CS410 Project Proposal

Team: Sentients
Term: Fall 2022
Members:
Raghavendran Ramasubramanian (rr26@illinois.edu) - Captain
Suganya Somu (ssomu2@illinois.edu)

What is your free topic?
Sentiment Analysis to predict positive/negative reviews of a movie in IMDB (movie dataset)

Please give a detailed description. What is the task?
Our object is to build a binary classification model to predict the sentiment of a review with a vocabulary size less than or equal to **1000**. The classifier will be built using regression models. We are planning to use **glmnet** and **text2vec** library to build and train the model. We will be using the same vocabulary for all five training/test datasets, and we should try to produce the AUC equal or bigger than **0.95** over all five test data.

Tasks are:

1. Get the data
2. Clean and transform the data
3. Exploratory data analysis
4. Feature engineering and selection
5. Machine learning using the training dataset
6. Evaluate the trained model
7. Document results/ Deploy the model

Why is it important or interesting?
It's interesting since our model will detect whether a review is positive or negative for a given movie from the list of movies in the IMDB dataset.

What is your planned approach?
The planned approach is to collect data from an open source and then split it as 70:30 to train and test the model. Once the model is built, we can evaluate it against a selected review for a movie and observe the results.

What tools, systems, or datasets are involved?

Tools: **glmnet** and **text2vec** library

System: Classifier (Ridge/Lasso regression with penalty)

Datasets: We are using the IMDB movie dataset. The core dataset contains 50,000 reviews split evenly into 25k train and 25k test sets. The overall distribution of labels is balanced (25k pos and 25k neg). We also include an additional 50,000 unlabeled documents for unsupervised learning.

In the entire collection, no more than 30 reviews are allowed for any given movie because reviews for the same movie tend to have correlated ratings. Further, the train and test sets contain a disjoint set of movies, so no significant performance is obtained by memorizing movie-unique terms and their association with observed labels. In the labeled train/test sets, a negative review has a score <= 4 out of 10, and a positive review has a score >= 7 out of 10. Thus reviews with more neutral ratings are not included in the train/test sets. In the unsupervised set, reviews of any rating are included and there are an even number of reviews > 5 and <= 5.

What is the expected outcome?
The expected outcome is a model which can predict whether the selected movie review is positive/negative. Inorder to make this project unique as compared to other implementations the goal would be to achieve AUC higher than **0.95**.

How are you going to evaluate your work?
We are planning to use the basic measures:
Classification accuracy
Precision and Recall
Area under curve (AUC)

Which programming language do you plan to use?
R

Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

Task	Duration
Identify dataset Research – Data collection	6 hours
Clean and transform the data – Data preparation	5 hours
Explore the dataset – Exploratory analysis	8 hours
Feature engineering and selection – Feature engineering	8 hours
Building the classifier - Development	10 hours
Evaluate the model – Testing and validation	5 hours
Overall documentation & presentation	10 hours