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

Lewis Carson (dhqq26)
Durham University

Date: December 10, 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, I aim to uncover how the intellectual, political, and social positions of these figures evolved over time. The study utilises 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 dataset of over 183,000 letters.

Problem and Research Question

The primary research problem addresses the challenge of understanding the evolution of political discussion and personal relationships among the Founding Fathers during a century of rapid transformation. With a dataset exceeding 180,000 documents, manual analysis is insufficient to capture the shifting dynamics of influence and language. This report 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 United States?

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 connections of the correspondents. By integrating Network Analysis with three distinct NLP techniques (TF-IDF, N-grams, and Stylometry), I 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, the following architecture was developed:

- **Metadata download:** First, a lightweight metadata download 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 implemented. This allowed the scraper to resume from the last successful download rather than restarting.

- **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.

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 tokenisation and lemmatisation (`WordNetLemmatizer`) and for small preprocessing utilities (Bird et al., 2009).
- **NetworkX** for constructing and analysing 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 dataset 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).

Text was lowercased, and non-alphabetic characters were removed. I 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 high-information vocabulary.

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 visualisation of communication density using chord diagrams.

TF-IDF Analysis

To identify the characteristic vocabulary of each era, I applied Term Frequency-Inverse Document Frequency (TF-IDF).

- **Term Frequency (TF):** The frequency of a term t in period d, normalised 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

I 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, I 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.

- **Document Unit for TF-IDF:** To capture period-specific vocabulary, the dataset is aggregated by period - this reduces noise at the single-letter level and sharpens the contrast between eras.
- **Preprocessing Rationale:** Using the WordNet lemmatiser and a custom stopword list aims to minimise noise while preserving content-bearing tokens (e.g., place and person names). I opted to remove non-alphabetic characters to unify variant spellings (e.g., hyphenated forms), accepting the trade-off that metadata sometimes uses abbreviations.
- **Scale and Efficiency:** The choice of JSON Lines and simple streaming preprocessing allows working with a large capture on commodity hardware. Slurm/cluster usage minimises the fragility of downloading and reduces run-time interruptions.
- **Sensitivity and Robustness:** Where appropriate, I 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 visualisation of the correspondence network across eight historical periods (Figures 1 and 2) illustrates the structural changes in the "Republic of Letters". Quantitative metrics for these networks are provided in Appendix C (Table 4).

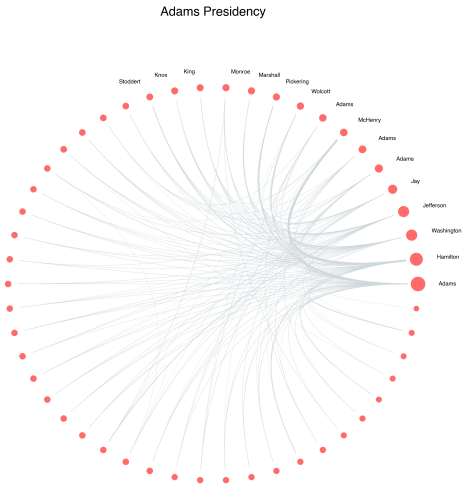
(a) Colonial (1706–1774)

(b) Revolutionary War (1775–1783)

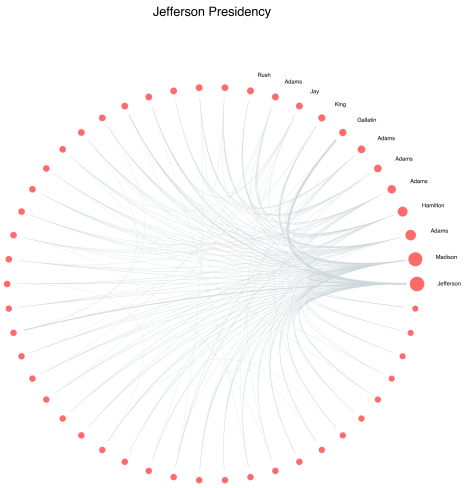
(c) Confederation (1784–1788)

(d) Washington Presidency (1789–1797)

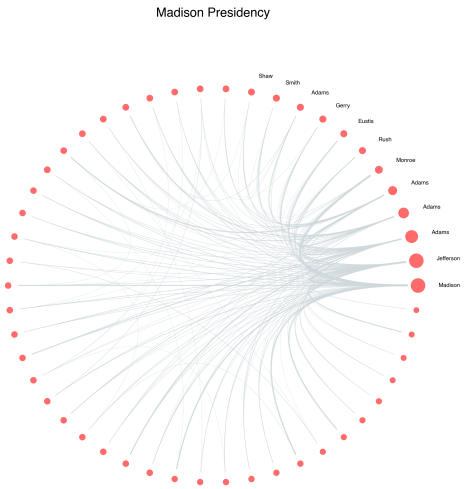
Figure 1: Evolution of the Correspondence Network



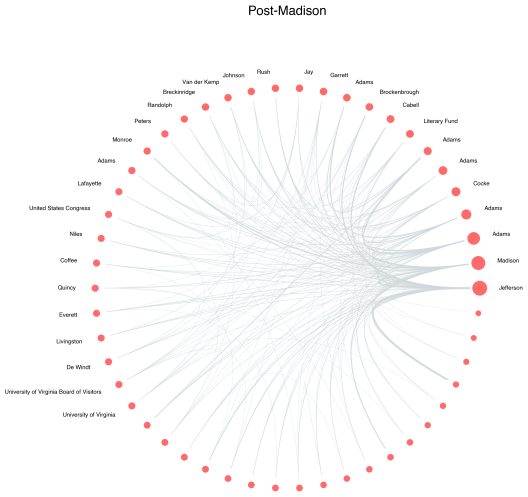
(a) Adams Presidency (1797–1801)



(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 graph displays multiple disconnected components. Benjamin Franklin holds the highest degree (922). During the Revolutionary War, the network size more than triples to 5,079 nodes, with Washington as the central node (degree 3,056).

