STAT 605 Final Report

Fangyang Chen, Guanqi Lu, Chong Wei, Youhui Ye,Runfeng Yong

12/5/2019

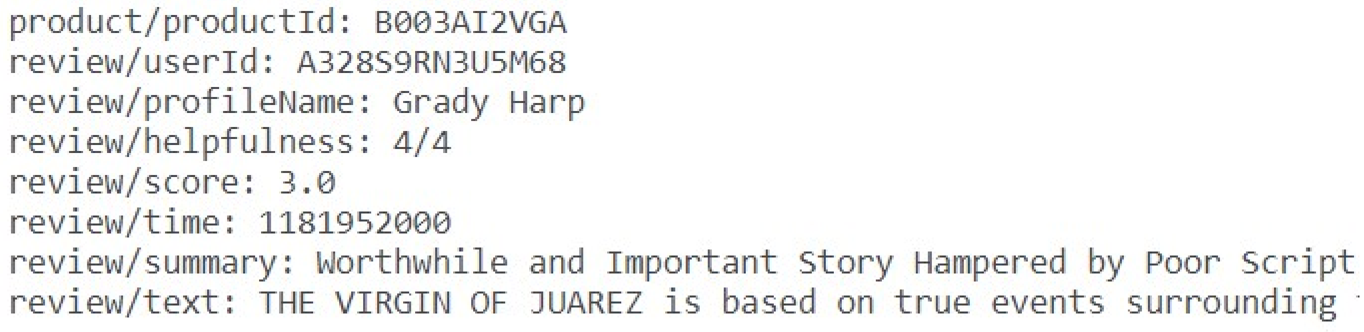
1.Introduction

1.1 Data resource

Our data is about movies' reviews on Amazon, the raw txt file can be found from kaggle by the link below:
https://www.kaggle.com/dm4006/amazon-movie-reviews (https://www.kaggle.com/dm4006/amazon-movie-reviews)

1.2 Description of the data

In this dataset we have 8 variables : the movie's Product ID, the review's User ID, the Profile name(username),helpfulness(number of clicks on 'helpful'), the score({1,2,3,4,5}) user rated the movie, time(timestamp),summary(wrote by user) and review text. Below is part of a record in the txt file.



```
product/productId: B003AI2VGA
review/userId: A328S9RN3U5M68
review/profileName: Grady Harp
review/helpfulness: 4/4
review/score: 3.0
review/time: 1181952000
review/summary: Worthwhile and Important Story Hampered by Poor Script
review/text: THE VIRGIN OF JUAREZ is based on true events surrounding
```

1.3 Question pursued

a. The relationship of scores and the length of review text.
b. The relationship of scores and the helpfulness.
c. The words frequency of high and low helpfulness rate.
d. The way that user's behavior changed over time.
   We'll use the bootstrap, regression and other statistics methods to achieve our goal.

2. Process

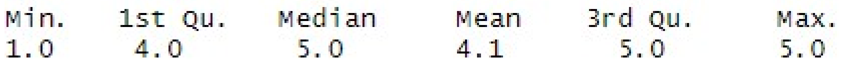
2.1 Data extraction and cleaning

We extracted the useful data from the 9GB txt file in HPC(parallel), and removed the duplications. In the "helpness" part, we removed the 0/0 items. For the timestamp, we converted this to real date. After the data cleaning, we have about 1000000 reccords.This step was done with shell script.

