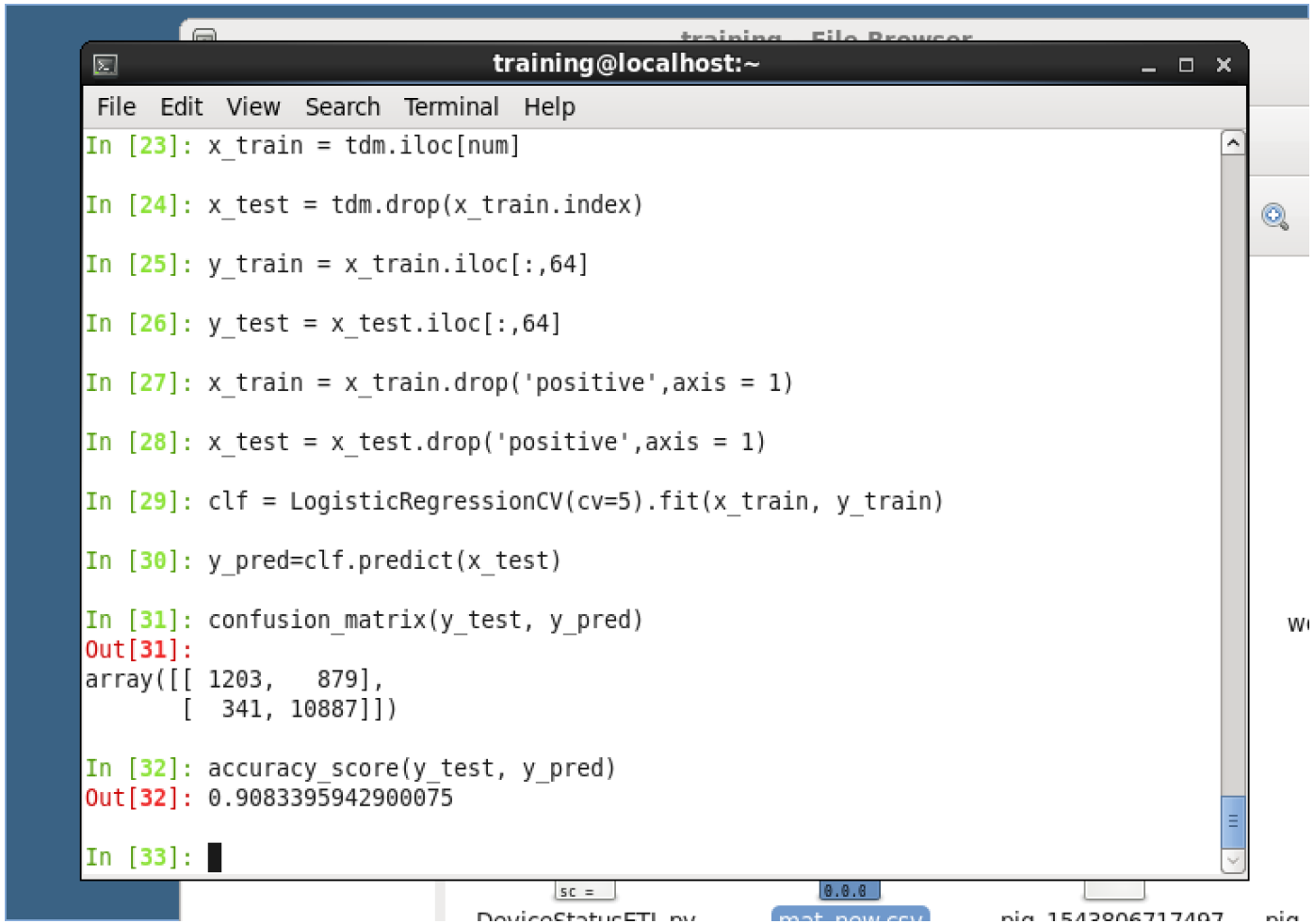
Open Save Undo

*logistic regression.py

```python
####logistic regression####
import os
import pandas as pd
import numpy as np
from sklearn.linear_model import LogisticRegressionCV
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
import random
import statsmodels.api as sm


tdm=pd.read_csv('mat_new.csv')
#create training and testing vars
num = random.sample(np.arange(0,len(tdm),1).tolist(),int(0.7*len(tdm)))

x_train = tdm.iloc[num]
x_test = tdm.drop(x_train.index)
y_train = x_train.iloc[:,64]
y_test = x_test.iloc[:,64]
x_train = x_train.drop('positive',axis = 1)
x_test = x_test.drop('positive',axis = 1)

#fit model
clf = LogisticRegressionCV(cv=5).fit(x_train, y_train)
y_pred=clf.predict(x_test)
confusion_matrix(y_test, y_pred)
accuracy_score(y_test, y_pred)

#baseline
sum(y_test)/len(y_test)

#summary result
logit_model=sm.Logit(y_train,x_train)
result=logit_model.fit()
print(result.summary2())
```

- Data Set: Term Document Matrix (TDM)
- Prediction model: Logistic Regression
- Prediction Tools: PySpark
- Classified negative: star 1,2
- Classified positive: star 3,4,5
- Predicted binary outcome: positive (1), Negative (0)

Prediction Model




```
In [49]: np.array(y_test).mean()
Out[49]: 0.8435762584522916
```

- Random splitting the training data 70% testing data 30%.
- Cross Validation: 5-folds in the training data set t avoid overfitting
- Testing data prediction accuracy: 0.9083
- Baseline:0.8436
- Model is fairly robust
- Random Forest model: 0.90104

