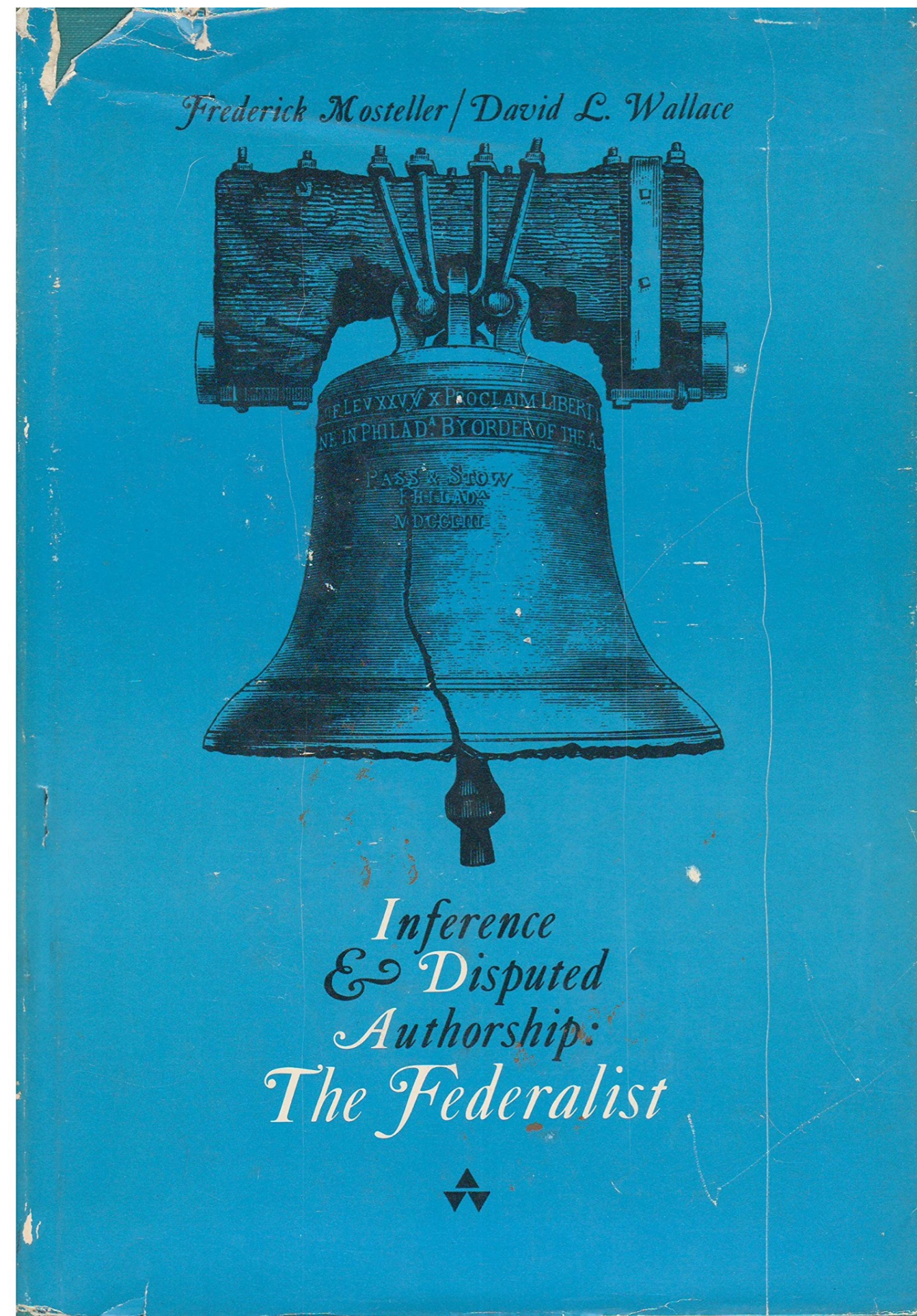


Quantitative Text Analysis

Meeting 3

