CPSC 4780/6780 Course Project

Introduction

The course project is to build a recommender system using Amazon data. You may also come up with any research topic/problem that you think is interesting by using the available datasets below or other datasets you collected. Talk with the instructor for approval first if you would like to propose your own project, including the datasets you selected.

Amazon Recommender System

Data: http://snap.stanford.edu/data/amazon-meta.html

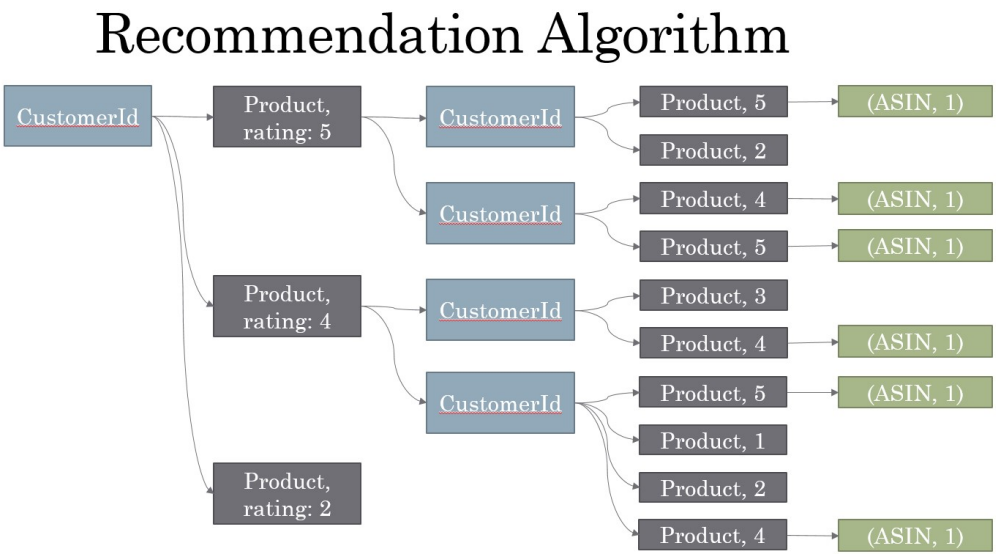
The data was collected by crawling Amazon website and contains product metadata and review information about 548,552 different products (Books, music CDs, DVDs and VHS video tapes).

Suggested Problem: Implement a data analytics engine that has the following functions:

1. Search.
   The search engine can answer simple and complex queries. Searchable attributes include but not limited to:
   - Best *n* sellers of a certain category
   - Number of reviews of a product
   - Queries with enriched operators: >, >=, =, <, <=; e.g., Select movie with average rating >=4.5
   - The number of customers co-purchasing same product of a user.
2. Recommender

   Analyze the co-purchasing patterns in the data. Given a customer, recommend products that the customer are most likely interested.

   The following recommendation algorithm was developed by two students (Angelina Gerbas & Michal Cudzich) in the past terms. This is just to illustrate how they approached this problem with their own understanding of the data and some simple reasoning. This is not the optimal solution but provide you some ideas for this project.



You can either use simple heuristics developed by yourself like those two students in the example above or use existing machine learning algorithms for collaborative filtering, clustering, or classifications.

Requirements

1. Examine dataset, formulate your problem, and review related work.
2. Prepare data collection and formatting. You may need to write a parser to extract the information you need to the data structure/platform you will be using – MapReduce or Spark. Most students choose Spark in the past.
3. Develop an algorithm.
4. Run tests to justify your methods.
5. You can either work alone or with one other students. 3 students a group maximum.

Bonus (5 points)

Bonus points will be given if one or both of the following two options is used in the project:

1. Projects that's launched and run on Cloud (Google, Amazon, Microsoft, etc.).
2. Research and use a third party data-mining library.

Deliverable-- Due April 28th end of the day

1. Code
2. Demo Video – record a video using screen recording software such as "screencasify", "screencast-o-matic", "zoom recording".
3. Project Report. The report should include introduction, related work, approach, results, and conclusion.