Correspondence Networks and Linguistic Change among the American Founding Fathers (1706–1836)

Lewis Carson
Durham University

Date: December 4, 2025

Introduction

This report presents a computational analysis of the correspondence patterns of the American Founding Fathers across distinct historical periods, ranging from the Colonial era (1706) to the post-Madison presidency (1836). By employing a combination of network analysis and natural language processing (NLP) techniques, specifically Term Frequency-Inverse Document Frequency (TF-IDF), n-gram analysis, and stylometric measures, we aim to uncover how the intellectual, political, and social preoccupations of these figures evolved over time. The study utilizes a "distant reading" approach to identify macro-level patterns in communication structure and vocabulary that would be indiscernible through traditional close reading of the massive corpus of over 183,000 letters.

Problem and Research Question

The primary research problem addresses the challenge of understanding the evolution of political discourse and personal relationships among the Founding Fathers during a century of profound transformation. With a corpus exceeding 180,000 documents, manual analysis is insufficient to capture the shifting dynamics of influence and language. This research investigates four key questions: (1) How did the vocabulary and prominent topics of correspondence shift across seven distinct historical periods? (2) How did the structure of the correspondence network change, and who emerged as central figures in different eras? (3) Are there observable trends in the formality and complexity of language used? (4) Do these computational findings align with established historical narratives regarding the development of the American nation?

Methodological Rationale

To answer these questions, this study employs a multi-modal computational approach. While network analysis can reveal *who* was communicating, it cannot explain *what* they were discussing. Conversely, linguistic analysis can track vocabulary but ignores the structural power dynamics of the correspondents. By integrating Network Analysis with three distinct NLP techniques (TF-IDF, N-grams, and Stylometry), we aim to triangulate the data.

Data Collection

Data was collected from the Founders Online archive (https://founders.archives.gov/), a comprehensive digital collection of the papers of major Founding Fathers. The dataset comprises metadata and full text for 183,673 documents spanning from 1706 to 1836.

The collection process was implemented in Python using the standard `urllib` library to interface with the Founders Online API. Due to the large volume of data, a robust scraping architecture was developed:

- **Two-Stage Harvesting:** First, a lightweight metadata harvest collected document IDs, dates, authors, and recipients. Second, a content download script fetched the full text for each document.
- **Checkpointing System:** To handle potential network failures during the long download process, a custom checkpointing mechanism (`download_checkpoint.json`) was imple-

mented. This allowed the scraper to resume from the last successful download rather than restarting, ensuring data integrity and efficiency.

- **Storage Format:** Data was stored in JSON Lines (`.jsonl`) format. Unlike a standard JSON array which requires loading the entire file into memory, JSONL allows for line-by-line processing, which is essential for handling a text corpus of this magnitude (hundreds of megabytes) on standard hardware.

The data was subsequently partitioned into eight historical periods (e.g., Colonial, Revolutionary War, Washington Presidency) to facilitate temporal analysis. The scale and rich metadata of this dataset make it highly appropriate for both network and linguistic analysis.

Remote Download and Transfer

Because the corpus is large and the full content download is I/O-heavy and may run for many hours, the initial data download was executed on the university HPC cluster "Hamilton" using a Slurm batch job. A small Slurm script (see `download.slurm`) was used to submit the process; the job requests moderate resources (48-hour wall time, 8GB memory) and runs the `download.py` script remotely. Running the download on Hamilton has two key benefits: (1) a stable, high-bandwidth connection to the remote API avoiding home-network interruptions; (2) the ability to restart or checkpoint long-running jobs using Slurm's job control.

Reproducibility and Code Archive

To facilitate reproducibility, the repository includes scripts for data download and analysis. Key points for reproducing the pipeline are:

- Dependencies: Core Python packages used include `nltk`, `pandas`, `networkx`, and `scikit-learn`. The exact versions used are recorded in the project environment (a `requirements.txt` is recommended in the archive for submission).
- Re-running the pipeline (example): First, ensure the dependencies are installed, then run the scripts in order:
  - `python3 download.py` (or submit via Slurm as above)
  - `python3 tfidf.py`
  - `python3 ngram.py`
  - `python3 stylo.py`
  - `python3 create_network.py`
- Checkpoints and one-line resumption means long-running tasks can be resumed via the checks recorded in the `download_checkpoint.json` file.

