Recommender Systems for Amazon

Data Science Institute, Columbia University

Instructor: Pappu, Vijay Applied Machine Learning Fall 2021

Group 29: Gupta, Parth Li, Ke Tyagi, Vishweshwar Wu, Xuanhao Zhang, Senqi

Introduction

Recommendation Systems have become a crucial factor in driving revenues of tech giants such as Amazon and Netflix. They help reduce the cost of finding and selecting items in an online shopping environment by accurately predicting whether a particular user would prefer an item or not. In this project, we set forward to build a recommendation system of our own.

Dataset

We will use user reviews and product metadata for various categories listed under the publicly available Amazon Product Dataset[1], which spans from May 1996 to October 2018 and includes 233.1 million reviews and 9.4 million products distributed over 29 main categories[2]. The review dataset consists of a reviewer's unique ID, their name, helpfulness of their review, review statement, overall rating given, and ID of the product being reviewed. The metadata consists of a product's unique ID, its description, price, sales-rank, brand name, subcategories, and lists of related products which are generally bought or viewed along with it.

Problem Statement

Given the users' reviews and product metadata, we will build a recommendation system that predicts, as accurately as possible, the lists of products related to a particular product that has a high probability of being bought or viewed with it.

Proposed Techniques

We will carry out extensive data exploration and visualization, followed by data cleaning and pre-processing. We will then incorporate item-based Collaborative Filtering techniques to create recommendations based on users' reviews which will link a given item to other items of similar kind, which are highly likely to be viewed or purchased by most users. To provide accurate recommendations, we will leverage both the users' reviews as well as products' metadata. We may select some keywords from each textual metadata feature, combine those features into one column and perform CountVectorizer or TF-IDF methods to convert the textual features into numerical columns in order to construct the profile table for all products. We will try to explore Neural Collaborative Filtering[3] in the context of this problem. We will use different similarity metrics like cosine similarity, Jaccard similarity, etc., to compute the similarity between two product vectors. After building the similarity matrix, we will select top-k related products as the recommendations based on the top-k highest scores.

References

[1] Amazon Review Data - Jianmo Ni, UCSD

[2] Justifying recommendations using distantly-labeled reviews and fined-grained aspects. Jianmo Ni, Jiacheng Li, Julian McAuley. Empirical Methods in Natural Language Processing (EMNLP), 2019

[3] Neural Collaborative Filtering. X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T. Chua. WWW. 2017.