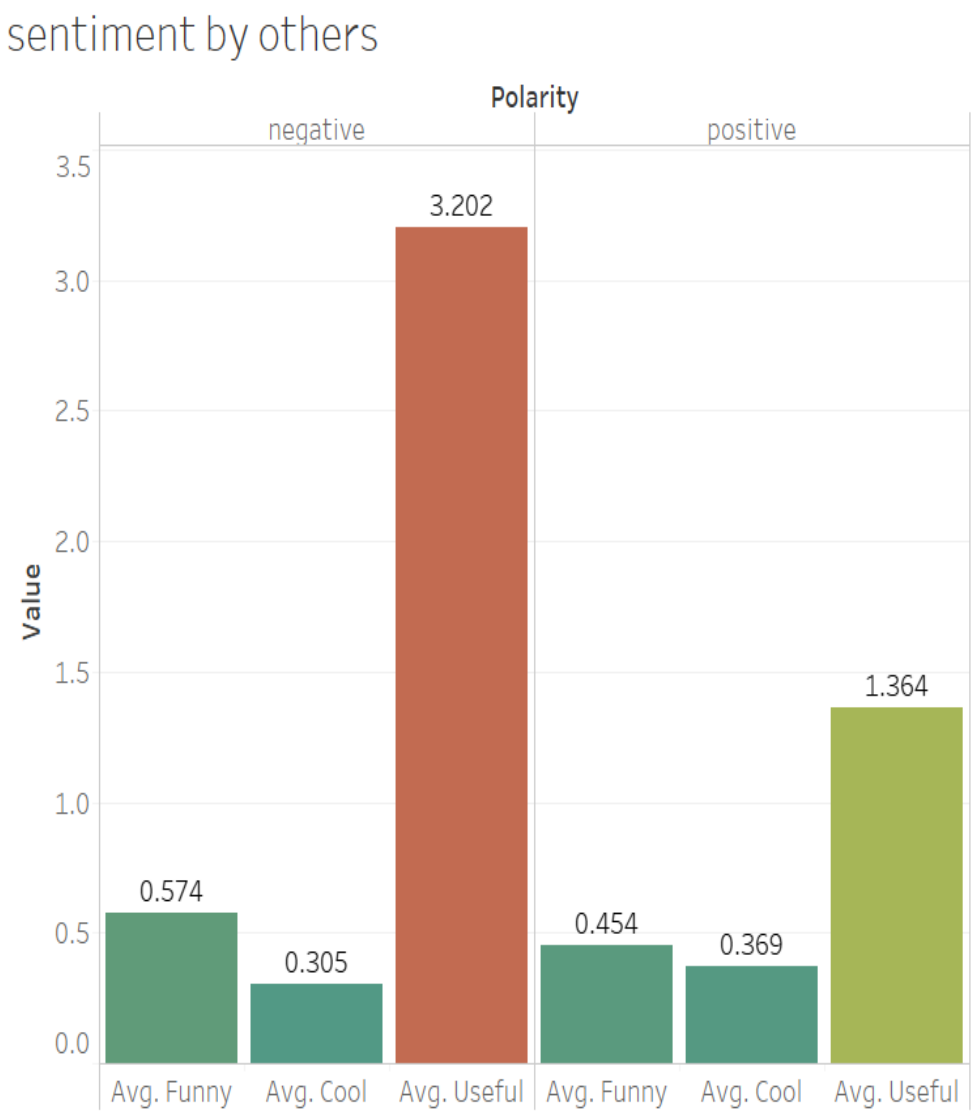
Prediction Model

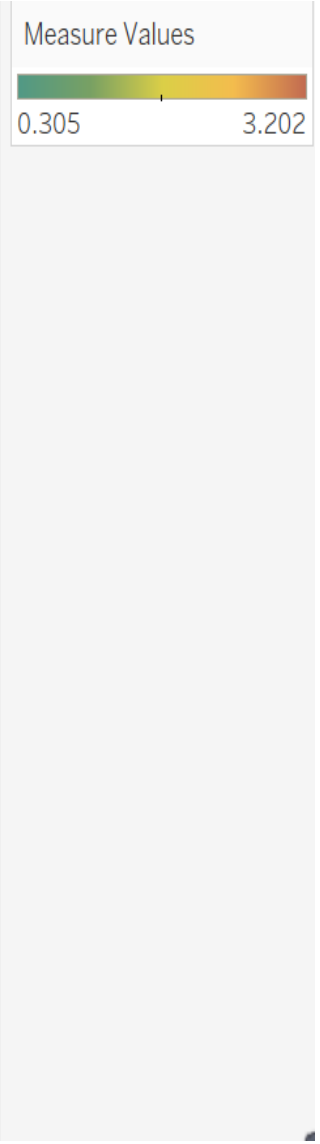
```
In [41]: print result.summary2()
                         Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: 0.464
Dependent Variable: positive         AIC:              14729.4268
Date:               2018-12-10 21:20 BIC:              15263.4138
No. Observations:   31056            Log-Likelihood:   -7300.7
Df Model:           63               LL-Null:          -13630.
Df Residuals:       30992            LLR p-value:      0.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     9.0000
------------------------------------------------------------------
                  Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
------------------------------------------------------------------
also               0.1499   0.0498   3.0068 0.0026  0.0522  0.2476
always             0.8731   0.0641  13.6120 0.0000  0.7474  0.9988
amazing            2.4330   0.1273  19.1173 0.0000  2.1835  2.6824
appointment       -0.2484   0.0296  -8.3949 0.0000 -0.3064 -0.1904
back              -0.1805   0.0353  -5.1150 0.0000 -0.2497 -0.1114
best               1.8005   0.0821  21.9405 0.0000  1.6397  1.9614
can                0.0202   0.0420   0.4822 0.6297 -0.0620  0.1025
care               0.1037   0.0425   2.4415 0.0146  0.0204  0.1869
cleaning          -0.0594   0.0291  -2.0388 0.0415 -0.1166 -0.0023
comfortable        1.6887   0.1168  14.4539 0.0000  1.4597  1.9177
dental             0.0757   0.0243   3.1193 0.0018  0.0281  0.1233
dentist            0.0163   0.0202   0.8052 0.4207 -0.0233  0.0558
didn              -0.2112   0.0436  -4.8419 0.0000 -0.2968 -0.1257
don               -0.4376   0.0442  -9.9082 0.0000 -0.5242 -0.3510
done               0.0349   0.0407   0.8577 0.3911 -0.0449  0.1147
even              -0.2838   0.0413  -6.8705 0.0000 -0.3648 -0.2029
ever              -0.3118   0.0599  -5.2051 0.0000 -0.4293 -0.1944
everyone           1.3567   0.0947  14.3239 0.0000  1.1710  1.5423
experience         0.1992   0.0481   4.1443 0.0000  0.1050  0.2934
family             0.4475   0.0673   6.6448 0.0000  0.3155  0.5795
feel               0.5210   0.0618   8.4301 0.0000  0.3999  0.6421
first              0.1586   0.0483   3.2866 0.0010  0.0640  0.2532
