Assignment 03 - APIs and Text Mining

Due Date

March 7, 2025 by 11:59pm. The learning objectives are to collect data from an API and perform text mining.

NASA API (Part 1)

The aim is to get more practice working with APIs. We will use and API maintained by NASA for retrieving information about near earth objects (NEOs), asteroids that pass close to Earth. The documentation is available at https://api.nasa.gov/ (scroll down to the API titled Asteroids NeoWs). First you will need an API key for accessing the service. You can get one at https://api.nasa.gov/ where you will need to supply an email address, which can be either your UofT email or a personal email address.

1. (6 points) Use a recent `start_date` and `end_date` along with the API key to submit an `HTTP GET` request to the NASA NeoWs Feed API. Pick a range of 7 days (maximum query is 1 week). There should be 3 elements in your retrieved query: a list of urls, the number of near earth objects that had their nearest approach during the time spanned by `start_date` and `end_date`, and an object whose attributes are the dates (represented by strings of the form YYYY-MM-DD) in the time spanned by `start_date` and `end_date`. Each date attribute has an array of values that represents a near Earth object. Write out your steps, and summarize how many near earth objects you retrieved along with the parameters that are available for them.

2. (3 points) Extract information on `estimated_diameter`, `is_potentially_hazardous_asteroid`, `miss_distance`, and `relative_velocity`, rename the variables to a neater format. Create a variable for mean `estimated_diamter` (from min and max). Keep the full dataset for the next questions, but also create a table summarizing these variables by date (take averages or medians and include a count of hazardous asteroids and total number of NEOs by date).

3. (2 points) Create a correlation heatmap of the numeric variables in the NEO dataset (you can use `ggcorrplot`). Describe the correlations between the variables.

4. (4 points) Are larger NEOs more hazardous? Are faster NEOs more hazardous? Create two histograms to explore these questions, and conduct two-sample t-tests. Summarize your results.

5. (4 points) Is the relationship between diameter and velocity different between hazardous and non-hazardous NEOs? Create a scatterplot to explore this. Fit a logistic regression model to see if we can predict whether an asteroid is hazardous based on its attributes. Summarize your results.

CFPB API and Text Mining (Part 2)

The Consumer Financial Protection Bureau maintains a database of customer complaints. We will use text mining to see if there are any trends in the reported consumer's narrative description of the issue/complaint. The data can be acquired here: https://www.consumerfinance.gov/data-research/consumer-complaints/. However, we will use the API to get data https://cfpb.github.io/api/ccdb/api.html.

7. (2 points) Using the API get consumer complaint data for the last year (i.e. since February 1, 2024), filtering for the product "Credit reporting or other personal consumer reports". Please note downloading the csv format and reading it with data.table is a good idea because this is a very large dataset. Summarize the dataset dimensions and variables.

8. (4 points) Tokenize the complaints (consumer complaint narrative) and count the number of each token. Do you see anything interesting? Remove stopwords, including numeric tokens and XXXX combinations (use grep). What words appear as the most frequent? Create a plot of the 20 most common complaint tokens and summarize.

9. (2 points) Tokenize the complaints into bigrams. Find the 10 most common bigrams, visualize and summarize.

10. (3 points) Calculate the TF-IDF value for the complaints in the top 3 issues. Here, the issue is like the document. What are the 5 tokens from each issue with the highest TF-IDF value? How are the results different from the answers you got in question 8?

11. (3 points) Use sentiment analysis to determine if consumer complaints express positive, neutral, or negative sentiments. Make a plot and summarize your results.