Code Reuse and Attribution

Some code and resources are reused from open-source packages and online tutorials. In particular, the project uses:

- **NLTK** for tokenization and lemmatization (`WordNetLemmatizer`) and for small preprocessing utilities (Bird et al., 2009).

- **NetworkX** for constructing and analyzing correspondence graphs.

- **Pandas** and standard Python libraries for data handling and file I/O.

Model Description and Implementation

The analysis was implemented using Python, leveraging the Natural Language Toolkit (NLTK) for linguistic processing and standard libraries for data manipulation. The corpus was first partitioned into eight distinct historical periods based on key dates: Colonial (pre-1775), Revolutionary War (1775–1783), Confederation (1784–1789), Washington Presidency (1789–1797), Adams Presidency (1797–1801), Jefferson Presidency (1801–1809), Madison Presidency (1809–1817), and Post-Madison (1817–1836).

Data Preprocessing

For all NLP tasks, the text underwent a rigorous preprocessing pipeline. Text was lowercased, and non-alphabetic characters were removed. We employed the NLTK `WordNetLemmatizer` to reduce words to their base forms (lemmas). A custom stopword list was created, combining standard English stopwords with common 18th-century epistolary terms (e.g., "thou", "hath", "servant", "obedient", "favour") to reduce noise and focus on content-bearing vocabulary.

Network Analysis

A directed graph was constructed where nodes represent historical figures (senders and recipients) and edges represent the volume of correspondence between them. The network data was extracted by parsing the metadata of all 183,673 documents. This structure allows for the calculation of centrality metrics and the visualization of communication density using chord diagrams.

TF-IDF Analysis

To identify the characteristic vocabulary of each era, we applied Term Frequency-Inverse Document Frequency (TF-IDF). Uniquely, we defined a "document" not as an individual letter, but as the *aggregated text of all letters within a single historical period.*

- **Term Frequency (TF):** The frequency of a term t in period d, normalized by the total word count of that period.

- **Inverse Document Frequency (IDF):** Calculated as $\log(N/df(t))$, where $N = 8$ (the number of periods) and $df(t)$ is the number of periods containing term t.

This approach highlights terms that are statistically over-represented in a specific period relative to the entire timeline.

N-gram Analysis

We extracted bigrams (2-word sequences) and trigrams (3-word sequences) to capture rhetorical patterns and compound nouns (e.g., "United States", "public good"). Unlike TF-IDF, N-gram analysis relied on raw frequency counts within each period to identify the most common phrases used in daily discourse.

Stylometric Analysis

To measure the complexity and richness of the language, we calculated three key metrics for each period:

1. **Average Sentence Length:** A proxy for syntactic complexity.
2. **Average Word Length:** A proxy for lexical sophistication.
3. **Yule's K:** A measure of vocabulary richness that is robust to varying text lengths. It is calculated as $K = 10^4 \times \frac{S_2 - S_1}{S_1^2}$, where S_1 is the total number of words and S_2 is the sum of the squares of the frequencies of each word. A higher K indicates a richer, more varied vocabulary [4].

Originality and Appropriateness

This project deliberately integrates network analysis with linguistic techniques (TF-IDF, n-grams, and stylometrics) to interrogate both the structural (who communicates with whom) and linguistic (what topics and styles are used) aspects of Founders' correspondence. The chosen methods are appropriate because they target the research questions: TF-IDF and n-grams detect changing vocabulary and rhetorical patterns, while stylometry captures the evolution of formal style across periods. The network analysis situates those linguistic features in a social structure, enabling historical interpretation beyond text-only analyses.

The implementation contains several decisions important for a reproducible and technically robust analysis.

