AML-3204 Social Media Analytics

Project (40% of the course grade)

Presentation: During the Lecture Hour

Project Title: Comparing Collaborative filtering based recommender and Hybrid (collaborative plus content) recommender system

In this project, you will compare collaborative filtering-based recommender and hybrid recommender systems. To build the collaborative filtering-based recommender, you will use the Matrix Factorization technique, and to build the hybrid recommender, you will use the Neural Embedding layer from the PyTorch package.

Dataset Description:

The dataset is from the MovieLens dataset collected by the GroupLens Research Project at the University of Minnesota.

This data set consists of:

* 100,000 ratings (1-5) from 943 users on 1682 movies.

* Each user has rated at least 20 movies.

Here are brief descriptions of the data:

u.data -- The full 'u' data set, 100000 ratings by 943 users on 1682 items. Each user has rated at least 20 movies. Users and items are numbered consecutively from 1. The data is randomly ordered. This is a tab-separated list of

user id | item id | rating | timestamp.

The timestamps are Unix seconds since 1/1/1970 UTC

u.item -- Information about the items (movies); this is a tab-separated list of

movie id | movie title | release date | video release date |

IMDb URL | unknown | Action | Adventure | Animation |

Children's | Comedy | Crime | Documentary | Drama | Fantasy |

Film-Noir | Horror | Musical | Mystery | Romance | Sci-Fi |

Thriller | War | Western |

The last 19 fields are the genres. A 1 indicates the movie is of that genre, a 0 indicates it is not; movies can be in several genres at once. The movie ids are the ones used in the u.data data set.

u.genre -- A list of the genres.

Tweets from Twitter:

For each movie, make a tagword using the movie name. Then, search Twitter using that tagword to download related tweets for that movie. Do the same for each movie. Clean the tweets, and then calculate the sentiment score for each tweet for each movie. Calculate the average sentiment score for each movie and save the average as the sentiment score for that movie.

Building the Collaborative filtering-based recommender system:

Algorithm to be used: Matrix factorization

Dataset to be used: u.data

Building the Hybrid Recommender System

To build the Hybrid Recommender System, we are combining the following:

i) Movie Ratings from **u.data**
ii) Movie genre information from **u.item**
iii) Movie sentiment from calculated from Twitter

To combine, you need to use the Pytorch embedding layer. The following article will help you implement the embedding layer using the Pytorch:

https://www.ethanrosenthal.com/2017/06/20/matrix-factorization-in-pytorch/

To better understand how the neural embedding layer works, please read the attached paper titled:

"Neural Collaborative Filtering [1]"

Please upload the following to Moodle by the deadline:

1. Project Report **as a PDF file**. (Please **precisely** follow the template given below, IEEE format is fine)
2. Python files
3. Presentation Slides **as a PDF file.** (During the presentation, you want to describe the methodology, dataset, and experiment results. The duration of the presentation will be **25** minutes max)

To complete the project, complete the following template:

Abstract: Here, using no more than 160 words, provide a summary of your work, and it should include your findings as well.

1. Introduction:

//Here, you want to write about the different recommender systems available. You need to describe your findings here as well. Please follow the Introduction section in the given paper [1].

2. Related Study:

//Here, you want to summarize at least **ten** papers related to recommender systems that use Collaborative filtering and Hybrid Recommendation Techniques. Please format this section similar to section **5** of the attached paper [1]. You have to do in-text citation for the referred papers as I have done here [1]. Also, their

references have to be included in the **References** section. Please follow the APA style of referencing. For example, use the style that I have used to refer to a paper in the **References** section.

3. Methodology:

3.1 The methodology section starts with an introduction to the methodologies you are applying in this experiment.

3.2 Insert a system architecture diagram for the Hybrid Recommender system. The diagram should include the workflow (from data collection to predicted rating generation) to design the Hybrid Recommender System.

3.3 Here you want to talk about the datasets that you are using. Note that here you do not want to describe the quantifiable Information (such as the total number of ratings, the total number of tweets, etc.) of the data

3.4 Write about Matrix Factorization and describe the matrix factorization algorithm along with the Matrix Factorization Equation. Do not forget to describe the different components of the equation
Insert a simple diagram here. The diagram should show how the latent features are generated using Matrix Factorization

3.5 Write about Neural Embedding Layer
Insert a diagram here. The diagram should show how the latent features are generated using the embedding layers

3.6 Write about Sentiment Score Calculation: Here, you want to talk about the sentiment score generation tool, which results from the tool you have used to calculate sentiment score, etc.

3.7 Here, you want to talk about you the activation function that you have used to build the hybrid recommender system

4. Experiments

Put an introduction using few lines about the structure of this section

4.1 Experiment Design: Here, you want to write down the steps you have taken to design the experiment. Please include the machine configuration, different APIs (such as APIs for Twitter, Vader, Pytorch, etc.)

4.2 Dataset Preparation: Here, you want to write the steps you have taken to prepare the datasets. You have to include the quantifiable Information (such as average tweets per movie, average ratings per movie, the highest number of ratings for a movie, the lowest number of ratings for a movie) of the datasets used in this experiment.

4.3 Evaluation Metrics: Here, you want to write about the evaluation metrics used in this experiment. You have to calculate RMSE, MAE, Precision, Recall, and F1-Score for the datasets. So, write about RMSE, MAE, Precision, Recall, and F1-Score. Include the formula to calculate each of the metrics

4.4 Results and Analyses

4.4.1 Hyperparameter Selection: Here, you need to run the experiment to select the different hyperparameters such as latent vector size for Matrix factorization and Embedding based Neural Networks (Pytorch) experiment. For both models, run the experiment using 20, 30, 40, and 50 as latent vector sizes. For each run, save the RMSE.

Include 2 line diagrams here:

One line diagram should show the RMSEs for the Matrix Factorization model using different latent vector sizes.

Another line diagram should show the RMSEs for the Neural Network model using different latent vector sizes.

Then choose the latent vector size that gives you the lowest RMSE

4.4.2 Results of Recommendation Models: Here, compare and discuss the results (RMSE, MAE, Precision, Recall, and F1-Score for top-10 and top-20 recommendations) for the two recommendation models using bar/line diagrams.

5. Conclusion: Conclude the experiment here.

References:

[1] He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. (2017, April). Neural collaborative filtering. In Proceedings of the 26th international conference on world wide web (pp. 173-182).