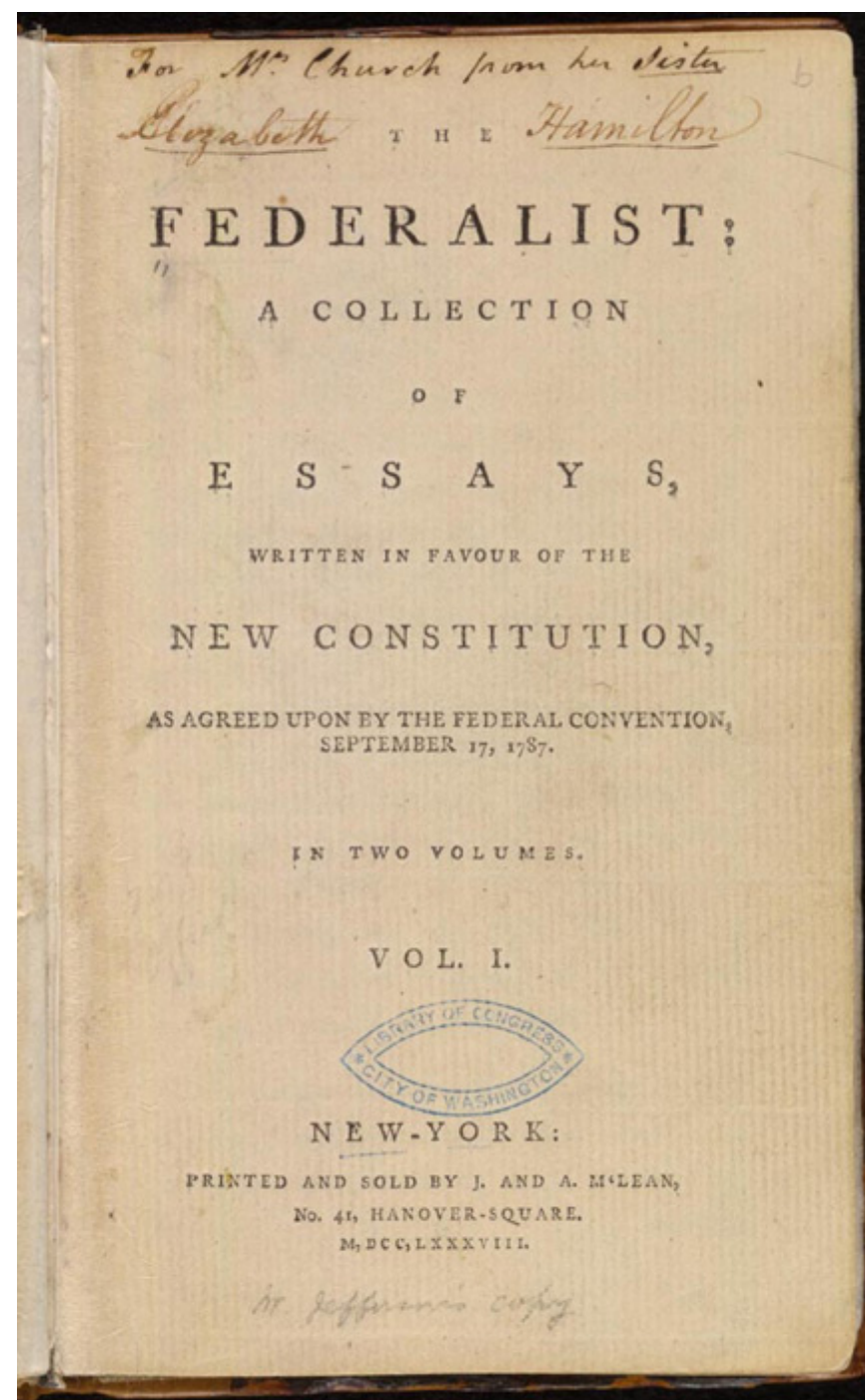
Petro Tolochko

Inference and