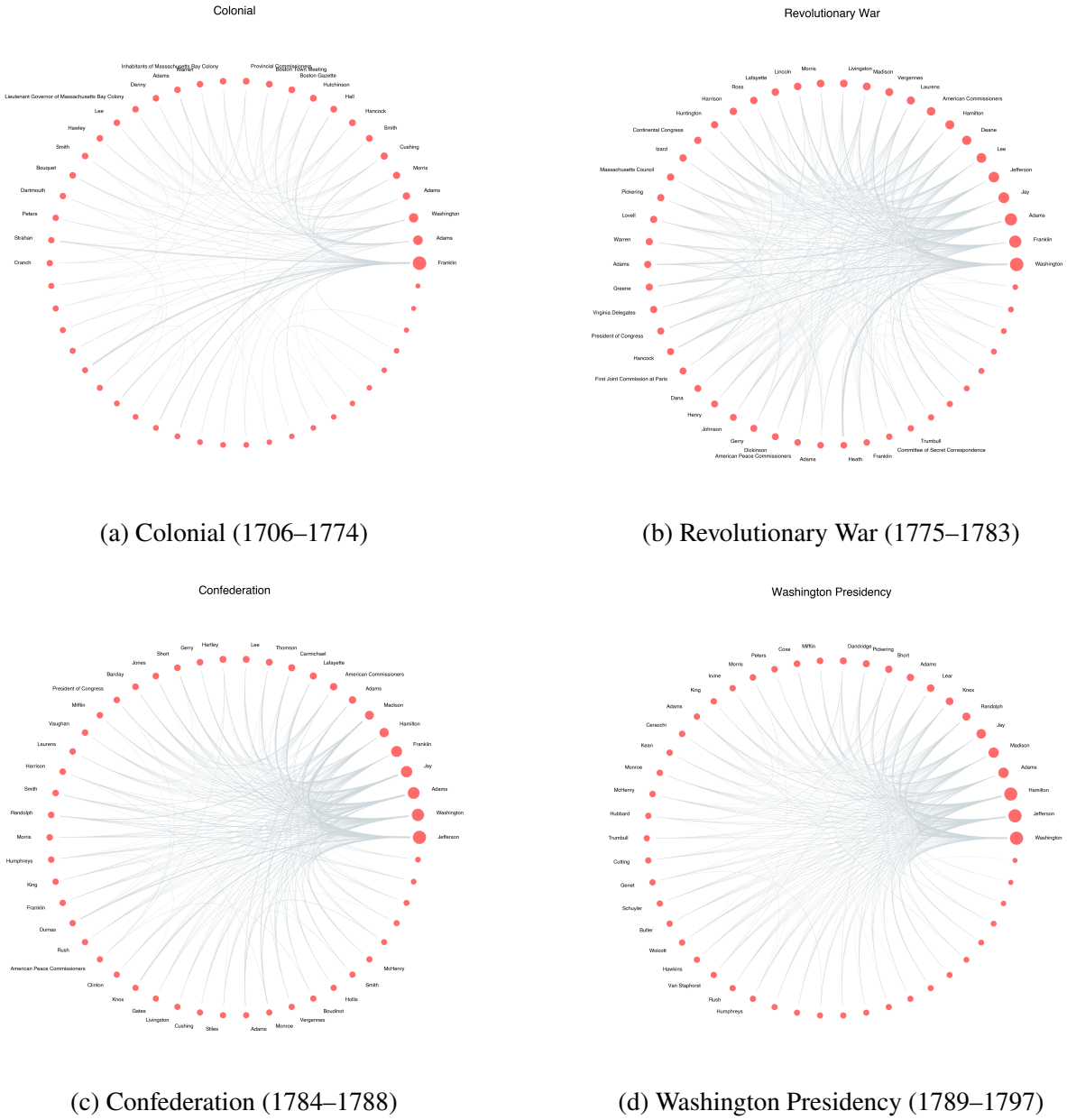
- **Document Unit for TF-IDF:** To capture period-specific vocabulary, the corpus is aggregated by period—this reduces noise at the single-letter level and sharpens the contrast between eras.
- **Preprocessing Rationale:** Using the WordNet lemmatizer and a custom stopword list aims to minimize archaic epistolary noise while preserving content-bearing tokens (e.g., place and person names). We opted to remove non-alphabetic characters to unify variant spellings (e.g., hyphenated forms), accepting the trade-off that editorial metadata sometimes uses abbreviations.
- **Scale and Efficiency:** The choice of JSON Lines and simple streaming preprocessing allows working with a large corpus on commodity hardware. Slurm/cluster usage minimizes the fragility of downloading and reduces run-time interruptions.
- **Sensitivity and Robustness:** Where appropriate, we cross-validate findings by comparing TF-IDF keywords with high-frequency n-grams and by manual inspection of outliers to identify OCR or metadata bleed.

Results

The computational analysis of the Founding Fathers' correspondence reveals distinct temporal patterns across network structure, vocabulary usage, and stylistic evolution. These findings map closely to the major political and personal phases of the founders' lives, from the colonial era through the early republic to their final years.

Network Analysis

The visualization of the correspondence network across eight historical periods (Figures 1 and 2) illustrates a fundamental structural transformation in the "Republic of Letters". Quantitative metrics for these networks are provided in Appendix C (Table 4).



