Team name: Deep Blue Deep Learning

Team members: Yunzhou Liu, Chuanrui Liu, Haoyu

Zhang, Dingyuan Liu, Yiyao Qu, Yumeng Zhang

Date: October 6, 2020 & October 9, 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, Dingyuan Liu**

Recorder(s): **Yumeng Zhang**

Deliverer(s): **Yiyao Qu**

Planner(s): **Yunzhou Liu, Chuanrui Liu**

Team contact: yiyao@live.unc.edu

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 online service. For example, we are looking at the dataset provided by a film review website Movielens, 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.

This week, our group has split up into 3 small groups to contribute to the team project with different tasks. During this process, the group members have an efficient discussion within the small groups and give useful feedback to the whole group, which makes our exploration consistent. At the end of this week, we also schedule a meeting with IA to discuss some of our problems. The discussion is intuitive, and we end up having more ideas about how to pre-process the dataset to adjust weights and simplify the modeling process.

Yunzhou Liu & Chuanrui Liu: We tested different models for transforming the features and measured similarity between transformed sentences and measured the similarity between raw tokenized sentences. We mainly used Python for the project, and the packages we utilized include transformers, sklearn.metrics. Pairwise, seaborn, and so on. We also researched the paper with the relevant topic that we are trying to utilize in our implementation and try to figure out the possible approach for us to deal with the obstacles we had faced during the model testing process. After testing each model, we implemented some comparison studies on different models to try to identify the better models we might consider in the future.

Haoyu Zhang & Dingyuan Liu: Starting from the first week, we begin cleaning the dataset and arranging it to the form we need to perform similarity modeling. The data on the websites are separated into different files; each has some useful information about films. We use techniques like group by, mapping, and join to extract the features: title, genres, time the film came out, keywords, average ratings to a new dataset. To arrange the data for future use, we combine all the features into one string with each feature separated by the space. We then used empty string to replace null values, deleted repeated words from combined text feature, and eliminated low_frequency words, eg. based_on_book, to further match the text requirement for document analysis.

Yumeng Zhang & Yiyao Qu: We tried to approach the cleaned data in another potential way. This will be our back-up plan if our approach to measure tokenized similarity failed. We would like to build a user-based model. Currently, we are using K-means clustering. Due to the size of the dataset, we took only Romance and Science fiction for efficiency. We will apply our algorithm to the whole dataset after we make the adjustment and finalize our model. We first calculated each user's average rating for Romance and Science Fiction movies. Then, we group the movies in these two genres using only the average rating.

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).

Haoyu Zhang & Dingyuan: Before any pre-processing, we observed that the dataset we found does not include the film's directors and descriptions. Without descriptions, it is hard to determine whether the result gives the true most-similar recommendation. Later we might want to web-scraping for more features. During data cleaning, we found that our newly created variable containing text from genre, cast, ratings, and descriptions has different lengths, and we are puzzled about whether to normalize the scale or not. Also, we did not agree about how weights should be assigned to each element of the combined text. For example, the keywords dataset contains more words than other features; the larger number of keywords may lead to a higher weight when considering the similarity. That explains why the films we found most similar in the demo belong to different genres. One possible solution

is to drop the low-frequency words in keywords. We already filtered out the low-frequency words with a threshold of 10. Another solution is to adjust the weights assigned to each feature beforehand. Some of the selected features like "time" were later found to have format problems and need further correction to be able to use.

Yunzhou Liu & Chuanrui Liu: Though all of the models transform sentences well and created the corresponding matrix & vectors, neither the transformed result nor the raw tokenized vectors appear to have "distinctable boundaries", i.e., the similarities between the features are too close for different categories of movies so that we cannot create a harsh boundary between the movies (i.e., we cannot give a similarity threshold that says "above this threshold are movies that are similar"), and we do not know if the similarity corresponds to the actual similarity (ground truth). After we research the similarity issue, we found that weight limitations on tags might be a possible approach to increase the accuracy of the similarity in the model results. However, we are currently not sure how to implement this idea in the actual model implementation.

Yumeng Zhang & Yiyao Qu: When we tried to clustering users into different groups, one obstacle is that it's hard to separate users based on existing data since there are no defined criteria for similarity and difference. We will have to explain, for instance, how many differences in those ratings will result in different clustering, subjectively by ourselves. Currently, the only thing we can do is to clustering users according to their watched movies and rate. However, with only movies as a classification standard, it is hard to find "similar users". Also, it is hard to choose what number k will work best for our purpose.

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 the next group meeting next Monday, and we will also meet with a Ph.D. student Lan D Carmichael who is specialized in text mining. We will continue working on two approaches: clustering on the user-side and clustering on films. We will also look for

techniques to speed up the modeling process and classify similar words into a single category beforehand.

Haoyu Zhang & Chuanrui Liu: We will first look through the work we have done so far and list the questions we need to ask during the scheduled meeting with the Ph.D. student Lan D Carmichael. Also, We will do research on how to adjust the weights of the texts when calculating similarity. We will try to find the proper approaches to combine the theory of weights of the texts and their application together with exploring whether this would improve our model. During the time, We will be in charge of communicating with different groups and manipulating the dataset as needed.

Yunzhou Liu & Dingyuan Liu: We will define the similarity and find a way to classify the keywords into meaningful categories before calculating the similarity distance and clustering, either through a designed algorithm or manual assignment.

Yumeng Zhang & Yiyao Qu: We will continue with the clustering model, using different criteria for similarity and seeing how the model will do. After clustering, we will check in person to see if those clusterings make sense.

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.