STAT 605: Project Proposal

Kou Wang kwang432

Shushu Zhang szang695

Xinyue Wang xwang2438

Yiqun Xiao yxiao85

Sixu Li sli739

1 Data Set

The web data is provided by Jure Leskovec, an associate professor from Stanford University. It's a data set containing about 35 million reviews from Amazon. The corresponding time spans from June 1995 to March 2013, a period of 18 years. The size of whole data set is 11 gigabtyes and it is available on website Amazon reviews. Note that Amazon merged some products' reviews, the raw data may contain some duplicates, so we need to do the data preprocessing carefully.

2 Project Goal and Statistical Questions

In Amazon's review dashboard, the rating score only provides the overall evaluations toward each product. But actually, among those text reviews, there are a lot of precious "words-level" information. So in this project, we are interested in *mining which words have strong relationships to the five star rating and which ones highly correlate to the negative reviews.* For example, we maybe would find out that the word "overprice" appears many times in the text reviews corresponding to 1 or 2 star rating. In such way, we may obtain some detailed information toward those business, which could be used to either improve the qualities of the products or guide the marketing strategies.

3 Descriptions of Variables

In general, the data set contains ten variables, including three product-related variables: product ID, product title, product price; and seven review-related variables: user ID, user name, the helpfulness, the score, the time, the summary, and the text of the review. As we can see, the data set has already been processed where the "summary" variable extracts the most useful words in the "text" that helps us identify the attitude of the reviewer, which is also reflected by the "score" that the reviewer gave to the product/business. Although the price of the product is highly related to the review, the variable contains a great proportion of missing value. Therefore, "summary", "score" are the most important variables in our analysis.

4 Statistical methods

In this project, we will be faced with high dimensional data given by customers' reviews. We are planning to adapt some traditional regression methods like logistic regression to deal with this problem. At the same time, we may also try to use methods based on machine learning. For example, it is reasonable to use some technics in Natural Language Processing (NLP) such as sentiment analysis for this task. We are looking for a model which is able to identify words that are strongly related to ratings and finally give a statistical summary indicating the advantages and disadvantages of a product or business (See project goal in Section 2). We are still exploring more possible ways to build up the model and looking for a better method. These are our preliminary thoughts.

5 Computational Tools

In order to answer the questions above and preprocess the raw data, we propose using both bash shell and R script. Bash shell is going to use as stream processing since the data to be handled is big enough that requires overwhelmed memories. R scripts are mainly used for textual analysis and modeling. The expected packages are "quanteda", "text2vec", "tidytext", "stringr", "spacyr". These packages are all for natural language processing. Besides, corpora packages may be used for data cleaning, such as eliminating stop words and revise abbreviations and miswritten words. HPC could be used in this case since our model may require extra computing power.