Project Milestones
Part 1: Project Description

By

Team One

Joseph Patten
Md Mahedi Hasan
Md Muhtasim Billah
Scott Walters

1. Problem Statement

Our primary project objective is to replicate a search engine on a set of Amazon products. This application will take user search terms and present appropriate items matching the search. Our (tentative) secondary project objective is to look at co-purchasing patterns among Amazon customers.

a. Input and output

The data that we will use is the Amazon Review Data from a 2018 dataset. Given the metadata on the products and the reviews given by users, we think that this input will assist us in generating meaningful search results. For our initial objective, the input will be the search query, and the output will be the query results. For our co-purchasing analysis, we would input a co-purchasing pattern, and identify customers (as output) that fit that pattern.

b. Importance of the problem and the application

Querying from a big dataset is challenging since the dataset is so large. The first (and obvious) application is that this would allow a consumer to quickly query items using searchable attributes like Amazon allows, as well as non-searchable attributes (like number of reviews, average rating, etc). We could then use this application to build our co-purchasing analysis. From a larger perspective this project is coupled with the importance of discovering ways to interpret large amounts of user and consumer data and finding relations or similarities between that data. Companies are faced with using the data they generate to provide better services and this project replicates that.

c. Difficulties to deal and the challenges

As stated in the previous section, querying the dataset presents a big challenge as the dataset is so large. Thus, we will have to use a distributive algorithm to quickly return results of the query. For the co-purchasing part, there are a lot of issues we will have to deal with, namely the sparsity of the co-purchasing data (consumers buy a few items out of millions of possible items). Thus, coming up with an efficient algorithm that satisfies our goal is paramount. Cleaning the data will also be a hurdle that will require us to learn tools and techniques for parsing the data.

d. Goal to achieve

Goal 1: Build an efficient querying algorithm to quickly query our large dataset. We plan on using existing distributive algorithms to do so, as this is a task that has been done to great efficiency before.
Goal 2: Analyze co-purchasing data using a custom algorithm that is based on computer science, economics, and statistics research.

e. Team profile

The team consists of 4 students from different backgrounds who are majoring in economics, statistics, mechanical engineering and computer science.

2. Datasets and tools

a. Link and description of the data

This Dataset is an updated version of the Amazon review dataset (2018). As in the previous version, this dataset included reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs). In addition, this version provides the following features: more reviews, more categories and product metadata. The data has been requested and collected from the following link.
https://nijianmo.github.io/amazon/index.html#code

b. Tools for parsing the data and data models

As we don't want to waste valuable resources and time, we will implement our initial ideas on how to implement our co-purchasing analysis in Python (pandas). Once we have a good idea of how we will implement things, we will use Apache Spark through Python.

c. Progress report:

We have started playing around with a very small subset of the data using Python (pandas) in order to better understand the dataset and also to see what would be possible. We have also been doing a lot of research on how to work with the co-purchasing data efficiently in order to output relevant and interesting results.

3. Tentative timetable for the milestones

Sept 20 - Submit proposal
Sept 27 - Having sufficiently played around with subsets of the data in Python, create an Amazon EC2 instance and start learning how to use it.
Oct 11 - Finish with initial related work review to give us ideas of how to implement both the query and co-purchasing subprojects.
Oct 7 - Have the data in the EC2 instance.

Oct 18 - Having the data in EC2 instance, using Apache Spark, come up with basic summary stats and EDA. We will also have a better idea on what to do (and a plan to go forward) with the co-purchasing analysis.
Oct 31 - Have the distributive query algorithm design complete.
Nov 8 - Have the distributive co-purchasing algorithm design complete.
Nov 14 - Implement both algorithms to our satisfaction in our EC2 instance.
Nov 15 - Submit the algorithm design for sequential and distributive algorithms (for both the query and co-purchasing parts).
Dec 13 - Finish experimental study/demo and submit