DS 410 - Project Proposal

Yuya Jeremy Ong & Yiyue Zou

Introduction

Web services in recent years have allowed consumers to purchase goods online from a huge catalog of products. This in turn has allowed companies like Amazon, Jet, and Walmart to accumulate data from their customers, which can be analyzed to further observe various different questions and hypothesis about their consumers, products and purchasing behaviors. With better insights and analysis, consumers, sellers and retailers can make better informed decisions and gain a better insight into purchase patterns, trends and behavior surrounding online shopping. In our project, we propose to analyze customer review data and product metadata from Amazon, curated by McAuley et. al [1, 2], to analyze and several different questions regarding the relationships between the consumers, products and the corresponding reviews associated between the consumer and the product.

We plan to utilize modeling techniques which allow us to leverage Apache Spark, Pig, MLlib through the MapReduce paradigm to better leverage the power and scalability these tools have to offer for the analysis of large scale datasets. We will also focus on analyzing the various performance metrics associated with big data analytics, including cluster configurations, runtime performance and scalability. If time persists, we will also look into analyzing the potential bottlenecks of our system and propose (and potentially implement) fixes remedy the performance issues.

Amazon Dataset

The dataset consists of approximately 142.8 million records (approximately 20 GB) of product reviews aggregated between May, 1996 to July, 2014 and split into three different key components including the reviews, product metadata and link information. In our work, we will be utilizing the preprocessed version of the dataset, where the author of the dataset has removed duplicates of the data. The newly scaled down dataset contains 84.3 million records (approximately 18 GB). The review component consists of the customer's ratings (scaled from 1 to 5), text reviews and the helpfulness vote associated with the product. The product metadata provides further information associated with the product's description, category, price, brand and image features. Finally, the dataset also consists of link information pertaining to the information on what consumers have also viewed or purchased Amazon. Permission to utilize the full-scale dataset has been authorized by the author.

Proposed Objectives

In the following section, we will describe some of the key questions we will be answering through our investigation with the Amazon Dataset. The analysis is divided into two different components where we will be analyzing the dataset and building predictive models through these three different phases.

0.1 Exploratory Data Analysis

To better understand the properties and attributes of the dataset, we will extract some key fundamental statistical characteristics of the dataset - including the mean, variance, standard

deviation as well as the kurtosis of the dataset. Furthermore, we will also perform some simple correlation analysis between the different features to better understand emerging relationship between things such as product reviews and review helpfulness. This analysis will later be useful for building more complex relational analysis between other features of the dataset.

0.2 Sentiment Analysis

In this phase, we will build a predictive modeling with the objective of determining the sentiment of the product based on the given text of the review. In constructing the model, we make the assumption that the higher the user rating, the more positive the customer's sentiment is and vice-versa. Thus in converting our ratings to appropriate labels, we will

To build our classifier, we will construct our feature vector utilizing word embeddings utilizing techniques from Mikolov et. al and Le et. al [3, 4] to extract features from unstructured text. Through this, we will utilize this feature vector to build a binary classification model, to predict discrete value of either a positive or negative sentiment or a regression based model for a continuous based output of the sentiment score. To complete this process, we will utilize some of the pre-existing packages already offered under Spark (Word2Vec) to construct our feature vector and modify it towards building review embeddings for embedding a single review as a vector in an embedding space.

By constructing our word embedding, we can perform further analytics on our review data set through performing several different analytics on this embedding. One notable analytics we can perform is k-means clustering. By applying this algorithm over our dataset, we can better observe and understand relationships of different clusters which emerge within consumer reviews and generalize typical characteristics on the types of responses consumers may potentially have on a product.

0.3 Cumulative Analysis

In our final analysis, we will devise a conglomerate analysis based off of the two different analytics we have performed in the previous phases to derive insight on the integrity of the reviewer's opinions on the product. In our analysis, we will utilize the "usefulness" of our dataset to analyze to find out which other features are significant in writing trustworthy and reliable reviews. In particular, we will observe the trends by product category, price, brand, image/presence and consumer sentiment.

References

[1] J. McAuley, C. Targett, Q. Shi, and A. Hengel. 2015b. *Image-based recommendations on styles and substitutes.* In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR15). ACM, New York, NY, 4352. DOI: http://dx.doi.org/10.1145/2766462.2767755

[2] J. McAuley, R. Pandey, and J. Leskovec. 2015. *Inferring Networks of Substitutable and Complementary Products*

[3] T. Mikolov, *Distributed Representations of Words and Phrases and their Compositionality*

[4] Q. Le. 2014. *Distributed Representations of Sentences and Documents*