Name of your project

Sergio Kazatzidis
sergio@example.com
Kamiel Fokkink
kamielfokkink@gmail.com

Baran İşcanlı
barantevitol@gmail.com
Tomas Kehus
tekehus@gmail.com

Abstract

This is where we write our abstract

1 Introduction

2 Dataset

The dataset that we used for the training of our model was Amazon review data, which was retrieved from here. This database contains review information about numerous kinds of products, ranging from groceries to clothes to video games. But reviews about for example kitchen appliances would say more about the quality of the product, and not about a consumers individual tastes. The most suitable kinds of products to train a recommender system on are those of which the review reflects the particular persons interests and tastes. Therefore, we chose to use the datasets of Kindle books, video games, and digital music. The database offers two options for the dataset: all aggregated reviews, or a filtered dataset to only include reviewers that rated 5 or more products. We chose the second option, because it reduces the cost of training, and it is more relevant to have several datapoints per user, so as to have more information for good recommendations.

2.1 Splitting the data

The sizes of the datasets are: 31MB and 64k lines for the music, 110MB and 231k lines for the video games, and 266MB and 982k lines for the books. For each review, there are several features that we can include. Some are necessary for our analysis, such as the reviewer ID, product ID, and the review score. Many features are also less relevant for our purposes, but can be included if needed, such as the review time or a review text. A printout sample

	reviewerID	asin	reviewText	overall
0	A3EBHHCZO6V2A4	5555991584	It's hard to believe "Memory of Trees" came ou...	5.0
1	AZPWAXJG9OJXV	5555991584	A clasically-styled and introverted album, Mem...	5.0
2	A38IRL0X2T4DPF	5555991584	I never thought Enya would reach the sublime h...	5.0
3	A22IK3I6U76GX0	5555991584	This is the third review of an irish album I w...	5.0
4	A1AISPOIIHTHXX	5555991584	Enya, despite being a successful recording art...	4.0
5	A2P49WD75WHAG5	5555991584	Who knows why I initially considered this to b...	5.0
6	A3O90G1D7I5EGG	5555991584	Enya is one of a few artists whom I consider s...	3.0
7	A3EJYJC25OJVKK	5555991584	Enya is one of of the most mysterious singers ...	5.0
8	A1DA8VOH9NR6C7	5555991584	This is not another lousy Celtic New Age album...	5.0
9	A33TRNCQK4IUO7	5555991584	Many times, AND WITH GOOD REASON, the "new age...	5.0
10	AWY3EPKEOUV1W	5555991584	I just recently purchased her "Paint The Sky ...	5.0
11	A1SCJWCMQ3W3KK	5555991584	Over the past twenty-odd years, Enya Brennan h...	4.0

Figure 1: Pandas dataframe of the data

of the data can be seen in figure 1.

For testing purposes, we want to split our data into a training and a testing set. This is not just a simple randomized split of the datapoints. To be able to evaluate our recommender system, we need to split based on users, taking out 10% of all users and putting them in a test set. This way, our test set consists of completely new users who have not been seen during training yet. By looking at a few of those new user's preferences, the model can then recommend them new items to buy.