2.2 Analysis

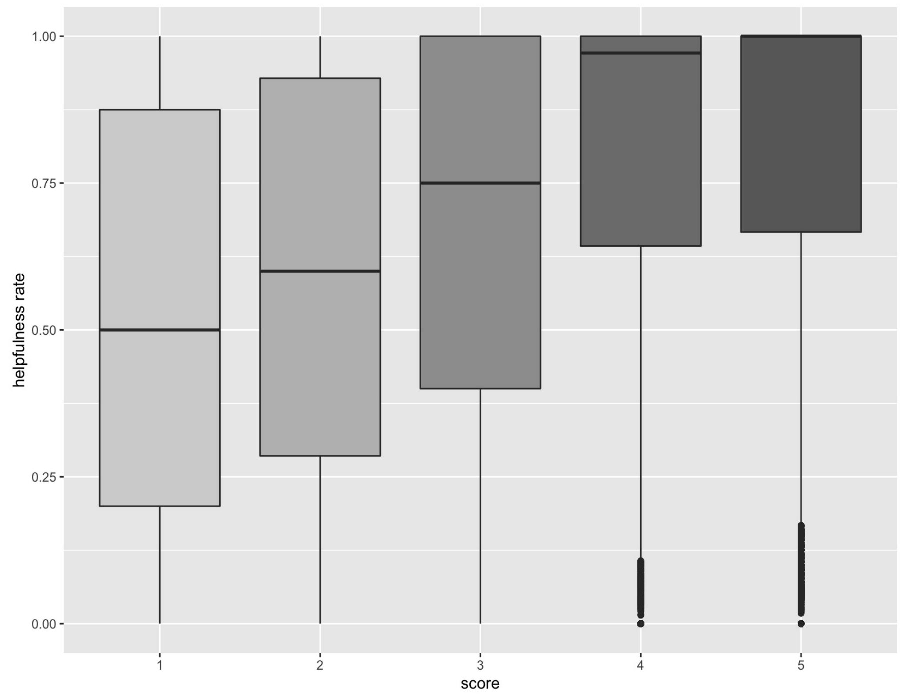
2.2.1 Score & Helpfulness

The average score is 4.1, the lowest is 1. From the summary we can see many users gave full score——5.

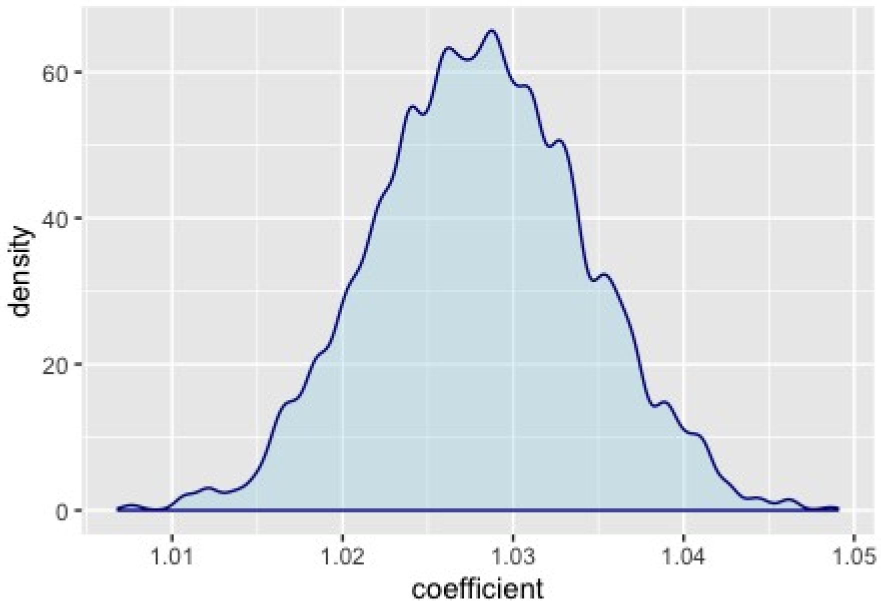


```
Min.   1st Qu.   Median    Mean    3rd Qu.    Max.
1.0     4.0       5.0      4.1      5.0       5.0
```

Here is a boxplot of helpfulness rate of each score category.



We use bootstrap(parallel) to caculate coefficient of the regression between each score and each rate of helpfuness.Here is the distribution of the coefficients.



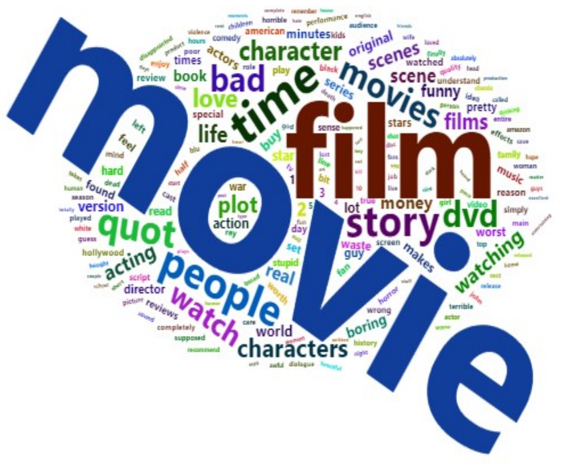
The t statistics is 166.1476, which is much larger than 1.96. The confidence interval of the coefficients is [1.016187, 1.040381]. It can be considered as a normal distribution, so we can do further test. The mean of the coefficients is 1.03, which means the score and the rate of helpfulness has positive relationship.

