*1. Dr. J. Starweather & Dr. A. Moske, 2011, Multinomial Logistic Regression* [*https://it.unt.edu/sites/default/files/mlr\_jds\_aug2011.pdf*](https://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf) *accessed 10 December 2016*

# Multinomial Logistic Regression on adverb frequencies of famous authors using R

For this project the team conducted a Multinomial Logistic Regression on adverb frequencies of famous authors using the R package “nnet”, function multinom(). The objective was to build a model using adverbs common to the selected authors that would predict the author of an unknown manuscript, assuming other works from the author were already in the model.

Practical applications of such a model could include, but are not limited to:

1. Analysis of homework assigned to university students working in teams to classify who wrote certain sections of a given report based on a model using known works from each student.
2. Analysis of unsigned manuscripts to predict the author based on a model built using data from a historical library of manuscripts with known authors
3. Analysis of “tweets” from infamous politicians, are they really writing their own tweets?

Procedure:

1. Text manuscripts from famous authors were found on <http://www.textfiles.com/etext/>
2. Team members were randomly assigned to choose an author and download three or more manuscripts.
3. A simple python script which took hardly any time at all for a novice programmer was written to count and do some math to get frequencies of the hard words and somehow get all the words in a .csv file
4. A github repository was created to store files and share results for the duration of the project
   1. <https://github.com/rlisbona/MSDS_6372_Project_3_Adverbs>
5. Several R functions were evaluated before settling on the Multinom() function in package nnet
6. Key steps in the analysis:
   1. Read the summary file from a local disk drive
   2. Partition the dataset into training and test subsets using createDataPartition()
      1. createDataPartition() was especially useful in this analysis due to the desire of keeping at least one book title from each Author in the training dataset. Unlike a simple random sample function, createDataPartition() includes at least one record from each group, in this case each Author.
   3. Build a model using all available adverb frequencies in the training dataset.
   4. Test the model against the test dataset to determine the accuracy of the model.
   5. Calculate p values for each predictor variable
   6. Rank the importance of variables in the model.
   7. Rerun with a reduced model to check accuracy with fewer predictor coefficients.

Definition of Multinomial Logistic Regression

Multinomial Logistic regression is a form of linear regression where the predicted variable has more than two levels. In this project the predicted variable is the Author, there are 16 Authors with no intrinsic sequence order to the Author list.

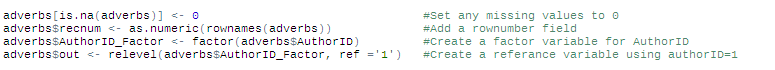
Multinomial logistic regression estimates a separate binary logistic model for each dummy variable, One predicted variable is chosen as the reference level, which results in N-1 models. Each model has its own intercept and coefficients for each predictor variable.



The books data set consists of 51 books from 16 authors showing word frequency for small, medium, and large words plus 24 common adverbs used in all manuscripts.

The data was partitioned into a training dataset and test dataset of two randomly chosen books from two authors.

Figure x shows an example test dataset.





The initial model was run on the “train” dataset consisting of 49 books from 16 Authors, with 24 adverbs shared between the books. This resulted in a very low residual deviance of 0.000096 and an Akaike Information Criterion (AIC) of 660. We want a low AIC so this will be checked against other models

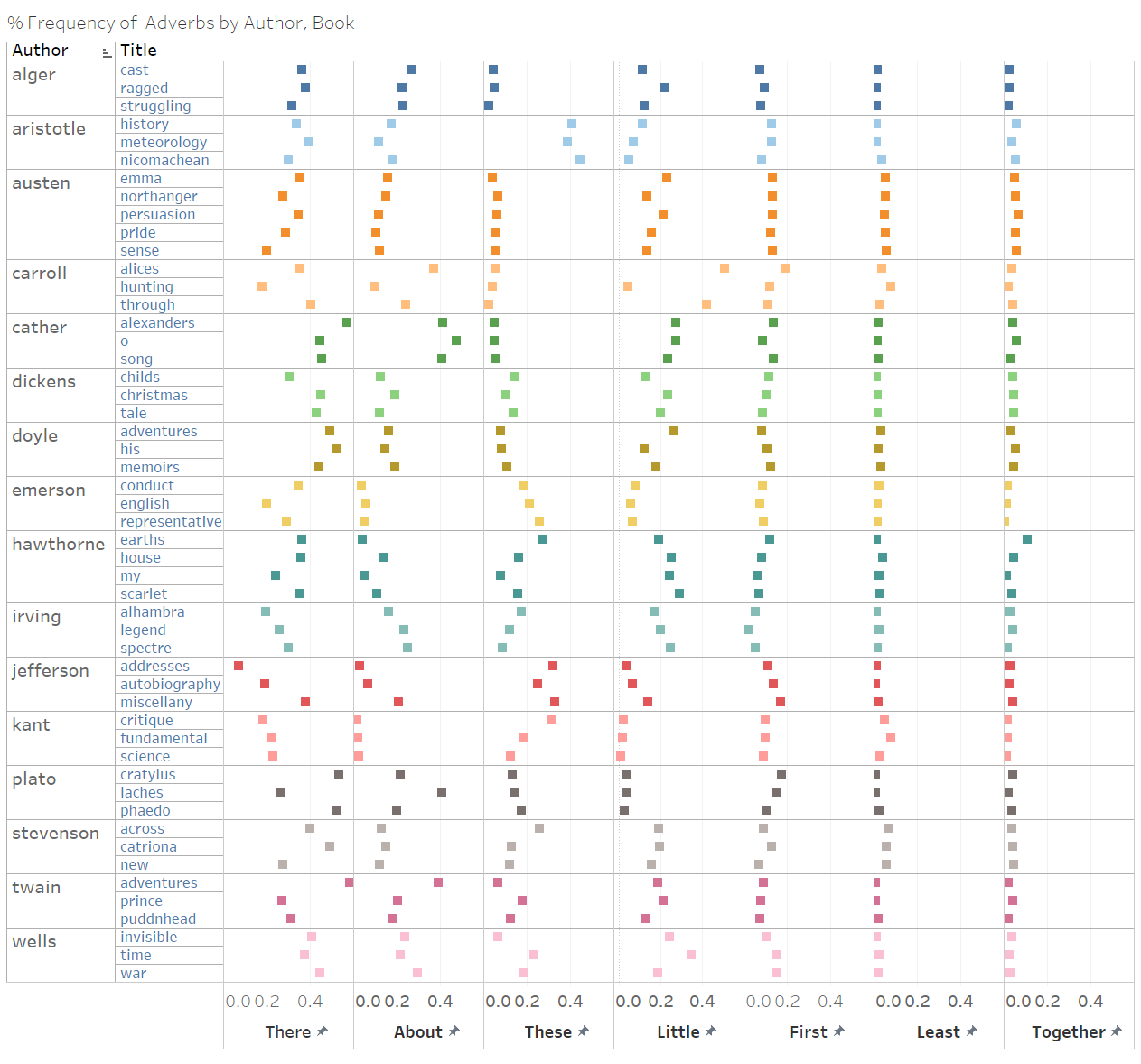




Table 1 - Authors



Figure 1- Adverb frequency by Author, Book

# Appendix A – Books and Author list