friendly           1.6131   0.0715  22.5551 0.0000  1.4729  1.7532
front             -0.1132   0.0525  -2.1555 0.0311 -0.2162 -0.0103
get               -0.1185   0.0327  -3.6262 0.0003 -0.1826 -0.0545
going              0.1885   0.0415   4.5456 0.0000  0.1072  0.2698
good               0.3294   0.0457   7.1992 0.0000  0.2397  0.4190
got               -0.0305   0.0471  -0.6476 0.5172 -0.1228  0.0618
great              1.4991   0.0529  28.3244 0.0000  1.3953  1.6028
highly             1.6288   0.1141  14.2742 0.0000  1.4052  1.8525
insurance         -0.3303   0.0327 -10.0860 0.0000 -0.3945 -0.2661
just              -0.2375   0.0366  -6.4961 0.0000 -0.3092 -0.1659
know              -0.1109   0.0533  -2.0807 0.0375 -0.2155 -0.0064
like              -0.0047   0.0375  -0.1251 0.9005 -0.0782  0.0688
love               1.8622   0.1194  15.5986 0.0000  1.6282  2.0962
made               0.0871   0.0520   1.6744 0.0941 -0.0149  0.1890
make               0.0242   0.0510   0.4750 0.6348 -0.0758  0.1243
need              -0.0571   0.0497  -1.1499 0.2502 -0.1544  0.0402
needed            -0.1066   0.0478  -2.2321 0.0256 -0.2003 -0.0130
never             -0.5404   0.0415 -13.0177 0.0000 -0.6218 -0.4591
new               -0.1667   0.0439  -3.7962 0.0001 -0.2527 -0.0806
```

```
don               -0.4376   0.0442  -9.9082 0.0000 -0.5242 -0.3510
done               0.0349   0.0407   0.8577 0.3911 -0.0449  0.1147
even              -0.2838   0.0413  -6.8705 0.0000 -0.3648 -0.2029
ever              -0.3118   0.0599  -5.2051 0.0000 -0.4293 -0.1944
everyone           1.3567   0.0947  14.3239 0.0000  1.1710  1.5423
experience         0.1992   0.0481   4.1443 0.0000  0.1050  0.2934
family             0.4475   0.0673   6.6448 0.0000  0.3155  0.5795
feel               0.5210   0.0618   8.4301 0.0000  0.3999  0.6421
first              0.1586   0.0483   3.2866 0.0010  0.0640  0.2532
friendly           1.6131   0.0715  22.5551 0.0000  1.4729  1.7532
front             -0.1132   0.0525  -2.1555 0.0311 -0.2162 -0.0103
get               -0.1185   0.0327  -3.6262 0.0003 -0.1826 -0.0545
going              0.1885   0.0415   4.5456 0.0000  0.1072  0.2698
good               0.3294   0.0457   7.1992 0.0000  0.2397  0.4190
got               -0.0305   0.0471  -0.6476 0.5172 -0.1228  0.0618
great              1.4991   0.0529  28.3244 0.0000  1.3953  1.6028
