CS598 - Project 3

Xiaoming Ji

Computer System

Hardware

- Dell Precision Tower 5810
- CPU: Intel Xeon E5-1607 @ 3.10GHz
- Memory: 32GB
- GPU: Nvidia GeForce GTX 1080 (2 cards)

Software

- OS: Windows 10 Professional 64bit
- R: 3.5.1
- R Packages:
  - text2vec
  - tokenizers
  - xgboost_0.71.2
  - glmnet_2.0-16

Overview: background, problem statement, objective, input and output of your model, ect.

The goal of this project is to build a binary classification model to predict the sentiment of IMDB movie review.

The given movie review dataset file `Project4_data.tsv` has 50,000 rows (i.e., reviews) and 3 columns. Column 1 "new_id" is the ID for each review (same as the row number), Column 2 "sentiment" is the binary response, and Column 3 is the review.

The given dataset file `Project4_splits.csv` is used to split the 50,000 reviews into Three sets of training and test data. The dataset has 3 columns, and each column consists of the "new_id" (or equivalently the row ID) of 25,000 test samples.

The required performance goal is is to achieve minimal value of AUC over the three test datasets. For model interpretability, the vocabulary size should less than 3000.

For this project, I used logistic linear/lasso regression model to select vocabulary and logistic linear/ridge regression model for training/prediction. Such approach achieved `0.965` AUC with only `1000` vocabulary size.

```r
load("EVAL.OBJ")

kable(eval.obj$perf, booktabs = T) %>%
  row_spec(4, bold = T, color = "white", background = "black")
```

Performance for Split 1	0.9657351	0.9771916	0.9810380
Performance for Split 2	0.9662497	0.9781839	0.9817788
Performance for Split 3	0.9652411	0.9776807	0.9814298
Vocab Size	**1000.0000000**	**2000.0000000**	**3000.0000000**