DS 410 - Project Proposal

Yuya Jeremy Ong & Yiyue Zou

Introduction

Online retail services on the web has recently dominated the digital landscape within the past decade, making it easier for consumers to have easy access to a wider variety of products without having to physically go to a retail location to make their purchases. Consequently, this has allowed for these web services to aggregate and record massive amounts of purchase records as well as product reviews in large scale data warehouse infrastructure. Companies such as Amazon, Jet, and Walmart have already aggregated massive amounts of data from their consumers, which can be analyzed to further observe various trends and patterns to provide a better customer experience - especially through the implementation of a better product recommendation system.

In our project, we propose to utilize the dataset aggregated from Amazon.com, curated by McAuley et. al [1, 2], to analyze data across products, reviews, meta-data information and related products towards the implementation of a recommendation system. The common form of recommendation systems rely on a variant of the collaborative filtering algorithm which rely on content-based systems have often been utilized to drive the recommendations. However, often times other features such as customer sentiment and other indirect features have not really been considered together as a ensemble model based on a hybrid collaborative filtering model to drive customer recommendations.

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 and evaluating the various performance metrics and program flexibility associated with these big data analytics tools, including cluster configurations, runtime performance and scalability.

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 - email proof has been attached as a supplementary material in the dropbox. Our only limitation to this would be to scale down our dataset and potentially work with only a small subset of the entire category of products as a proof-of-concept.

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 assume that anything above 3.5 stars will be considered positive and anything below or equal to 3.5 stars will be negative.

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 Recommendation System

In our final phase, we will utilize some of the features we have identified in the first phase, as well as our review embedding feature vector generated from the second phase to construct our hybrid collaborative filtering model. We will look into further literature pertaining to the augmentation of the original collaborative filtering algorithm to utilize other metrics for distances based on these features.

Some notable algorithms we will attempt to experiment with is Matrix Factorization and K-Means Clustering. These algorithms and methods will serve as a baseline for some other algorithms and methods we will look into through more literature investigation. As for our comparison metrics, we will utilize the provided meta data from the Amazon dataset as our

weakly labeled samples to compare our performance and analyze differences or similarities in the recommendations each of our algorithms provide.

Project Deadline Estimation

We propose to finish all of the above objectives based on the tentative schedule formed below. Much of the dates and objectives are not solid due to further research in the area to figure out if any other methods can be implemented. Due to time constraints and circumstances we will notify the instructor for any changes in direction and plans if necessary.

Target Date	Objectives
2/7 - 2/8	Upload dataset to V-Lab Cluster. Begin literature search on word embedding architecture for sentiment analysis on Spark, recommender system algorithms (including k-means and matrix factorization based CF).
2/8 - 2/10	Perform basic exploratory data analytics, write source for data preprocessing and compile general report of dataset statistics. Finish all basic literature search and devise plan for model pipeline.
2/10 - 2/16	Begin implementation of document embedding for Spark architecture.
2/16 - 2/23	Start compiling report for mid-term project report. Begin prototype implementation of k-mean based recommender system.
2/23 - 3/2	Prepare for mid-term project report. Begin work on CF based algorithm for recommender system.
3/2 - 3/16	Seek out other recommender system algorithms to implement. Begin setup for full experimentation on performance and evaluation of models.
3/16 - 3/30	Perform experiments on cluster with hyperparameter tuning and other repeated evaluation of performance.
3/30 - 4/6	Performance tuning of algorithms and further evaluations.
4/6 - 4/20	Compile final project report and prepare for presentation.

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*