highly             1.6288   0.1141  14.2742 0.0000  1.4052  1.8525
insurance         -0.3303   0.0327 -10.0860 0.0000 -0.3945 -0.2661
just              -0.2375   0.0366  -6.4961 0.0000 -0.3092 -0.1659
know              -0.1109   0.0533  -2.0807 0.0375 -0.2155 -0.0064
like              -0.0047   0.0375  -0.1251 0.9005 -0.0782  0.0688
love               1.8622   0.1194  15.5986 0.0000  1.6282  2.0962
made               0.0871   0.0520   1.6744 0.0941 -0.0149  0.1890
make               0.0242   0.0510   0.4750 0.6348 -0.0758  0.1243
need              -0.0571   0.0497  -1.1499 0.2502 -0.1544  0.0402
needed            -0.1066   0.0478  -2.2321 0.0256 -0.2003 -0.0130
never             -0.5404   0.0415 -13.0177 0.0000 -0.6218 -0.4591
new               -0.1667   0.0439  -3.7962 0.0001 -0.2527 -0.0806
office            -0.1388   0.0234  -5.9294 0.0000 -0.1847 -0.0929
one               -0.1644   0.0374  -4.3917 0.0000 -0.2378 -0.0910
pain               0.0386   0.0341   1.1320 0.2577 -0.0282  0.1054
patient            0.1093   0.0507   2.1542 0.0312  0.0099  0.2088
place             -0.3839   0.0389  -9.8793 0.0000 -0.4600 -0.3077
professional       1.1064   0.0753  14.6954 0.0000  0.9589  1.2540
really             0.4956   0.0503   9.8567 0.0000  0.3971  0.5942
recommend          0.5237   0.0625   8.3793 0.0000  0.4012  0.6462
see               -0.0414   0.0505  -0.8212 0.4115 -0.1403  0.0575
service           -0.0906   0.0528  -1.7178 0.0858 -0.1941  0.0128
staff              0.7429   0.0365  20.3285 0.0000  0.6713  0.8145
teeth              0.0048   0.0260   0.1861 0.8523 -0.0462  0.0559
time              -0.0231   0.0304  -0.7610 0.4466 -0.0826  0.0364
told              -0.7362   0.0394 -18.7071 0.0000 -0.8133 -0.6591
took               0.2870   0.0562   5.1056 0.0000  0.1769  0.3972
visit              0.1611   0.0539   2.9905 0.0028  0.0555  0.2666
well               0.2657   0.0551   4.8180 0.0000  0.1576  0.3738
went              -0.3528   0.0421  -8.3834 0.0000 -0.4353 -0.2704
will              -0.1384   0.0360  -3.8415 0.0001 -0.2091 -0.0678
work               0.0358   0.0337   1.0630 0.2878 -0.0302  0.1019
years              0.3941   0.0458   8.6150 0.0000  0.3045  0.4838
==================================================================
```

