AMAZON BOOKS REVIEWS DATASET ANALYSIS

Engineering Big Data Systems

April 2016

BY:

AAYUSH SHAH

PALASH KOCHAR

SAURABH GOYAL

OVERVIEW

We have selected Amazon Review dataset because we wanted to learn more about Machine learning, Text Mining and Information extraction techniques. The dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014.

This dataset includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, group, and image features), and links (also viewed/also bought graphs). This data set gave us perfect platform to learn about new technologies such as Pig, Hive, HBase, MAHOUT, Python, IBM SPSS and R studio. In order to perform analytics, we have created a cluster on 1 Master and 9 Slave nodes on Amazon Elastic Map Reduce with Pig, Hive, HBase, MAHOUT and R on it.

The metadata part of the dataset was not in a format feasible to query through so we wrote a java code in order to parse the data into a json structure using GSON library.

ANALYSIS

SENTIMENT ANALYSIS

We did this analysis in order to determine attitude of the person who was writing review with respect to overall contextual polarity of document. To do this analysis we used a dictionary of 20000 positive and negative words. Based on the words from this dictionary we assigned +1 score to each positive word and -1 score to each negative word. After assigning score to each word in the review we grouped all the words and calculated the sum and reduced it to a scale.

First, we have uploaded our JSON file into HBase with the help of Hive script. Then we retrieved relevant data from HBase through Pig and wrote a Pig script to calculate sentiment of each review. We exported this data in R extracted required columns using Sqldf and plotted it in a graph using ggplot2 to further analyze the trend.

TEXT MINING AND INFORMATION EXTRACTION

In this analysis we did text mining and information extraction to find out various categories of words about which people have used to write the reviews. To do this analysis, we used IBM SPSS which uses inbuilt Natural Language Processing and Machine Learning algorithms to extract information from various documents and analyze common trends between documents.

For this we generated a statistical file with help of IBM Statistics Modeler. This file indexes data such that it can be accessed in minimal amount of time. We imported this statistical file in IBM SPSS text analytics for the tool and built categories. After building categories we generated category web graph which was very essential in telling us which categories we should combine or create new categories that better account shared responses.

Book Recommendations

Purpose of doing this analysis is to recommend books to users based on their similarity with other users. This is User-based recommendation. In order to do this, we generated a csv file with mahout compatible format from our dataset in HBase using Hive and exported that csv file in HDFS. After exporting that file in HDFS we ran MAHOUT to find top 10 recommendations for the users.

To interpret mahout output we wrote a python script which takes few input arguments and gives out the rated or reviewed books and recommended books for a particular user as an output. We have also written a MapReduce program to find out most recommended books among all users. We exported this data in R and Plotly to better analyze the recommendations and find out which book has been recommended most to users.

Helpfulness of Reviews

This analysis shows to what extent a review is helpful to other users. In order to analyze the helpfulness of reviews we devised an algorithm and implemented that in Pig to calculate the degree of helpfulness of each review. After calculating the degree, we exported that data in R, Plotly and QlikSense to analyze and understand the Helpfulness of reviews with various other dimensions such as rating of the book, length of review etc.

Analysis of Common Syntax in Reviews

In order to determine most common syntax among the reviews we analyzed the first word that the review begins with. It gives us the basic initial understanding of the tone of review, whether it was written with the perspective of first Person or third person etc. To perform this analysis, we wrote a MapReduce program that takes in input of reviews and outputs first word and its count. After getting the output, we used R to further analyze the output data. Basic purpose of this and similar kind of analysis is that it can be used in linguistics analysis.

Conclusion

After the analysis, we were successfully able to find the top recommended books from the dataset, also we were able to research deeper on the Information Extraction techniques. We also concluded that average helpfulness of the reviews on the books provided by customers on Amazon is accurate and can be trusted. Also, our analysis proved that sentiments of people writing the reviews are usually on the scale of 0 to 5 where the mean of the statistics rounded off around 1. All these analysis, though already done, can be used by ecommerce websites and plays a very important role in increasing the business and customer base.