Mücahit Furkan Yıldız 21400425
Kerem Ayöz 21501569
Group 18

Term Project Proposal

In our term Project named as Retro Movie Recommender Project, we will try to develop a program for recommending a movie to a person. Movie enthusiasts always search for a good movie to watch, they search it for hours and sometimes their choice might be bad. A system which could give instant and accurate advices would be really helpful to that kind of people. As a movie enthusiasts, we thought that developing a program would be better since we have a domain knowledge about the topic that we have chosen. For instance, we could decide which data is valuable when recommending a movie to a specific person. Therefore, the theme of the program is suitable for us. There are many approaches to develop recommendation systems however we will use the learning algorithms with the movie dataset that we found. We found different datasets with different attributes hence the learning algorithm that we will develop is not decided exactly.

We found 2 methods for deciding what will be the input of our program. In first method, we will use the first dataset which contains the information about the user. We will use that information to use features of the user that program will recommend a movie. These features might be gender of user, age of user and job of user. We may determine the quality of the recommendation by comparing the ratings of users with similar features. In the second method, we will use the second dataset which does not contain any information about users. In that method, we may want user to choose a category and tag or to select his 3 different movies to understand the taste of user that we will recommend a movie. These selections might become our inputs in our model if we use this method. Thus, the learning algorithms that we will choose affect the selection of the inputs, labelling and evaluation of the data in different ways.

As a group, we searched over some data sets related to our task and reduce the data sets to two. They are both collected from the MovieLens web site[] by GroupLens Research[]. First data set[] contains three different data groups which are ratings file, users file, movies file. Rating file contains user ids with the range from 1 to 6,040, movie ids with the range from 1 to 3,952 and the rating of the users for movies. It is guaranteed that each user vote for 20 movies. User file contains same user ids from rating file and gender, occupation and age information of the users. Movies file contains the same movie ids from the ratings file, titles of the movies and their genres. Totally it contains 1,000,029 ratings. Second data set is bigger than the first one in terms of the number of ratings. It includes again three files. Movies file and ratings file have the same structure with the files in the first data set but they are different in numbers. It contains 20,000,263 ratings across 27,278 movies by 138,493 users. However the third file is not users file, it is tags file and it includes the tags of movies given by users which are generally a single word or a short phrase.

The first data set is relatively smaller than the second one. However, they both include different attributes and different type of information and they could shape the algorithms that we are going to use in our projects. Powerful side of the first data set is users file. Age, gender, and occupations are important indicators for the movies which are interested by users. Some movies could be popular between some age range or maybe some movies directly connected with different occupation types or some movies could be chosen by the users only because of their genders. We can use them as input parameters with weights calculated by algorithms. The strong side of the second data set is tags file. At first sight, users judge the movies by tags and categories. Tags give us the most general information of the film by the point of the users. Therefore tags are also important in terms of recommendation. We can use tags file by labeling tags as positive, neutral or negative and train our algorithm with this additional information. All positive tagged film would be a good suggestion for a new audience.

In spite of the powerful sides of the data sets, there are difficulties we will encounter. For instance, one movie can be highly gender-related that even the other gender doesn't watch it and doesn't rate it or the movie can be in the intersection of subsections of age or occupations. We may have to find these critical points. Another difficulty is the weights of parameters. They can change dynamically with the society. For instance, teenagers prefer dramas today, however, maybe the same age range preferred sci-fi movies ten years ago. Therefore determining the weights may be very complex or even impossible. In addition to difficulties of the first data set, the second one has different types of difficulties such as labeling. Without labels, tags are just a word and algorithm cannot understand the meaning of them. We need to find a way to label them and there is a high amount of data to label. Even if labeling procedure ends up successfully, yet they will be a subjective judgment. We need to turn them into reliable information to use them. In both cases, the rating number of each user is different and judging capability of the users are different which may need to be considered. In addition to that there, maybe variations or inconveniences in between the rating of the same user. He or she may be in an unusual condition while watching the movie and these unpredictable situations affect the data. All of these can be seen as noises in the data and there are many possibilities to create noise.

We will use Python programming language to implement our project. The main reason for us to use Python is the available libraries to manipulate the data. Since the datasets contain also the irrelevant information, we need to separate the data that we will use while training and testing the algorithms. Python provides libraries such as pandas, numpyy, mathplotlib, xlwt, etc. for manipulating and visualizing the data. Also many resources and examples for Python are available on the Internet which will also support us while we are developing our project. We will use PyCharm programming environment to work in a more synchronized way. We will use Git as a version control system for our project while we are developing our project.