Team name: Deep Blue Deep Learning

Team members: Yunzhou Liu, Chuanrui Liu, HaoyuZhang, Dingyuan Liu, Yiyao Qu, Yumeng Zhang

Date: October 21, 2020 & October 23, 2020

(This journal has to be uploaded to your Sakai group folder by 11:59 pm every Friday and your team webpage till and including October 30). Change the file name to "Team_name_MM-DD-2020" where MM = month and DD= day of upload and then upload to your Sakai team folder).

Team roles for this week (write down name):

Facilitator(s): Haoyu Zhang

Recorder(s): Chuanrui Liu, Yunzhou Liu

Deliverer(s): Yiyao Qu, Dingyuan Liu

Planner(s): Yumeng Zhang

See last page for description of roles. Obviously one person can take more than one role or there can be more than one person per role or make your own roles!

> **0. Describe briefly what the main goal of your team is (so the peer reviewer has some context). E.g. we are working on image classification for blah de blah. Our goal is blah de blah etc. In the initial part of the semester before your proposal it is ok to put down "we are still coming up with ideas on team project".**

Our team decided to focus on building a recommendation system for a user-based onlineservice. For example, we are looking at the dataset provided by a film review websiteMovielens, which contains the movies' genre, titles, the ratings given by individual users, etc.Our goal is to analyze the users' interactions with the interface, clustering the movies or users to build a well-defined recommendation system.

In our recommendation system, we aim to identify different situations where user ratings could be a useful variable to implement movie recommendations and where user ratings are

less useful to predict other movies. Instead of taking user ratings into account, we hope to recognize the user ratings extremely low or high to better understand what the users really want to see and give accurate feedback back to the users. It's important to understand our users' preferences and the similarities between the movies in order to provide useful suggestions to the user and improve the service accordingly.

I. What was done this week regarding the project: If you want to include code include this in the Appendix. Describe what the group did (including contributions of individual team members) with regards to the group project this week. Give enough details so I understand what you folks have been doing over the week. Include dates of your meeting(s) and who met on these days.

Last Monday, our group met with the Ph.D. student Lain D Carmichael to look for inspiration and solutions for our obstacles. Lain suggested that we can try preprocessing the data or include structured texts instead of single unrelated keywords to improve our current model's performance. He also recommended other algorithms to check out and try, including glove, transformer, and pre-trained word2vec. After the meeting, we divided our work to search for new resources on these topics and improve the model performance and build up a cohesive recommendation system based on different measurements, including ratings, popularity, and texts.

Yunzhou Liu: Applied our previous algorithm to a new dataset with descriptions. We observe that different models can easily understand the descriptions, and the result is significantly better than our previous run on the "tags + genre" combo.

Dingyuan Liu: Followed the K-means idea proposed by Yiyao Qu and Yumeng Zhang. We went through setting up functions finding the most appropriate k for clustering, and tried clustering based on two categories of films. Later we tried to utilize k-means in a higher dimension by setting up the data so that each row shows different ratings from different users. Each column shows a user's ratings of all the top 1000 most rated movies. However, we cannot apply k-means in a sparse dataset with lots of NaN, and we went to Weibin's office-hour during for suggestions.

Yiyao Qu & Yumeng Zhang: searched online resources for K-means and K-NN; to prepare alternative solutions in case our current approach is not valid.

Haoyu Zhang & Chuanrui Liu : We worked on our team assignment question 1 and combined the team's efforts after discussing the results. Since preprocessing data plays a very important role in our text mining model, we wondered whether another dataset with structured texts could perform well. We found another movie dataset on Kaggle with detailed descriptions of

each movie and popularity, actors, and directors. We decided to search for ways to use a single vector representing the vectors' collection for each individual vector. We have explored some questions regarding the weights of tags and the relevant implementation. Also, we together did some research on the concept of multivariate median that the Ph.D. student has suggested to our team. In this way, we considered and evaluated the usability of this concept in our project since we are having some issues with the criteria of the similarity during our project implementation.

II. What were obstacles faced if any in working on the project? This could be technical (like not being able to implement or understand particular techniques) or time issues (midterms for other courses etc).

Yunzhou Liu: Though the model runs well on the descriptions, we cannot apply this algorithm to our original dataset because we only have about 5k movies with description. In contrast, we had 10k+ original movies in our dataset. Hence, many movies do not have descriptions, so we cannot apply this method to them. Furthermore, we observe that running the entire dataset algorithm is too slow (and even crushed my computer during calculation). Due to these two reasons, we decide to still try and build another bag-of-words model for the tags and measure the cosine similarity between them. If a movie had a description, then we would combine the two results by averaging them or a max-pooling to achieve the final result; otherwise, we would solely rely on the tags similarity to obtain the result.

Dingyuan Liu: The main obstacle we met when trying to cluster by users' rating is that the dataset contains too many NaN values because most users will not watch and rate the films. Also, we consider that it is not appropriate to simply delete those Nans. K-means cannot deal with sparse dataset. The converting methods we found so far did not provide a nice solution, but Weibin's suggestion of looking into matrix completion offers new directions.

Haoyu Zhang & Chuanrui Liu: The obstacles we encounter during this project period include that we are not so sure about the execution of the concepts of tag weights and multivariate median in reality and how the execution would contribute to the results. Though the two possible ideas could help us solve some of the difficulties by making the linkage of similar films to make more sense, we have not decided how to implement these ideas.

III. What is the plan for the next week including what each team member is planning to work on in the next week. Describe goals and potential timelines (" I plan to finish understanding x to see if it can be implemented for our project by Wednesday etc".)

We plan to have our next regular project meeting this weekend and the following Monday. Our data and modeling choices have given us some great insights in our data. We decided to continue with what we have done to proceed with an appropriate conclusion or explanation and construct a comprehensive final report. Specifically, Yunzhou Liu will continue using text mining to find the most similar movies for a given movie. At the same time, Dingyuan Liu, Yiyao Qu, and Yumeng Zhang will try clustering on user's ratings to recommend movies watched by similar users. Chuanrui Liu and Haoyu Zhang will look for ways to assign weights and combine the two results with appropriate final recommendations.

Yiyao Qu & Yumeng Zhang & Dingyuan Liu: We plan to solve the Nah values in our dataset. We have collected several research papers about matrix completion and prepared to go through them for valid programming methodologies.

Chuanrui Liu & Haoyu Zhang&Liu Yunzhou: We plan to apply the multivariate median concepts to the different techniques we have tried. Trying to incorporate more measurements in considering what to recommend. We also need to transform our current demo to a more comprehensive project, using all the dataset and managing the results in a more presentable format.

While in the weekly document above you will describe what your team did with regards to the team project (with proper attributions of who did what in the week) there are 4 pre-defined roles. I urge you to have different people do these jobs every week so that you gain experience in each of the jobs. There can also be more than one person per job for example 2 people recording the weekly journal.

Facilitator: Manages the group for this week including setting up times for group members to meet, making sure everyone has a say in the meetings etc.

Recorder: Person in charge of recording the meetings as well as the happenings of the week and describing what was accomplished in the meeting and writing up this report.

Deliverer: Person in charge of checking the entire journal and uploading the doc file to Sakai in the group folder as well as the representative of the group getting in touch with the instructor.

Planner: Person in charge of what will be happening next week as well as thinking about longer term goals (what more needs to be done for the project).

Team contact: Person I can email if I see any issues in the weekly log instead of mass spamming everyone in the team.