Homework 1: Authorship Analysis of the Federalist Papers

Deadline: November 3, 2025

Setup

You may use either **Python** or **R** or anything else for this assignment.

- In R: use R Markdown and submit the knitted output (.html or .pdf) or the `.Rmd` source file.
- In Python: use a Jupyter notebook and submit the `.ipynb` file.
- If using any other language, submit code and short word-based interpretations (where required).

Task 1: Repeat the Authorship Analysis for All Disputed Federalist Papers

1. **Identify Disputed Papers:** Inspect the `author` column and find rows labeled `"HAMILTON OR MADISON"`.
2. **Word Frequencies:** Choose a set of words, start with the three used in class (`"man"`, `"by"`, `"upon"`) and add five additional words you believe help distinguish authorship.
3. **Author Probabilities:** For each author (Hamilton, Madison, Jay), compute their word probabilities (relative frequencies) $\hat{\mu}_h, \hat{\mu}_m, \hat{\mu}_j$.
4. **Posterior Computation:** For each disputed paper, use the word counts to compute the posterior probability of each author.
5. **Prediction:** Assign each disputed paper to the most likely author based on the posterior probabilities.

Task 2: Logged Odds Differences for Bigrams

1. Tokenize text into **bigrams**.
2. Either remove every bigram that contains a **stopword**, or remove stopwords *before* tokenizing into bigrams.
3. Remove the bigrams that occur only once.
4. Separate the bigrams by author (focus only on **Hamilton** and **Madison**).
5. Compute the **relative frequencies** of bigrams for each author.
6. Compute the **log odds** for each bigram:

$$\log O_b^i = \log\left(\frac{f_b^i}{1 - f_b^i}\right)$$

   where f_b^i is the relative frequency of bigram b for author i.

7. Compute the **log odds differences** for each bigram:

$$\Delta_{bigram} = \log\left(\frac{f_b^i}{1 - f_b^i}\right) - \log\left(\frac{f_b^j}{1 - f_b^j}\right)$$

   where i and j are the authors being compared.

8. **Plot** the log odds differences of the top discriminative bigrams.

Task 3: Dictionary Analysis

Use a simple sentiment or style dictionary to characterize writing differences between authors. For example:

- Use any of the existing/built-in sentiment lexica (`bing`, `nrc`, `afinn`) or a self-created dictionary.
- Compare the average sentiment scores across authors.
- Plot the sentiment distribution for the authors.

Submission

Please submit your file to my email: petro.tolochko@hotmail.com.