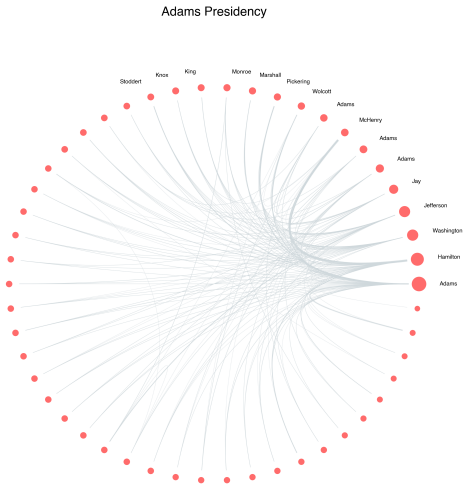
(a) Colonial (1706–1774)

(b) Revolutionary War (1775–1783)

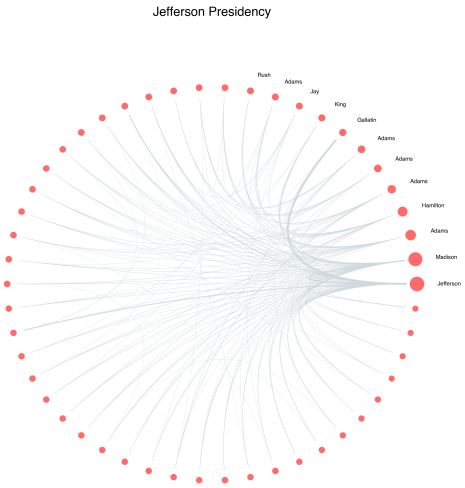
(c) Confederation (1784–1788)

(d) Washington Presidency (1789–1797)

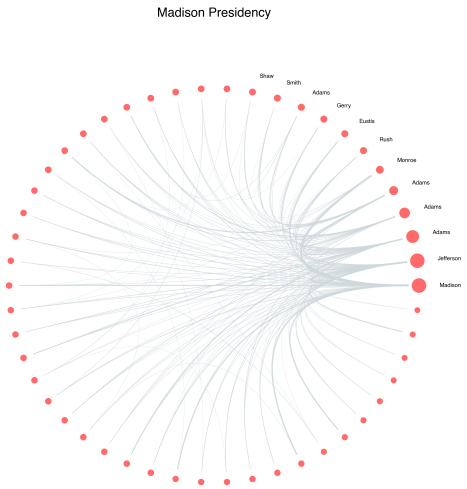
Figure 1: Evolution of the Correspondence Network



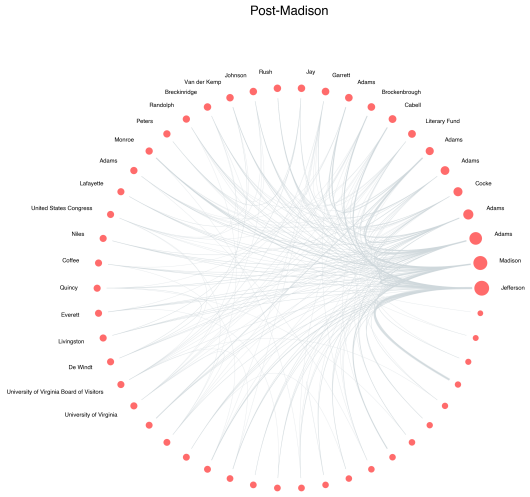
(a) Adams Presidency (1797–1801)



(b) Jefferson Presidency (1801–1809)



(c) Madison Presidency (1809–1817)



(d) Post-Madison (1817–1836)

Figure 2: Evolution of the Correspondence Network

The Colonial Era is characterized by fragmentation. The graph shows distinct, loosely connected clusters representing regional elites (e.g., the Virginia planters vs. the Boston intellectuals) with no single dominant center. Benjamin Franklin holds the highest degree (922), reflecting his unique position as a colonial agent in London connecting disparate groups.

The Revolutionary War forces a dramatic centralization and expansion. The network size more than triples to 5,079 nodes. It adopts a ”star topology” or ”hub-and-spoke” model, with Washington as the undisputed central node (degree 3,056), mediating information between Congress, the Army, and the states. This is the visual signature of a command economy of information.

