Homework II

PLSC 497 – Text as Data

Professor Kevin Munger

This homework is due electronically by 11:59 p.m. EST on Wednesday, April 7 2021. You can submit your homework by emailing copies both to Prof. Munger (kmm7999@psu.edu) and Mr. Villegas-Cruz (amv5718@psu.edu). Late work will incur penalties of the equivalent of one third of a letter grade per day late.

It must be your own work, and your own work only—you must not copy anyone's work, or allow anyone to copy yours. This extends to writing code. You may consult with others, but when you write up, you must do so alone.

Your homework submission must be in one of the following formats: (1) A set of answers and a clearly commented R code appendix (use comments to identify code relevant to each answer you produced), (2) A report consisting of clearly marked answers, each accompanied by the relevant code (e.g., a report generated using *rmarkdown, knitr,* or similar). **In either case, your code must be included in full, such that your understanding of the problems can be assessed.**

PART 1

1. We would like you to perform some Naive Bayes classification **by hand** (that is, you may use math functions or DFM-creating functions, but not any built-in naive Bayes functions). Make sure to show your work!

   (a) Imagine a situation in which you receive emails from the two main U.S. parties in antici- pation of the 2020 election. The contents of those emails after all relevant preprocessing are displayed in Table 1. Using the standard Naive Bayes classifier without smooth- ing, estimate for each party the posterior probability (or rather, the prior multiplied by the likelihood) that the following email was sent by the respective party: "immigration voter aliens help economy". Report these estimates. Based on these results, which party would you predict sent the mystery email? Explain whether you trust your findings and why.

   (b) Now impose Laplace smoothing on the problem and re-estimate each party's respective posterior probability. Report your findings. Based on these new results, which party would you predict sent the mystery email? Beyond computational reasons (i.e. avoiding log(0)'s), can you think of any theoretical reason why smoothing might make sense (hint: the above data is but a sample of each party's shared language).

email	content
republican1	immigration aliens wall emergency country
republican2	voter economy president growth security
republican3	healthcare cost socialism unfair help
democrat1	immigration country diversity help security
democrat2	healthcare universal preconditions unfair help
democrat3	economy inequality opportunity voter help
democrat4	abortion choice right women help

Table 1: Training set of presidential candidate emails.

PART 2

For this exercise you will use a database of Yelp reviews gathered for a Kaggle challenge (source). Each user left a star rating of 1-5 along with a written review. You'll be asked to use some of the supervised learning techniques we've discussed in class to analyze these texts.

Download the most recent version from the course GitHub. The data are available in the file "yelp .csv". Before we get started, be sure to actually read a few of the reviews, to get a feel for the language used, and any potential imperfections in the text created during the scraping process.

For each task (3) through (6), begin with the raw version of the text, and briefly explain which pre-processing steps are appropriate for that particular task.

2. Before we apply any classification algorithms to the Yelp reviews, we will need a general classifier that tells us whether the review was positive or negative—also referred to as the "actual score."

   (a) Divide the reviews at the empirical median score and assign each review a label as being "positive"—if the user score was greater than the empirical median score—or "negative"—if the review is less than or equal to the empirical median (you can use "1" and "0" as labels if you prefer, just be consistent as you do the exercises below).

   (b) For some tasks, we will need "anchor" texts at the extreme of the distribution. Create a character variable (name it "anchor") that has value "positive" if the user star rating given to a review is equal to 5, "neutral" if the user rating is less than 5 but greater than 1 and finally "negative" if the user rating is equal to 1. Report the proportion of reviews that are anchor positive, neutral and negative.

3. Next, we'll train a Naive Bayes classifier to predict if a review is positive or negative.

   (a) Use the "textmodel" function in quanteda to train a smoothed Naive Bayes classifier with uniform priors, using 80% of the reviews in the training set and 20% in the test set (Note: features in the test set should match the set of features in the training set. See quanteda's **dfm match** function.). Report the accuracy, precision, recall and F1 score of your predictions. Include the confusion matrix in your answer.

   (b) Were you to change the priors from "uniform" to "docfreq," would you expect this to change the performance of Naive Bayes predictions? Why? Re-estimate Naive Bayes with the "docfreq" prior and report the accuracy, precision, recall and F1 score of these new results. Include the confusion matrix in your answer. In terms of accuracy, how would you evaluate the performance of this classifier?

   (c) How is accuracy affected if you fit the model without smoothing? Why might this be?

4. Finally, let's try out a Random forest classifier. For this question use the first 500 reviews in the dataset.

   (a) As we did for the Naive Bayes model, split the dataset into a training (80%) and a test set (20%) and construct a document feature matrix for each (Note: features in the test set should match the set of features in the training set).

   (b) Using the **randomForest** package fit a random forest model to the training set using the package's default values for **ntree** and **mtry** (also, set the **importance** argument to TRUE). Having fitted the model, extract the mean decrease in Gini importance for the feature set and order from most important to least important. What are the top 10 most important features according to this measure?

   (c) Using the fitted model, predict the sentiment values for the test set and report the con- fusion matrix along with accuracy, precision, recall and F1 score.

PART 3

4. Collect all of the novels written by Charles Dickens from Project Gutenberg.

   (a) Calculate the following statistics:

      i. Which one has the most tokens?

      ii. Which one is the most lexically diverse?

      iii. Using cosine similarity, which two are the most similar to each other?

   (b) Calculate the Flesch Reading Ease score for each book, and plot these scores on the y-axis with the dates of publication for each book on the x-axis. Is Dickens getting more or less complex over time?

5. Using James Joyce's "A Portrait of the Artist as a Young Man" (gutenberg id = 4217) and Mark Twain's "The Adventures of Tom Sawyer" (gutenberg id = 74), make a graph demonstrating Zipf's law. Include this graph and also discuss any pre-processing decisions you made.

6. Both James Joyce's "A Portrait of the Artist as a Young Man" and Mark Twain's "The Adventures of Tom Sawyer" broach the topic of religion, but in very different ways. Choose a few Key Words in Context and discuss the different context in which those words are used by each author. Give a brief discussion of how the two novels treat this theme differently.