Disputed Authorship: The Federalist

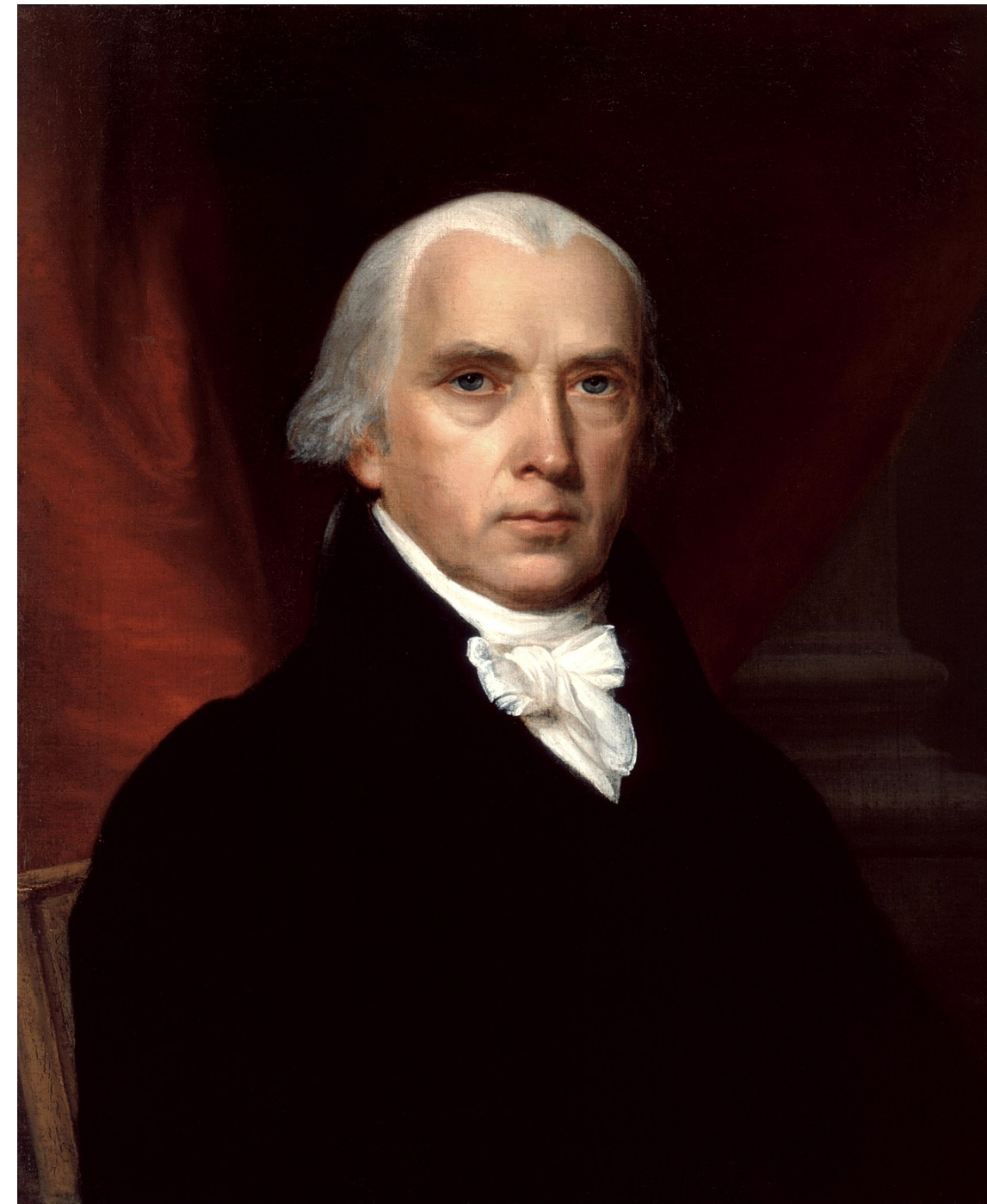