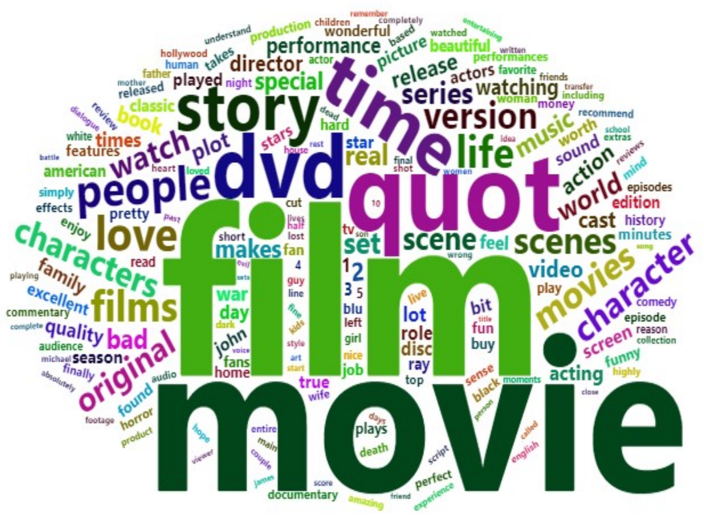
2.2.2 Score & Length of review text

We do ANOVA of score and the length of the review text.The F statistics is 1961.403.It is significant. Then we do permutation(parallel) of the score, after the permutation, the F statistics is still large and the p-value is smaller than \(10^{-5}\). The regression shows that the score and length of text has negative relationship, and the coefficient is significant.

2.2.3 The Contents of reviews & Helpfulness rate

We caculated the wors's fequence of occurences for these reviews that "helpfulness" greater than 10. Here are the wordcloud of the high(right) and low(left) helpfulness rate.

