Homework 3 - More APIs, Classification, NLP

Due Date

March 1, 2026 by 11:59pm (total 53 points)

Learning Goals

- Use two more APIs, get familiar different datasets.
- Train and test logistic regression and random forest for classification.
- Conduct text analysis (natural language processing), topic modeling, and sentiment analysis.

NASA API (Part 1)

The aim is to get more practice working with APIs and to conduct logistic and random forest classification. We will use and API maintained by NASA for retrieving information about near earth objects (NEOs), asteroids and comets that pass close to Earth. The documentation is available here (click on Asteroids - NeoWs). General information about NEOs can be found here. First you will need an API key for accessing the service. You can get one here - you will need to supply an email address, which can be either your UofT email or a personal email address.

1. (3 points) The NASA NeoWs Feed API only allows pulls of 7 days of data at a time, so create a loop to pull all NEOs observed so far this year (i.e. `start_date` and `end_date` will be for a week but we want to get all asteroids since January 1, 2026). Your query will pull a list of asteroids based on their closest approach date and will have various attributes. Summarize the number of NEOs you pulled, the how many days had NEOs since January 1 of this year.

2. (5 points) Extract information on `estimated_diameter`, `is_potentially_hazardous_asteroid`, `miss_distance`, and `relative_velocity`. Create a variable for `avg_estimated_diameter` (take the mean of min and max). Make a summary table of these variables, a correlation heatmap of the numeric variables, and create a stacked barplot of the number of asteroids by day, distinguishing those that are hazardous `is_potentially_hazardous_asteroid`.

3. (6 points) Are larger NEOs more hazardous? Are faster NEOs more hazardous? Create two histograms to explore these questions, and conduct two-sample t-tests. Is the relationship between diameter and velocity different between hazardous and non-hazardous NEOs? Create a scatterplot to explore this. Summarize your results.

4. (5 points) Split the data into 70% train and 30% test, and fit a logistic regression model (on train) to see if we can predict (on test) whether an asteroid is hazardous based on its attributes (include diameter, velocity, miss distance). Summarize and interpret the results (include a confusion matrix, accuracy, precision, recall, and F1; plot the ROC curve).

5. (5 points) Repeat 4 but using random forest. Summarize and include a comparison of your results in 4.

CFPB API (Part 2)

The Consumer Financial Protection Bureau maintains a database of customer complaints. We will use natural language processing to see if there are any trends in the reported consumer's narrative description of the complaints. The data can be acquired here. We will use the API to get data.

6. (3 points) Use the API to get consumer complaint narratives (i.e. `has_narrative`) from the last 3 months (i.e. `date_received_min` is November 1, 2025) across all product categories. I suggest using the csv `format`. Please note this is a very large dataset so may take some time to query (you can use a longer timeout). Once you have pulled the data, summarize the dimensions and variables. Create a bar chart of complaint counts by product category.

7. (2 points) Tokenize the complaints (consumer complaint narrative) and count the number of tokens. Make a plot of the top 20 tokens. Summarize.

8. (3 points) Remove stopwords, including numeric tokens, xxxx combinations, and customize to remove any other words that look strange. After cleaning, what words appear as the most frequent? Create a plot of the 20 most common complaint tokens in your cleaned data and summarize.

9. (10 points) Use Latent Dirichlet Allocation (LDA) topic modeling to discover topics in the complaint narratives. First, use perplexity and log-likelihood on a held-out sample of 50,000 to evaluate different numbers of topics (e.g., 2 to 10) and select an appropriate

number. Plot perplexity and log-likelihood as a function of the number of topics. Then fit an LDA model with your chosen number of topics. Display the top 10 words for each topic. Do the discovered topics align with the actual product categories? Create a visualization to compare. Summarize and interpret your findings.

10. (5 points) Use the `vader` lexicon to conduct sentiment analysis on the word tokens to determine if consumer complaints express positive, neutral, or negative sentiments. Create a histogram of the sentiment scores, and make a barplot to show the counts of tokens by sentiment category. Make another barplot to compare sentiment category counts by product categories. Make another barplot to show which words have top positive sentiment scores and which words have most negative sentiment scores. Summarize and interpret your results.

11. (2 points) Conduct sentence-level tokenization, summarize the number of sentences, and the average number of sentences by complaint.

12. (4 points) Conduct sentiment analysis of the sentence tokens. Like above, visualize the histogram of the sentiment scores, categorize sentiments into negative, neutral, positive and visualize counts with a barplot as well as a barplot by product categories. Interpret, summarize, and discuss differences from word token sentiment analysis in 10.