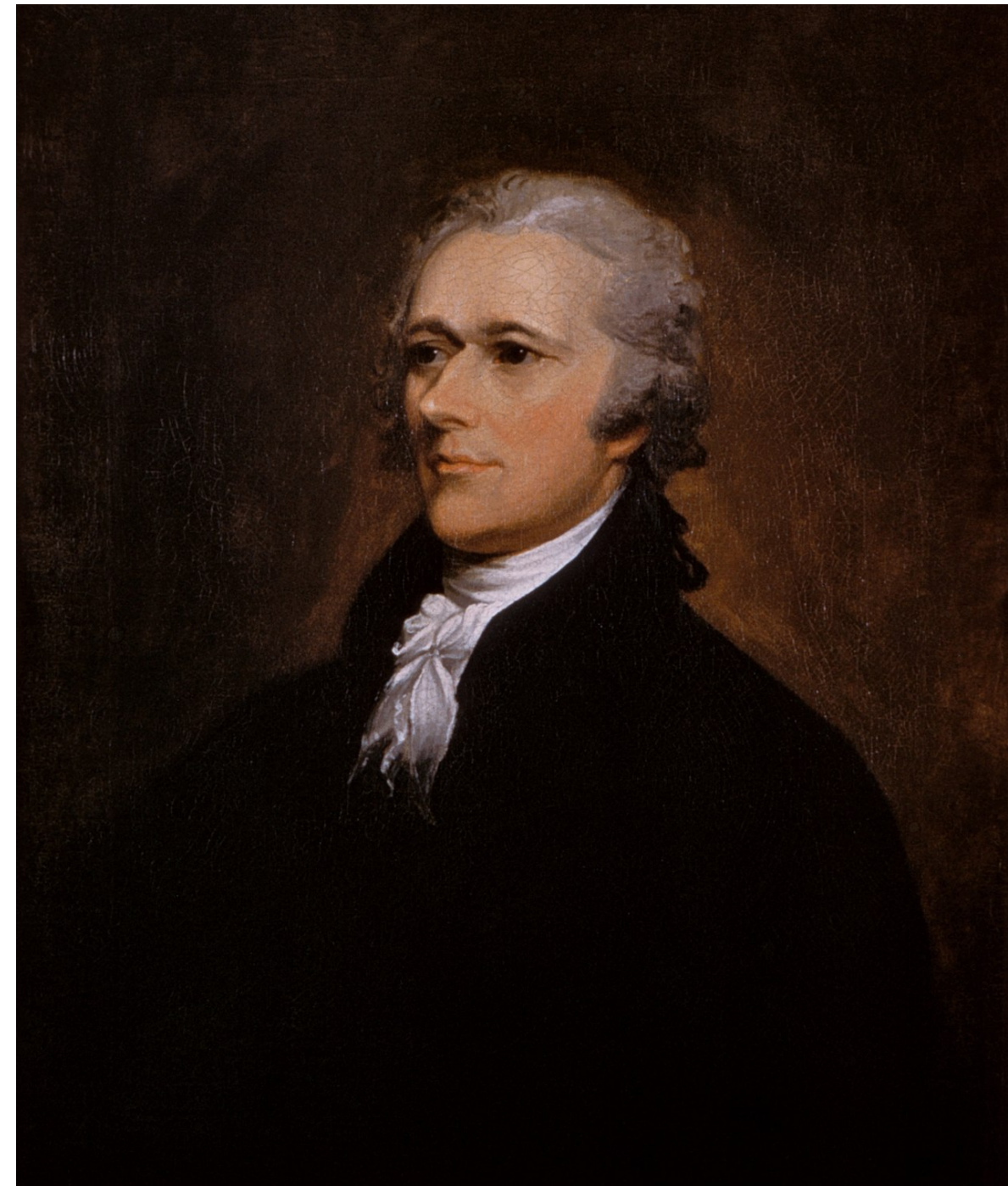
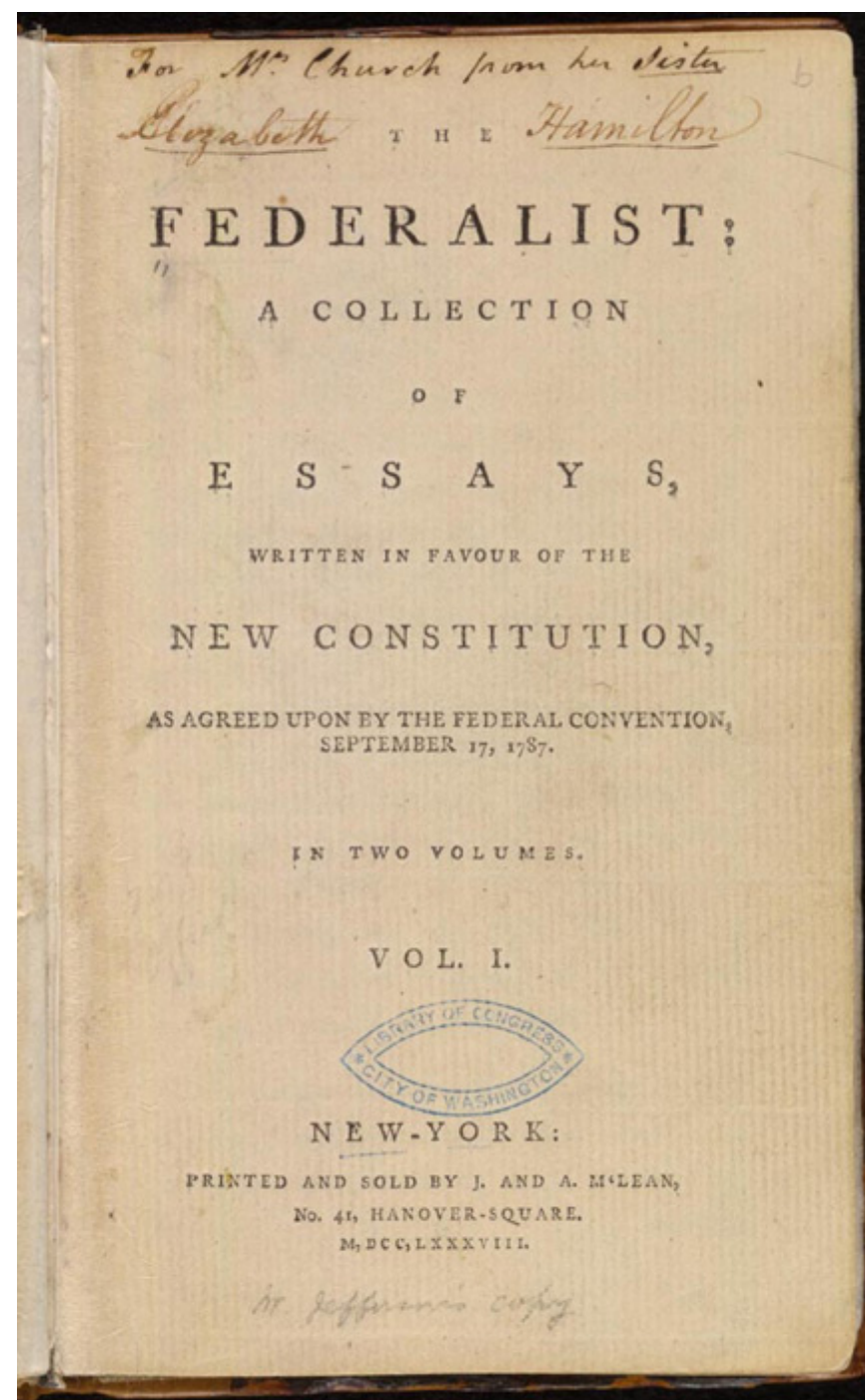