The Confederation Period shows a partial relaxation of this centrality, with new nodes of influence appearing (e.g., Jefferson and Adams in Europe), reflecting the diplomatic dispersion of the era.

The Presidential Eras (Washington through Madison) see the re-emergence of the ex-

ecutive branch as the gravitational center. Notably, Jefferson's presidency exhibits the highest individual degree centrality in the entire corpus (4,391), surpassing even Washington's wartime command, indicating an immense personal investment in managing the administration via correspondence.

Finally, the Post-Madison network shows a "retirement diffusion." While key figures like Jefferson remain central due to their vast correspondence (degree 2,477), the network structure loosens, reflecting a shift from administrative command to intellectual exchange (the "University phase").

TF-IDF Analysis

Term Frequency-Inverse Document Frequency (TF-IDF [1]) analysis highlights the unique vocabulary defining each historical period. The results (see Appendix A) show a clear trajectory from personal and local concerns to national administration and finally to Durham Universityal legacy.

- **Colonial (1706–1774):** The vocabulary is intensely local and personal. Terms like "doeg" (Doeg Run) and "muddy hole" reflect Washington's focus on land surveying and estate management. "Cravenstreet" points to Benjamin Franklin's London residence, while "teedyuscung" indicates specific, localized diplomatic engagements with Native American leaders, contrasting sharply with the abstract "Indian affairs" of later periods.

- **Revolutionary War (1775–1783):** The discourse shifts abruptly to military logistics. Top terms include "middlebrook", "peekskill", and "pompton"—strategic encampments that dominated the writers' daily reality

- **Confederation (1784–1788):** Vocabulary bifurcates between the domestic and the diplomatic. We see "dogue" and "plowed" (Washington returning to Mount Vernon) alongside "calonne" and "thulemeier" (Jefferson and Adams in Europe).

- **Washington Presidency (1789–1797):** The language becomes administrative and Durham Universityal. Terms like "assignats" (French revolutionary currency) and "genest" (Citizen Genêt) mark the intrusion of global economic and political instability into the fragile new republic.

- **Adams Presidency (1797–1801):** The Quasi-War with France drives the vocabulary. "Artillerists", "talleyrand", and "toussaint" (L'Ouverture) rise to prominence, indicating a fixation on external threats and the Haitian Revolution.

- **Jefferson Presidency (1801–1809):** The focus shifts to continental expansion. "Pichon", "turreau", and "yrujo" (diplomats) are key, alongside "natchitoches", reflecting the administrative context of the Louisiana Purchase.

- **Madison Presidency (1809–1817):** The War of 1812 dominates. "Napoleon" is the top term, but economic terms like "merino" (sheep) also appear, signaling the drive for domestic industrial independence (wool production) necessitated by the British blockade.

- **Post-Madison (1817–1836):** The vocabulary turns remarkably academic. Terms like "dormitory", "bursar", "dunglison" (a professor recruited by Jefferson), and "bonnycastle" dominate.

Stylometric Analysis

Beyond vocabulary, the structure of the language evolves significantly (see Appendix B, Table 3).

- **Syntactic Complexity:** Average sentence length jumps from 18.8 words in the Colonial period to 26.4 words during the Revolutionary War. This 40% increase suggests that the exigencies of war and statecraft required a more complex, qualified, and precise mode of expression than private business correspondence.
- **Lexical Richness:** Yule's K, a measure of vocabulary diversity, peaks at 119.5 during the Washington Presidency. This is the highest value in the entire corpus. It suggests that the task of inventing the American presidency required an unprecedented expansion of the political lexicon—the Founders were literally finding new words to describe a new form of government.

N-gram Analysis

N-gram analysis (Appendix B) corroborates the TF-IDF findings while revealing structural changes in language.

- **The Disappearance of the Local:** In the Colonial period, top bigrams are specific and local, such as "muddy hole" (a farm on Washington's estate) and "doeg run". These vanish entirely in later periods, replaced by abstract political entities.
- **The Rise of the Nation:** The bigram "united state" is virtually absent in the Colonial period but becomes the dominant political entity from the Revolutionary War onwards.
- **Durham Universityalization:** The Revolutionary War is characterized by "head quarter", "court martial", and "commander chief"—Durham Universityal terms that replace the personal salutations of the earlier era.

