For the final project, your goal is to predict which subreddit certain “posts” origin from. This is a supervised classification problem. I have many recommendations and will offer some guidance, however many of the tools you have already learned this semester. Below I have my suggestions on where to start.

1. Familiarize yourself with Reddit. Mainly figure out terms such as “Subreddit” and “post”.
2. Download the data from our Kaggle.
3. Familiarize yourself with the embedding code which is on our canvas page.
4. Embed the posts. Consider this embedding as a set of 512 features.
5. Use XGboost for classification. We haven’t used xgboost for classification before, so use the documentation to find the correct objective etc.
6. (this is really step 5 but you’ll run into this problem after reading documentation) XGboost requires the data in a very specific format, namely the class of the train examples need to be changed. Do this in a feature file then return to the previous step.
7. After you have successfully been able to do classification, you’ll notice that the prediction accuracy is not great. Even though xgboost has an inherent feature selection, with 512 features this task is nearly impossible. When in this type of situation, use a dimension reduction technique. We have learned 2 methods to do dimension reduction, with some variations mixed in there. Attempt one or both. This is the toughest step, in my opinion, and I expect many people to get stuck here. Ask for help when you figure out the issue.
8. Lastly, xgboost itself has some tuning parameters. For the final you are required to tune these parameters using a loop. I only require you to tune 2 parameters with a loop. You will be graded on the presence of the loop as well as the grid itself. Further, the table (we created one for the previous xgboost assignment) needs to be saved and turned in. This can go in the “interim” folder.