Jeremy Mednik, Stephen Owen, Jiayi Du, Karim Naous

Project 2 Report:

We can classify authors fairly well based on just word usage. Classification rate on the validation data set is 68% with a model trained on unigrams. Using bigrams/trigrams improved performance very much. Our model, trained on trigrams, scores a classification rate of 78% on the validation set.

Using these models we can learn many interesting things about the authors in our data set. From our coefficient analysis, we learned what the most likely gaming platforms were for several of the authors in our data set. For example, **blackaciddevil** likes Gameboy Advance and other portable gaming consoles; **thomas henry** likes portable gaming consoles too like Playstation Vita and Gameboy. We were also able to associate some misspellings and/or abbreviations with certain authors. For example, the misspelling "tho" for "though" is used by **Michael Hunt** and **Ishmael**.

Firstly, we used stri_split function to split the title variable and obtain the name of the game in lower cases and the consoles it implemented. Our goal was to explore whether consoles as a covariant would help to predict the authorship. Then, we implemented the basic penalized regression model on the text data with consoles covariant. The classification rate for training score and validation score were 0.89/0.59. The scores were close to the results of the penalized regression model without adding covariant. Pursuing this further, we looked up the Segmented Error Rates and the results were interesting. Some users had way above average classification rate, yet some users had way below average classification rate. Based on this observation, we made two hypotheses: (1) Consoles as a covariant is better to help the model to predict the authorship of those who mainly play one type of console. For example, the percentage of PlayStation he/she owned was above 50 percent of the whole collection. (2) Those who mainly play one type of console may hit a higher classification rate than those who did not. To prove our hypothesis, we compared two users, RJ the Great Cat Lover and Dimons. RJ the Great Cat Lover had above average classification rate and over 60 percent of the games she/he owned was PlayStation. By contrast, Dimons had way below average classification rate and he/she had a diverse game consoles collection. There are many limitations on the short project, but it could be interesting to combine consoles and genre in the future to predict the authorship, which would be much more accurate.

We wanted to create a model using a covariate of the average star rating between reviewers with the hopes to see the model accurately predict the reviewer based on the rating a review had. The reviewer with the lowest average rating was Inspector Gadget and the reviewer with the highest average rating was Ivan Orosco who typically leaves 5-star reviews. This model was unfortunately highly inaccurate where the training data had a classification rating of 93% and the valid data had a classification rating of 62%. The particular products each person reviews is a driving factor for a model based on what each user commonly said in each of their reviews.

For example, NeuroSplicer mentioned 'DRM', 'BULDUR' and 'CIVILIZATION' in a lot of their reviews. The particular products do not affect the model completely and there are other factors in predicting the reviewer, such as a signature or writing style. If we wanted to remove it we could create a multinomial model and filter out proper nouns to not include. That would prevent words like Nintendo or Gameboy from predicting a reviewer.

One prominent detail we noticed while looking through this dataset was that the overwhelming majority of the reviews are positive. In fact, 70% of the reviews in the dataset are either 4 or 5 stars. Because of this, there is potential for some voluntary response bias, because people are more likely to review products they feel strongly about (this dataset is comprised of people that write lots of reviews, so the effect is likely lesser than if it were instead a sampling of all amazon reviews, but we believe this still plays a part). Because of this, we are exposed to much more, on average, positive phrasing. We believe that if there were more negative reviews, we would be able to attain higher accuracy due to the exposure to more varied language that would come with a larger number of negative reviews.