PROJECT PROPOSAL

Problem to solve

Nowadays, in this global world there is a vast number of things to select, where people might be lost. That is the case at the moment of choosing the next book to read. That's why this project proposal is a book recommender system to help people choose better among a huge set of options.

The provided solution will be valuable since a recommender system can be useful in different ways and also it can be replicated for other topics like music, movies, retail items, etc.

Goal and Milestones Project

Milestones:

- Install python libraries
- Data loading
- Data cleaning
- Data exploration
- Feature selection
- Cost function definition
- Model selection
- Model evaluation
- Conclusions
- Project Presentation

Goal:

- Provide an accurate book recommendation system

The data source

Originally, I found the data source at Kaggle website, but it was obsolete. The current version is in github and data source is composed by .csv files that I can download in a few seconds.

Dataset is composed by three main csv files:

Books.csv with metadata for each book. It includes book id, title, author, number of editions, publication year, language, average rating and others

Ratings.csv contains the most important variables for this project, user id, book id and rating for books. All these variables are integers, specifically the given rating is into a range from 1 to 5, being 1 for the lowest rates and 5 for the highest rates for a book.

To_read.csv, is the file that provides user IDs and the book IDs that the user has already read.

Solution Techniques

First at all, for data cleaning I will determine how to handle missing values and outliers based on what I will find into the dataset. I might disregard rows with missing values, or I might use interpolation or imputation. Same case for dealing with outliers, I might use winsorize or log transformation techniques.

Since there are no more than 25 variables, I am assuming that dimensionality reduction won’t be necessary. Anyway, I am considering PCA, if needed.

For feature selection I am planning to use random forest technique to select features based on their importance.

I will use mean squared error (MSE) as the cost function to have an idea of the goodness of the model.

For developing the recommender system, I am planning to use Restricted Boltzmann Machines, a common form of neural network with two layers, the first one is the inputs or visible nodes and, the second one is the hidden layer which is the new set of derived features.

Once the recommender system is created, the product could be deployed as a mobile or web application, where user will introduce her/his reading story and then receive recommendations.

Challenges

The biggest challenge I’ll probably face, it is going to be the fact I have never done a similar project before but there is a good documentation outside, books, serious blogs and magazines that I can use to complete my project successfully.