Hypothesis: The more closely the distribution of words in a book match the distribution of words in the overall English language, which is ziphian, the more likely the book is to be a best seller

Procedures:

* Take a large sample of books that are best sellers and are not and break each of them down into a vector of the distribution of words, which will be the predictors. The response will be binary, whether the book was classified as a best seller or not.
  + The sample has to be as unbiased as possible, and a 50-50 distribution of best sellers to non-best sellers would be ideal.
  + In the future we could use a numeric response based on the total number sold as an extension of the work.
  + The data set will be broken into a 80% training set and 20% testing set
* Sweep the distribution of words in the general English language out from the predictors, leaving the predictors as deviations from the overall frequencies
* Regularize based on the total number of words in the books – books with less words will have a greater variability than books with more words.
  + Use cross validation to optimize regularization parameter
* Some major dimension reduction will be needed – first remove predictors with variance under a certain threshold, then PCA. We want to keep like 10-50 predictors.
* Use several machine learning techniques to try to predict whether the book is a best seller or not – use the train function from the caret package for all
  + LDA
  + QDA
  + Logistic Regression
  + LOESS/LOWESS
  + KNN
  + Random Forest
  + Potentially an Ensemble of all of the above
* Potentially cluster the data – hierarchical and k-means
* Evaluate accuracy on the test set. For this analysis, there will be no difference in the relative importance of specificity and sensitivity.