Frederick Mosteller &
David L. Wallace, 1963

One of the first (if not the first) text-as-data study

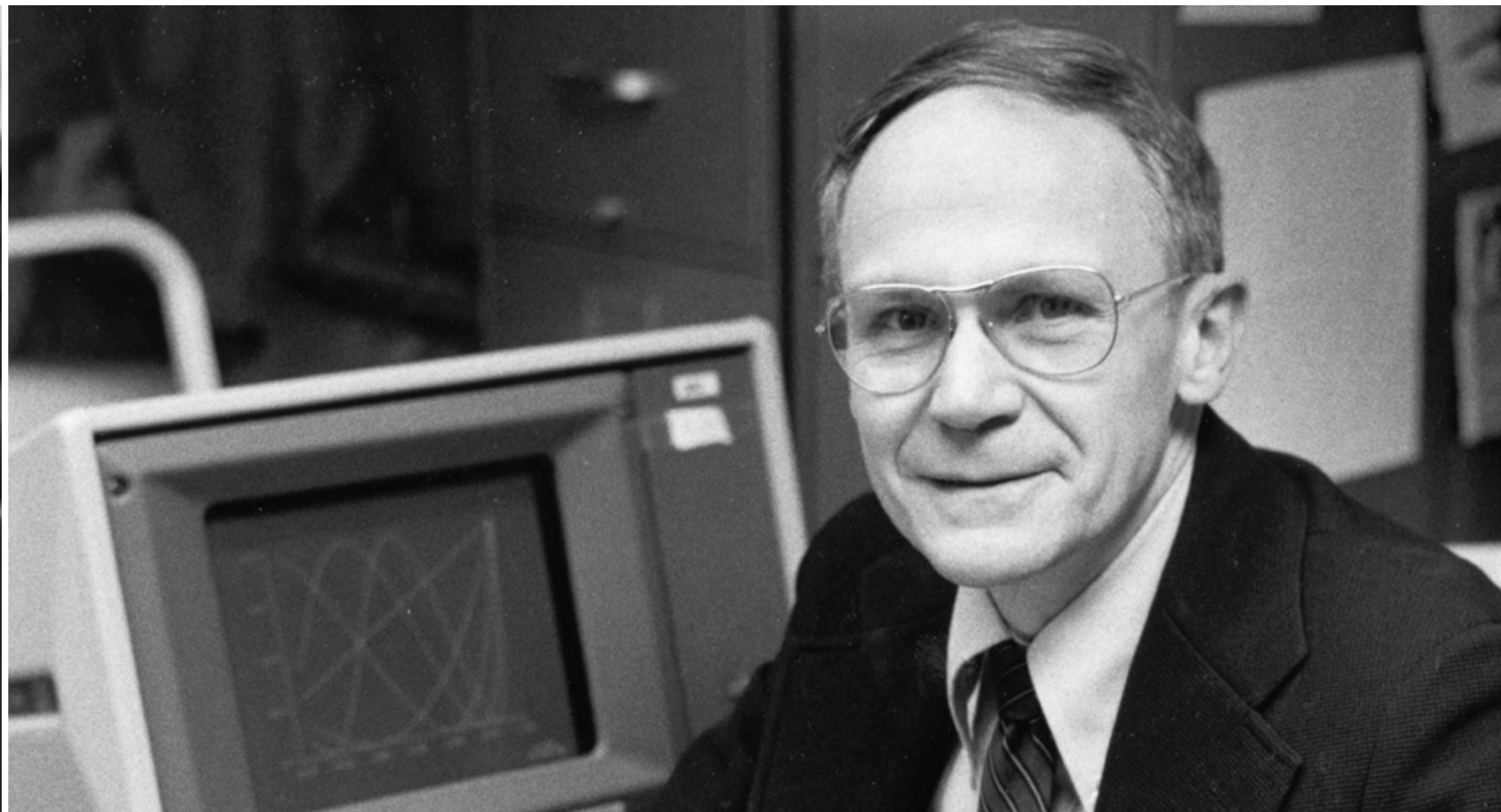


One of the first (if not the first) text-as-data study



Who wrote them?

- 71 of the essays have a fairly certain authorship
- 12 are disputed
- Big historical debate as to how to ascribe authorship



Computer-assisted text analysis!

Computer-assisted text analysis...?



Dimension Reduction

- Remove all the stop-words!

Dimension Reduction

- Remove all the stop-words!
- Still... too many words!
-

Dimension Reduction

- Remove all the stop-words!
- Still... too many words!
- Remove all the words BUT the stop-words

Dimension Reduction

- Remove all the stop-words!
- Still... too many words!
- Remove all the words BUT the stop-words
- Maybe there is information in them!

Simplified example from Grimmer et al., 2022

- Focus on:
 - “Man”
 - “By”
 - “Upon”
- The rates with which the authors use these words may indicate authorship

Word Rates

| | man | by | upon |
|----------|-----|-----|------|
| Hamilton | 102 | 859 | 374 |
| Madison | 17 | 474 | 7 |
| Jay | 0 | 82 | 1 |

Word Proportions

| | man | by | upon |
|----------|------|------|------|
| Hamilton | .076 | .643 | .28 |
| Madison | .034 | .952 | .014 |
