DS5110 Homework 6

Kylie Ariel Bemis

31 March 2024

Instructions

Your solutions should include all of the code necessary to answer the problems. All of your code should run (assuming the data is available). All plots should be generated using `ggplot2`. *Make sure that you answer all parts of the problem.*

Submit your solutions on Canvas by the deadline displayed online. For full credit, your submission must include exactly two files:

- R Markdown (.Rmd)
- PDF report (.pdf)

Problems must appear in order, and problem numbers must be clearly marked. Any written responses should appear outside of code blocks and use Markdown for text formatting. Code comments are encouraged, but will be ignored for grading purposes. Solutions that are especially difficult to grade due to poor formatting will not receive full credit.

All solutions to the given problems must be your own work. If you use third-party code for ancillary tasks, you **must** cite them.

Part A

Problems 1–3 ask you to visualize tweets from @realDonaldTrump, as collected by https://www.thetrumparchive.com. Download the data files from "twitter.zip" on Piazza.

Problem 1

Import "realDonaldTrump-20201106.csv" making sure that the tweet IDs are preserved. Structure the tweets into a tidy text format using the `token="regex"` option. Process the data as follows:

- Do not include re-tweets
- Do not include tweets without any spaces
- Remove stop words and the token "rt"
- Remove variations on Donald Trump's name
- Remove URLs and twitter @usernames
- Replace all punctuation (EXCEPT for #) with empty strings
- Remove zero-length tokens (after applying above replacements)

(Some special characters like emojis may get through this processing. This is acceptable.)

Then visualize the top 20 most common terms in Donald Trump's tweets.

*Hint: The regular expression [[:punct:]]+ can be used to match **all** punctuation. The `str_replace_all()` function may be helpful. There are many potential ways to accomplish the above preprocessing.*

Problem 2

Visualize the top 20 most common terms in Donald Trump's tweets for each year from 2015-2020, and comment on the visualization.

Problem 3

Treat year as a "document" to calculate the tf-idf for each term and year. Visualize the top 20 most characteristic terms in Donald Trump's tweets for each year from 2015-2020, and comment on the visualization.

Part B

Problems 4–5 ask you to fit and interpret a model to the tweets analyzed in Part A.

Problem 4

Filter the data to include only tweets from 2016-2020, and then use the `glmnet` package fit sparse regression models to predict the number of retweets that a tweet will get. Use cross-validation to select the sparsity parameter lambda. Report the selected value of lambda and the number of non-zero coefficients in the regression model. (You do not need to partition the dataset beforehand or report the error.)

Hint: Use the `rownames()` and `colnames()` of the sparse matrix to extract the terms and IDs.

Problem 5

Extract the coefficients from the best model from Problem 4, and visualize the terms with the strongest positive relationship with the number of re-tweets. Comment on the visualization.