Critical Evaluation

Methodological Assumptions and Limitations

The models rely on several key assumptions. First, Network Analysis assumes that the frequency of correspondence equates to the strength of a relationship. This ignores the qualitative nature of the letters; a single long, intimate letter may signify a stronger bond than ten brief administrative notes. Second, TF-IDF assumes that term frequency correlates with thematic importance. However, in 18th-century epistolary style, formulaic politeness (e.g., "your most obedient humble servant") is highly frequent but semantically empty. While we mitigated this with a custom stopword list, some noise inevitably remains.

Furthermore, the dataset is subject to **survival bias**, as not all historical correspondence has been preserved, and **selection bias**, as Founders Online focuses on prominent figures ("Great Man" history). As noted in the lecture material regarding n-gram assumptions, **OCR errors and metadata artifacts** significantly impact results. For instance, the appearance of "hamiltonsecy" in the TF-IDF top terms reveals that editorial annotations (e.g., "Hamilton, Secy.") were not fully separated from the body text during the scraping process. Similarly, spelling variations (e.g., "Doeg Run" vs. "Dogue Run") fragment the counts for single concepts, a common issue in pre-standardized 18th-century English that simple lemmatization cannot fully resolve.

Comparison with Historical Narratives

Despite these limitations, the computational results align remarkably well with established historical narratives, such as Gordon Wood's account of the early republic [2], validating the "distant reading" approach.

Synthesis of Methods

By employing multiple methods, we can observe how structural power correlates with linguistic change. The Network Analysis shows a centralization of power around Washington during the war and presidency. The tripling of the network size during the Revolutionary War (1,454 to 5,079 nodes) coincides perfectly with the 40% increase in sentence length, suggesting that managing a larger, more complex network required more complex linguistic structures. This structural shift is mirrored in the TF-IDF results, which transition from diverse local terms to a unified, administrative vocabulary. However, the methods also diverge in revealing ways: while Stylometry shows a relatively constant sentence length after the Revolutionary War spike, the vocabulary richness (Yule's K) peaks specifically during the Washington presidency.

This alignment suggests that while the models may be noisy at the micro-level (individual words), they are robust at the macro-level (historical trends). The Network Analysis shows a centralization of power around Washington during the war and presidency. This structural shift is mirrored in the TF-IDF results, which transition from diverse local terms to a unified, administrative vocabulary. However, the methods also diverge: while Stylometry shows a relatively constant sentence length, the vocabulary richness (Yule's K) peaks during the Washington presidency, suggesting that while the *complexity* of sentence structure remained stable, the *lexical diversity* required to build a nation increased significantly. This multi-modal approach provides a more nuanced view than any single method could offer.

Evaluation Methodology and Validation

To satisfy the assignment's requirement for critical evaluation and to validate claims derived from computational models, several evaluation strategies were used:

- **Manual verification** of top terms and sample letters: a historian-authored spot-check of the top TF-IDF terms in each period ensured that salient words correspond to the contexts inferred (e.g., military or administrative contexts).
- **Sensitivity analyses**: To test the stability of TF-IDF rankings, we re-ran TF-IDF with and without the custom stopword list and with a minimal lemmatization pipeline to measure term stability across preprocessing choices.
- **Metadata consistency checks**: Indexes for place names, person names, and editorial notes were checked for bleed-through (e.g., "hamiltonsecy"), and where these occur they were handled by either removal or separate analysis and flagged in the appendices.

Conclusions

This study demonstrates the utility of computational methods in enriching historical research. The analysis confirms that the Founding Fathers' correspondence patterns were not static but

evolved dynamically in response to the changing political landscape. We observed a clear trajectory: from the personal and local concerns of the Colonial era, through the urgent military logistics of the Revolutionary War, to the complex administrative statecraft of the early Republic, and finally to a reflective focus on education and legacy in the Post-Madison years.

A key finding is the peak in vocabulary richness during the Washington Presidency, suggesting that the task of establishing a new national government required a uniquely complex and varied lexicon. These findings underscore the value of "distant reading" as a complement to traditional historical scholarship, offering a quantitative backbone to qualitative historical narratives and revealing the linguistic footprint of nation-building.

Further Work and Reflection

To address remaining limitations and to deepen interpretation, a number of further steps could refine and extend this project:

- **Formal hypothesis testing:** Explicit testing—for example, whether sentence length distributions differ significantly between periods—would strengthen claims about stylistic change.
- **Improved OCR and spelling normalization:** Preprocessing improvements could reduce term fragmentation (e.g., Doeg/Dogue) and editorial artifacts (e.g., hamitonsecy). This includes custom spelling unification or editorial cleaning.
- **Authorship attribution and contextualization:** Applying supervised classification to determine if stylistic changes are driven by authorship differences or by broader social changes would refine causal claims.

