Team Centrosaurus Milestone 2 Report

State of the Project

Currently, our main goal is to test our current implementation of the Reinforcement Learning Model that we designed and implemented. The method that we have devised is quite simple and may work to some degree, however we are currently also researching other state of the art techniques that we could use to accomplish the task we need.

We are also currently working to optimize the embedding model that we implemented for Milestone to reduce loss. We have explored numerous different training hyper parameters (see training-notebook.ipynb), but our lowest loss has remained similar to that achieved in milestone one, around 0.67 mean square error. This represents a predicted rating that is within 1 star of the actual rating. While we would like to further drive this down, achieving this has proved difficult. Exploration of different hyper-parameters will continue as we work towards Milestone 3, however with less emphasis as during this milestone working period.

Regarding our dataset, we currently have a working dataset, from Milestone 1, that is working well with our model and implementation. Our team has also been experimenting with a second dataset that is much larger. However due to technical limitations (covered in proposal changes) we will not be working further with this dataset.

Proposal Changes

The only proposal change in this milestone has to do with which datasets we will be using. In the proposal we observed we may run into issues with some of our datasets, and may need to abandon them. We are doing this with the Amazon Dataset. This dataset did not provide a method to quickly retrieve book titles from the provided identifiers, so we had to use the ids to link to and scrape from the Amazon store front. We quickly realized that in order to provide new users with an accurate representation of books in the model when they signed up[1], we would need to retrieve all of the titles for the books in the dataset before launch, and not during use as we speculated we could. With a dataset of 51 million entries we would be dealing with a significantly large number of titles, and the slow down we would need to include in order to scrape from the Amazon website without interference means that it would be unfeasible to do so. Per the proposal back up plan section, we will be moving forward with the project using just the one dataset, and will be sure to implement Reinforcement Learning. Exploration of this dataset can be seen in: amazon_dataset_work.ipynb.

Current Challenges

There are two challenges we are currently facing: 1. Driving down loss in the embedding model, 2. Implementing Reinforcement Learning. For the former we will continue to examine ways to drive down loss, further testing different training parameters and methods of combating overfitting. There is also a possibility that predictions can be finer tuned by our solution to the

[1] So we could match them to an existing user

second problem. This problem being how we implement reinforcement learning. We currently have a “hack” of how to do this thought up which we will be testing in the coming milestone. We will also be researching how reinforcement learning is used in state of the art recommendation systems, and look to use similar methods in our project.

Team Member Breakdown

Byrne, Declan

Declan brainstormed possible solutions to the Reinforcement Learning model, and helped develop the solution Harvir is working on. He also spent time exploring ways to drive down loss in the embedding model, trying different hyper parameters for training. For Milestone 3 he will continue to try to drive down the loss of the embedding model, research different techniques for Reinforcement Learning, and possibly look at working on the backend integration of the app.

Dhaliwal, Harvir

Harvir worked on the development of the Reinforcement Learning Model that will be used to recommend books to new users. Harvir is currently testing and experimenting with the solution that we have designed for this problem. Specifically, he is trying to develop a loss function that may be used to train the “Reinforcement Learning” Model. However, this solution will likely not be used for the final product, so his current job is to also research different techniques that we can use for this task, while also testing with our proposed solution.

Riley, Anna

Anna worked on finding solutions to issues with the Amazon dataset by scraping ISBN numbers from amazon website using ASIN urls. She ran into problems with accessing the HTML of the amazon dataset, so after other team members tried to find a solution, we decided to stop working with this dataset. Anna researched effective ways of saving models and began working on the web app.

Xia, Ritchie

Ritchie worked on extracting book data from their asin number, but in the end the team decided to not use the Amazon dataset due to its large size. The code used is still available in the Amazon notebook and can extract the title of a book from its asin number, but he could not link together the Amazon and Goodbooks datasets since not every Amazon page includes an ISBN number. For the next milestone Ritchie will be working on developing the web app using React.