Team name: Deep Blue Deep Learning
Team members: Haoyu Zhang, Yumeng Zhang, Yiyao Qu, Chuanrui Liu, Yunzhou Liu, Dingyuan Liu
Date: Thursday, September 24, 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, Yumeng Zhang

Recorder(s): Yiyao Qu, Chuanrui Liu

Deliverer(s): Yunzhou Liu

Planner(s): Dingyuan 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!

I. 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.

II. 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.

For this week, we have started to research and explore possible ways to build our model.

One way we found focuses on the concept of cosine similarities. First, we should import the dataset and choose our preferred features. Then we use the function to calculate the cosine similarities between our user-like movies and other movies based on our preferred model. Thus, we can recommend movies according to the similarity values. However, this method may cause misinterpretation in exceptional cases.

We also found an alternate model that uses package BERT to measure the general similarity between sentences and phrases. In general, for each movie, we concatenate its name, genre, and casts into a long string, and run the Tokenizer and the BERT model on this transformed dataset to measure the similarity between sentences. This model would help us accomplish the goal that, when provided a movie that a user watched, it will find movies with the highest similarities to that movie and recommend it to the user.

In this case, we plan to split up the two situations where the movies' user ratings are involved in the similarity metrics and where the user ratings for the movies are not involved in the similarity metrics to help our recommendation system implement different goals. Also, we are considering using rating as the criterion to narrow down the similar movies' recommendation set to display a proper amount of movies for users.

III. 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).

We met two main obstacles. One is that every member was busy with one or two midterms these weeks, so we have limited time to work on the project. To resolve this, we split the work for everyone, so that we do not have to set a standard time for the team to work together.

The other is that though we found some techniques that could be applied to this project, not everyone is truly familiar with the rationale. Each member might have a different understanding of our task and what we should do next, especially at the beginning stage of the project. To familiarize ourselves with the common techniques we have researched, we might need to spend more time exploring the dataset to identify the appropriate techniques and algorithms to start with.

This week we decided that we could make use of cosine similarities as a standard to categorize the movies. To apply the algorithm, we need to pre-process the data. Some minor obstacles appear before we get to the serious work, such as combining multiple strings to a single string with different string lengths. To understand this, we searched for professional videos lecturing about this cosine similarities and how others are pre-processing their dataset to fix the problems. The next step would be applying the mathematics algorithm onto the data set and making adjustments that work best for our data.

IV. 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 clean and create a text-based variable in the dataset provided by a film review website Movielens. The new feature should contain information about file cast, genre description, plot tags, etc. We can use cosine similarities to predict predilections based similarities of existing text-based features. For the BERT method, we will search for pre-trained base coefficients and incorporate them into our model.

Also, we decided to set a meeting time fixed at 10 am Wednesday and Friday. Thus, everyone in the teams can catch up on projects in the middle of the week and wrap up at the end of the week.

Work distribution for every member as follows:

Yiyao Qu & Yumeng Zhang: clean and merge the dataset that we would like to use. Then create a text-based variable using the features.

Dingyuan Liu & Yunzhou Liu: research for some base model that could be used for context analysis in English and use the text-based feature to measure similarities between movies

Haoyu Zhang & Chuanrui Liu: research for some base model that could be used for context analysis in English and help assess if the work already done is reasonable.

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.