Team Centrosaurus Milestone 1 Report

State of the Project

Currently, we have achieved all of the goals we set for Milestone 1 in a form that offers function.[1] Much of the work for this milestone centered around the embedding model component and it's dataset. Currently we have developed a functional model that works with the goodbooks-10k dataset.[2] This model can take a book id and user id from this dataset and predict the rating with a mean square error of 0.67, which translates to within 0.82 stars of the actual 5 star rating. This was achieved after running the model for 25 epochs, and it appears that if we had run the model for more epochs we could have further driven down the mean square error. We will touch on this further below.

We have also examined additional datasets such as the Amazon Review dataset,[3] but we have not fully processed these as of yet. Integrating this dataset with the other dataset remains an issue which we will address below and in our work for Milestone 2

Proposal Changes

Most of the plans for the project remain consistent with the original proposal, however, we are reevaluating the Reinforcement Learning model and thinking of ways to successfully implement it into the project. For example, the proposal says that we will use the Gym library, however this will likely not be the case. We are still researching Reinforcement Learning and are searching for other tools that may work better in our situation. We also made a slight adjustment to the milestones in that the milestones have us working on the Web Application throughout the lifetime of the project, but we want to move this task to a later milestone (initial implementation in Milestone 3, and functional implementation in Milestone 4).

Current Challenges

As previously mentioned our model is achieving a test MSE loss of 0.67, after 25 epochs. This loss is consistent with the example embedding model from lecture 6, but while the model places the book on average within 1 star of the users rating, we would like to bring this down to within 0.5 stars or less (meaning an MSELoss of 0.25 or less). We have already implemented drop out and tested different embedding dimensions, but are still above our goal. To address this we plan to test further training parameters including looking into the use of other optimizers and schedulers.

On the issue of the Amazon dataset, we have found that in order to extract the title from the data provided, we need to scrape it from its corresponding amazon storefront. Doing so when processing may not make sense, as we would need to make thousands of requests to Amazon.com to do so. We believe the best approach for this dataset is to leave it in it's Amazon identifiers and only process it when recommending it to a user, reducing the quantity and

---

[1] We have not achieved all the goals set out in the proposal, as part of the reworking of our proposal as detailed in a later section, we changed the expectations for different milestones.

[2] [http://fastml.com/goodbooks-10k-a-new-dataset-for-book-recommendations/]

[3] [https://nijianmo.github.io/amazon/index.html]

frequency of requests to Amazon.com. This approach does raise concerns over whether the integration of multiple datasets is feasible, (or even necessary if we have a dataset of 54 million entries). These concerns will be explored and acted upon as we work towards Milestone 2.

Finally the last challenge we are facing is the implementation of Reinforcement / Online learning. Since our use of our RL is different to that examined in class, we have had to explore the area beyond the content covered. We have been researching different methods, and have come across a couple libraries that could help implement it, as well as a potential “hack” that would implement it without the use of a new library. Per the proposal this does not need to be implemented until Milestone 3 at the earliest so our approach to resolving this problem as we work towards Milestone 2 will be research and small scale testing of different methods

Team Member Breakdown

Byrne, Declan

Declan worked on the development of the embedding model, writing part of the model code, as well as testing different training parameters and addressing overfitting. He also put in some research into different approaches to reinforcement and online learning, though nothing has been implemented yet. Presently he is working to further decrease the loss the embedding model is producing, and will continue to do so as he works towards the next milestone.

Dhaliwal, Harvir

Harvir has worked on the development of the initial embedding model, which is currently working. This model may be improved for later milestones (e.g. optimizations). He is currently working on researching Reinforcement Learning tools and implementation strategies, so we can have a working Reinforcement Learning model for the next milestone.

Riley, Anna

Anna mainly worked on obtaining and processing the dataset(s). She tried out multiple datasets to see which ones would be feasible and helpful for the project. She then helped write the dataset module using the Goodreads dataset. Since a large dataset of Amazon book ratings was too large for Google Colab, she split the data into eight smaller csv files that would be small enough for Google Colab to process. She is currently working on finding a way to collate the Goodreads and Amazon datasets and translate the ASIN Amazon labels to book titles.

Xia, Ritchie

Ritchie helped with setting up the dataset. He helped extract data for the Goodbooks entries by using a separate lookup table to get the book titles and Goodbooks IDs which help by combining different editions of books. For the next milestone he will try and help extract the element on Amazon’s web page to get a book’s title and other information.