References

References

[1] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.

[2] Wood, G. S. (2009). Empire of Liberty: A History of the Early Republic, 1789-1815. Oxford University Press.

[3] Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

[4] Yule, G. U. (1944). The Statistical Study of Literary Vocabulary. (Cambridge University Press).

[5] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.

[6] Hagberg, A., Schult, D., & Swart, P. (2008). Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference (SciPy2008).

[7] Bird, S., Loper, E., & Klein, E. (2009). Natural Language Toolkit (NLTK). http://www.nltk.org

A TF-IDF Results

Table 1: Top 10 TF-IDF Terms by Historical Period

Colonial	Rev. War	Confederation	Washington	Adams	Jefferson	Madison	Post-Madison
doeg	artaud	dogue	hamiltonsecy	subdistricts	pichon	napoleon	dormitory
ageneral	middlebrook	williamos	ustates	artillerists	cevallos	batture	dunglison
eund	zullen	deral	assignats	mchenry	turreau	merino	montezillo
dismd	tyonderoga	mansn	hamiltonsecretary	dittovice	reibelt	correa	bursar
patd	sartine	sjl	jaudenes	rivardi	osage	gothenburg	montpr
empld	eenige	tabacs	ischem	talleyrand	yrujo	ticknor	brockenbrough
dind	peekskill	shipton	chiappe	bewell	ustates	bonaparte	raggi
waterson	pompton	deslon	clingman	dittojohn	polygraph	bankhead	capitels
cravenstreet	hazen	grd	whitting	subdistrict	gallatin	nonag	ticknor
magowan	billingsport	doradour	genest	tousard	gavino	dallas	bonnycastle

B N-gram Results

Table 2: Top 10 Bigrams by Historical Period

Colonial	Rev. War	Confederation	Washington	Adams	Jefferson	Madison	Post-Madison
new york	head quarter	united state	united state	united state	united state	united state	united state
home day	new york	new york	new york	new york	new york	new york	new york
little wind	united state	noon night	secretary treasury	secretary war	new orleans	john adam	john adam
benjamin franklin	respect excellency	morning noon	secretary state	mount vernon	department state	respect jefferson	james madison
george washington	que vous	muddy hole	president united	john adam	james madison	esteem respect	respect jefferson
every thing	every thing	mount vernon	house representative	major general	respect jefferson	accept assurance	accept assurance
went away	beg leave	dogue run	obedt servt	general hamilton	beg leave	every thing	esteem respect
muddy hole	servt washington	thermometer morning	servt jefferson	obedt servt	every thing	jefferson monticello	jefferson monticello
mount vernon	obedt servt	que vous	treasury department	war department	president united	james madison	thomas jefferson
fort cumberland	west point	every thing	every thing	president united	secretary state	beg leave	university virginia

C Stylometric Results

Table 3: Stylometric Analysis by Period

Period	Letters	Avg Word Len	Avg Sent Len	Yule's K (Richness)
Colonial	16,151	4.39	18.81	93.11
Revolutionary War	48,188	4.46	26.43	96.38
Confederation	17,810	4.43	21.72	97.44
Washington Presidency	27,156	4.51	23.02	119.51
Adams Presidency	13,570	4.49	23.55	109.19
Jefferson Presidency	29,499	4.51	24.02	110.65
Madison Presidency	15,477	4.48	26.53	106.81
Post-Madison	15,445	4.47	26.14	109.56

D Network Statistics

Table 4: Network Statistics by Historical Period

Period	Nodes	Edges	Density	Most Central Figure (Degree)
Colonial	1,454	2,211	0.0011	Benjamin Franklin (922)
Revolutionary War	5,079	9,755	0.0004	George Washington (3,056)
Confederation	2,396	4,454	0.0008	George Washington (1,343)
Washington Presidency	3,505	6,196	0.0005	George Washington (2,398)
Adams Presidency	1,994	3,388	0.0009	John Adams (1,081)
Jefferson Presidency	4,764	7,232	0.0003	Thomas Jefferson (4,391)
Madison Presidency	3,268	4,943	0.0005	James Madison (2,217)
Post-Madison	2,815	5,346	0.0007	Thomas Jefferson (2,477)