| Jay | 0 | .988 | .012 |

Word Proportions

Multinomial Model of
Language



| | man | by | upon |
|----------|------|------|------|
| Hamilton | .076 | .643 | .28 |
| Madison | .034 | .952 | .014 |
| Jay | 0 | .988 | .012 |

Disputed Paper

| | man | by | upon |
|----------|-----|----|------|
| Disputed | 2 | 15 | 0 |

Disputed Paper

$$p(D|H) = \frac{17!}{2!15!0!} (.076)^2 \times (.643)^{15} \times (.28)^0$$

Disputed Paper

Total Words

Raw Rates

$$p(D|H) = \frac{17!}{2!15!0!} (.076)^2 \times (.643)^{15} \times (.28)^0$$

Hamilton Rates

The diagram illustrates the components of the multinomial probability formula. Red arrows point from the labels to specific parts of the equation: 'Total Words' points to the 17 in the numerator's factorial; 'Raw Rates' points to the (.28) term; and 'Hamilton Rates' points to the (.076) and (.643) terms. The 0 in the denominator's factorial is also pointed to by an arrow from the 'Hamilton Rates' label.

Calculate Jay and Madison

$$p(D|H) = \frac{17!}{2!15!0!} (.076)^2 \times (.643)^{15} \times (.28)^0 = .001$$

$$p(D|M) = \frac{17!}{2!15!0!} (.034)^2 \times (.952)^{15} \times (.014)^0 = .076$$

$$p(D|J) = \frac{17!}{2!15!0!} (0)^2 \times (.988)^{15} \times (.012)^0 = 0$$

Federalist Vector Space Model (And More!)

- In the Markdown file...