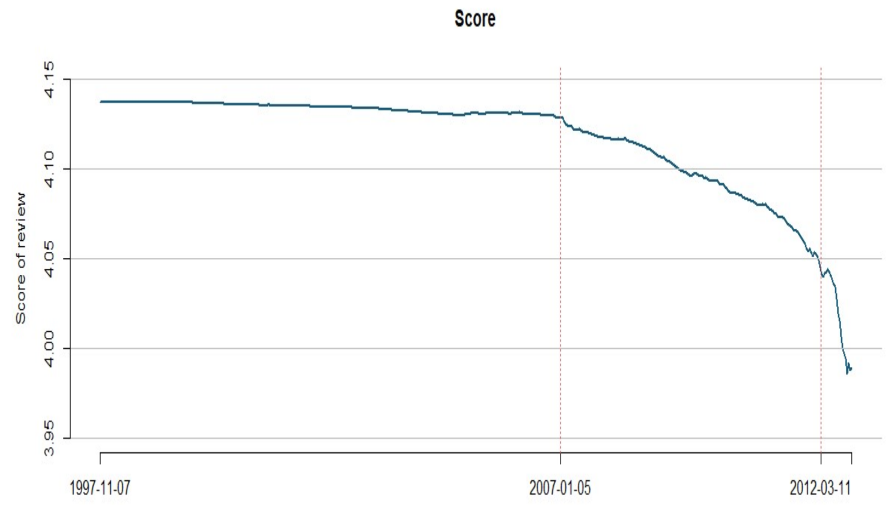


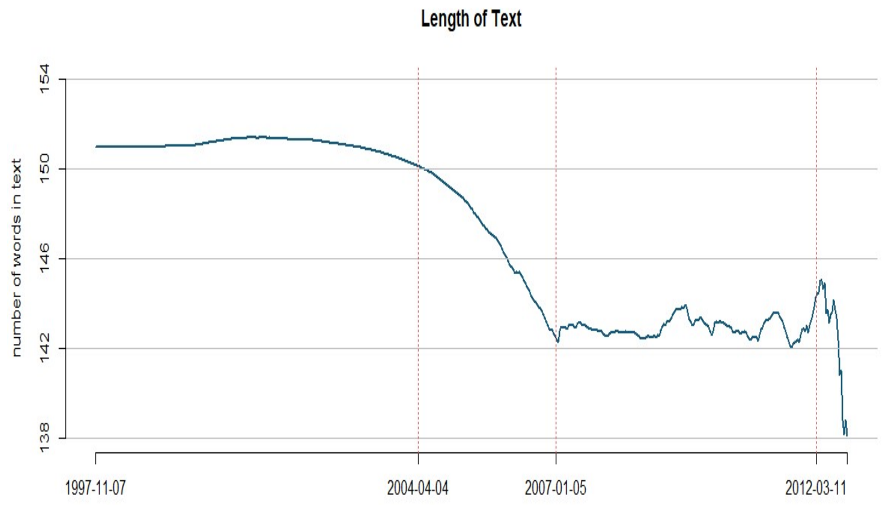


2.2.4 Changes over Time

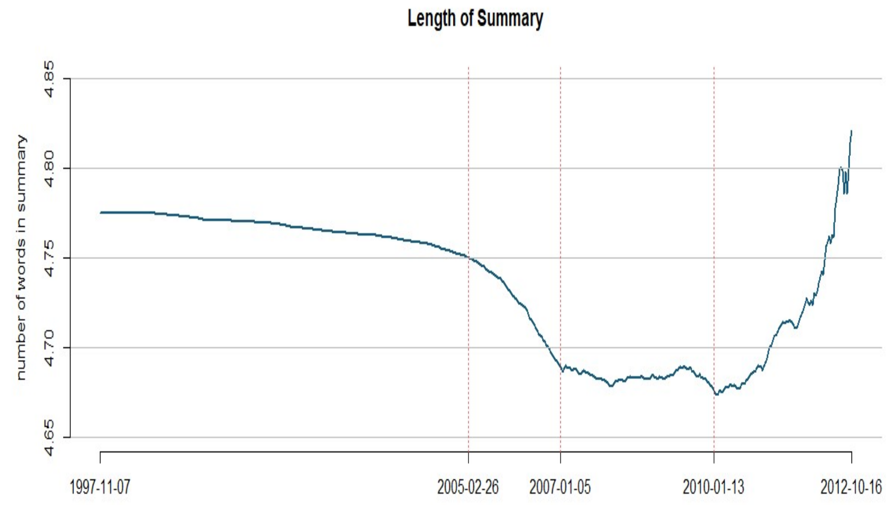
To see how the behaviors of the users changed over time, we plot the scores, length of text, length of summary according to the timeline. These plots contains 500 data points. For each point the y axis was caculated by the average of previous 15 days and next 15 days.

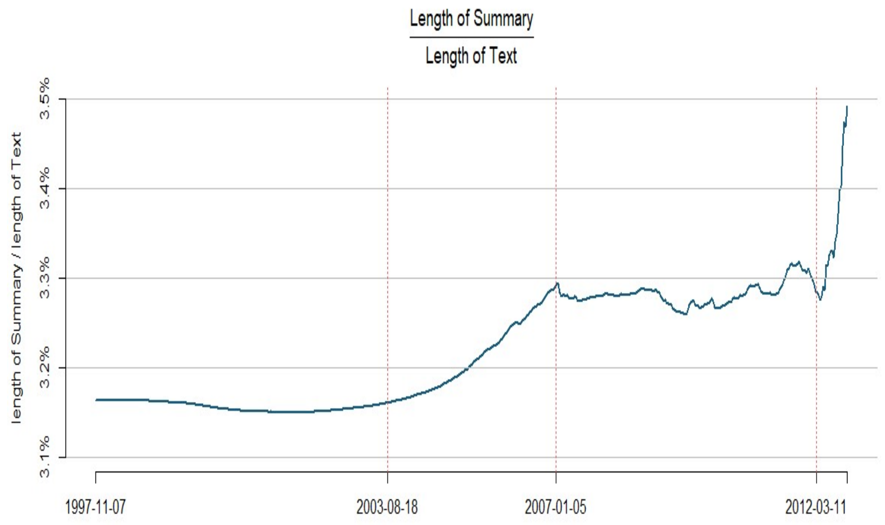
As we can see from the score plot, the score is steady before 2007. After 2007, the score dropped over time. People became stricter when they give score to the movies.
The length of review text also dropped a lot between 2004 and 2007, then it became steady till March 2012, after 2012, the length of review started to drop again.





However, the length of summary only dropped before 2010. After 2010, the length of summary became larger.
Then we checked the ratio of length of summary and length of review text, we can conclude that when users give comments to the movies, they are more likely to write long reviews in the past. Nowadays, people tend to write shorter reviews, and focus on much shorter version of comments——summary.





3. Conclusions and Weaknesses

According to our analysis, we can found out that the score of the movie has negative relationship with the length of the review text.It makes sense that when viewers dislike a movie, they will use more words to express their unsatisfaction.

The score of the movie has positive relationship with the rate of helpfulness, which means that the positive reviews will be easier to receive “helpful” feedback.

Over the 12 years, the scores that users gave to the movies dropped , which may suggest that people became more strict nowadays due to the development of movie industry. The users are too “busy” to write long reviews nowadays. The summary would be another choice when they have comments on movies.

Our weaknesses are: The regression we did is only linear regression, there may be other better models. The analysis of time is only done by plots, we didn’t get into the quantity relationships.