Building a Hybrid Product Recommendation System for Amazon using Customer Ratings and Reviews (Sentiment)

Introduction:

Amazon is a multinational technology company based in Seattle, Washington, that focuses on e-commerce, cloud computing, digital streaming, artificial intelligence and more. One of its most popular services is the e-commerce website 'Amazon.com.' The website offers a plethora of different types of products ranging from automotive to sports to groceries to hardware tools to digital products and more. A shopper can purchase almost anything on Amazon.com.

One of the ways Amazon keeps its shoppers engaged is by way of its product recommendation system. The product recommendation system uses content and collaborative based filtering, looking at a shopper's past purchases and purchases by shoppers who purchased similar items and then recommends items that the shopper may like based off of these two behaviors. Collaborative filtering enables the shopper to see items from other categories, while content based filtering allows shoppers to see domain specific items.

According to Neil DeGrasse Tyson, "If you track what I buy at a store, and then send me coupons based on what you think I'm going to buy next based on what I bought before. You have denied me the chance of stumbling upon something that I never thought of buying and that takes away my freedoms and I don't like that. It's the art of browsing. If I have to look up this word in the dictionary and get through 6 other words. You learn other words in route to the word that you're targeting." Based on these comments, which are similar to how many people feel when they shop or look at content on the internet, we want to build a hybrid recommendation engine that will not only recommend similar products but also recommend products in other categories, genres or fields to a shopper in order to help them find what they might not have been looking for.

Problem Statement:

"If you know my previous habits, you're assuming I'm going to stay that way for the rest of my life. You're trying to channel me into buying a product. I want to experience this world by stepping where I've never stepped before and buying something I never thought of buying."

- *Neil Degrasse Tyson*

Recommendation systems are great at recommending similar products to shoppers, but they tend to be too concentrated, which does not let a shopper go outside their comfort zone or see something they may not have known they wanted. For example, watching a video on

Youtube will start to flood a user's browser with several similar videos and not enough diversity in content. As for the case with Amazon, we want to build a recommendation system that recommends a diverse selection of items or content to users as needed.

The Client:

People who like to shop online, stream or read content, support their favorite charities or try out certain products before they buy are our targeted clients. Essentially users of Amazon.com or Amazon Prime members. These people like the randomness of what shopping online may entail. The randomness of it enriches their lives. They are looking for certain products but love the idea of finding items or content they weren't looking for or didn't know they may have needed. Our clients know what they want but know they won't be the same person tomorrow as they are today.

Data Analysis:

The Amazon Customer Reviews Dataset is a robust set of tsv files that house over a hundred million reviews that express opinions and experiences of shoppers regarding products on the Amazon.com website. This makes Amazon Customer Reviews a rich source of information for Natural Language Processing (NLP), Information Retrieval (IR) and Machine Learning (ML). Reviews range from 1995 to 2015 and are compiled from customers in 5 countries.

Most of the tsv files only contain reviews for products in one category (toys, software, shoes, lawn and garden, etc.), but there are millions of reviews in each tsv file. We are going to be using a multilingual reviews dataset (US) which contains multiple categories and several million reviews.

The dataset contains 6.9 million+ reviews on a 5-star rating scale from 4 million+ different customers. There are 86,000+ products in 16,000+ product parents ranging in 39 different categories from August 8th, 1995 to August 31, 2015. Some of the categories are mobile apps, digital ebooks, home entertainment, video games, home improvement, automotive, software, beauty and much much more.

Solution and Approach:

We will be making a hybrid recommendation engine that will take in content and collaborative based filtering in order to recommend a diverse selection of products or content that a user may like. Some recommendations will be very similar to the user's past purchases while others will be from other categories, genres or domains. This is to help give the user a look at various items they may not have known they wanted.

Deliverables:
Jupyter notebook for streamlining the analysis and prediction model and report with slide deck for a quick overview. Also a recommendation engine as well.

Source:
Amazon Customer Reviews Dataset Readme
Amazon Customer Reviews Dataset Downloads
Search for Amazon Products Using Asin Number
Neil Degrasse Tyson Speaks on Shopping