- Summary of the coefficient and significance
- The frequencies of “amazing”, “best”, “everyone:, “comfortable”, “friendly”, “great“, “highly”, “love”, “professional” have a significant positive relationship with the positive review
- The frequencies of “ever”, “insurance”, “never”, “just”, “even”, “ever”, “place”, “went” have a significant negative relationship with the positive review

Interesting Findings





Customer tends to find negative review are more useful than positive review.

Generally, Customer do not personal favor too much on the attitude of the reviews.

Interesting findings

text	business_id	cool	date	funny	review_id	stars	useful	user_id
June 28, 2017	4S358eahT9dfeKpnB1U46Q	0	2017-06-28	0	Om7M3gtKmONI0_sFSDIByw	1	5	4ZQ0gvQ9393Nf36Z32HD3A
June 28, 2017	8pC7jQY8MNHVL0W6pRItfg	0	2017-06-28	0	yXzJ5qzRkgZMhTWK-eE--g	1	1	4ZQ0gvQ9393Nf36Z32HD3A
I have to travel a long way to get…	YPyzzWmKH49ZeAE5XhPjNw	0	2017-03-10	0	CLFJwHtG3U7YLReXeqdfYg	5	0	6bVZqcnQAHui7dtKQYEOQg
I have to travel a long way to get…	YPyzzWmKH49ZeAE5XhPjNw	0	2017-03-10	0	w4q1eWWF8hei_wUI33dvwQ	5	0	Bb70ZSqTDCKKzDstqFXKVQ
Dr. Todd and his team are absolute…	-3K14kIKBH3gBOLf8-XFsg	0	2016-01-12	0	fuv4jteyXS4GDYwdVPe1QA	5	1	Z3nwYQzDUimoUl8LNDIxjQ
Dr. Todd and his team are absolute…	3jB7VWZf1_1V-CXuQN-YoA	0	2016-02-03	0	FHtDfw6mA0TfDKFZ0XTbDA	5	0	Z3nwYQzDUimoUl8LNDIxjQ

Q&A

THANK YOU

24Slides