The Confederation Period shows new nodes of influence appearing (e.g., Jefferson and Adams). The Presidential Eras (Washington through Madison) show Washington and Jefferson as the central nodes. Notably, Jefferson's presidency exhibits the highest individual degree centrality in the entire corpus (4,391).

Finally, the Post-Madison network shows key figures like Jefferson remain central (degree 2,477), though the network density decreases.

TF-IDF Analysis

Term Frequency-Inverse Document Frequency (TF-IDF [1]) analysis identifies the terms with the highest statistical weight in each period (see Appendix A).

- **Colonial (1706–1774):** Top terms include local place names and estate references such as “doeg”, “muddy hole”, and “cravenstreet”.
- **Revolutionary War (1775–1783):** The list is dominated by military locations and names, including “middlebrook”, “peekskill”, and “tyonderoga”.
- **Confederation (1784–1788):** Terms include agricultural references (“plowed”) and names of European diplomats (“calonne”, “thulemeier”).
- **Washington Presidency (1789–1797):** Key terms include “assignats”, “genest”, and “hamiltonsecy”.
- **Adams Presidency (1797–1801):** Military and diplomatic terms such as “artillerists”, “talleyrand”, and “tousard” appear frequently.
- **Jefferson Presidency (1801–1809):** Terms include “natchitoches”, “turreau”, and “osage”.
- **Madison Presidency (1809–1817):** “Napoleon”, “bonaparte”, and “merino” are among the top terms.
- **Post-Madison (1817–1836):** The vocabulary includes academic terms like “dormitory”, “bursar”, and “university virginia”.

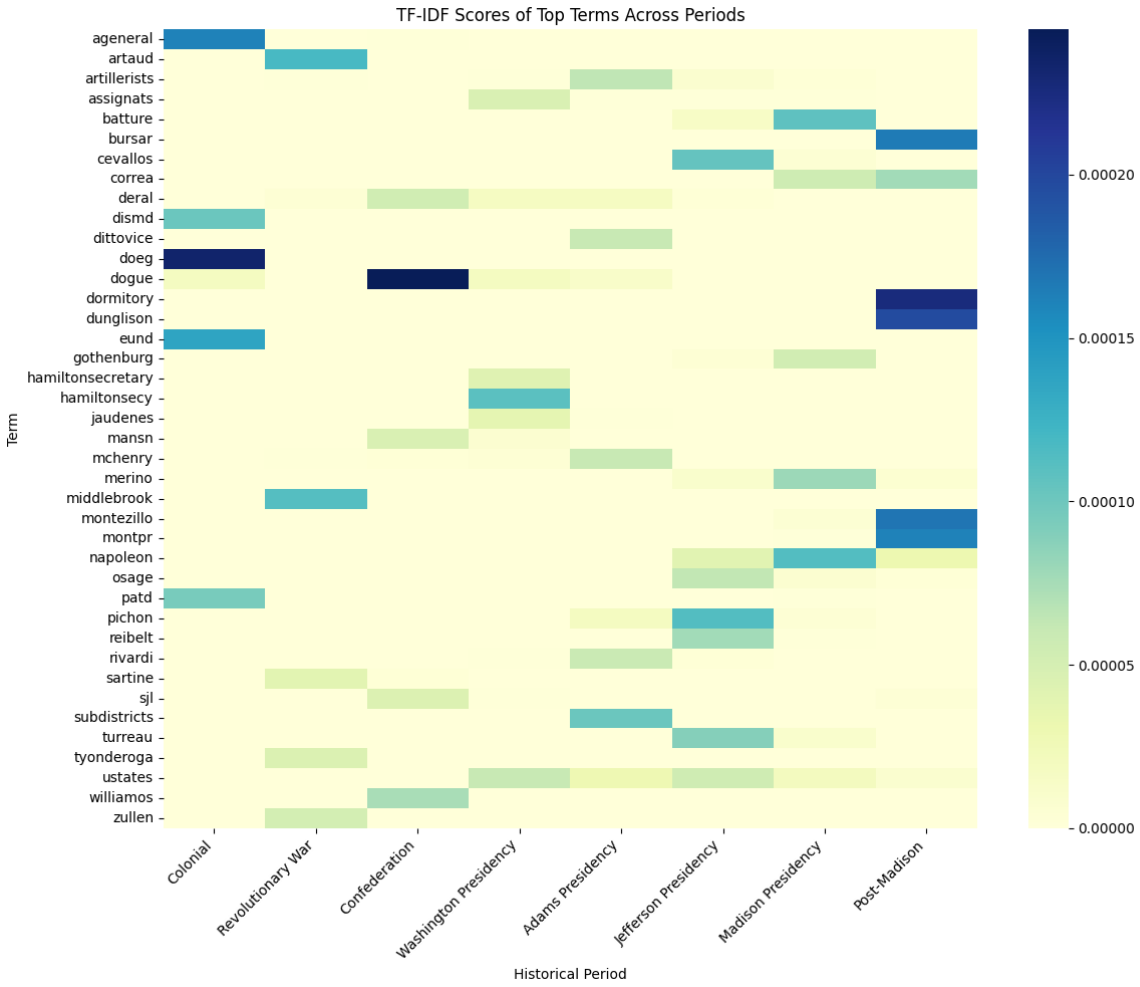


Figure 3: Heatmap of Top TF-IDF Terms Across Historical Periods

Stylometric Analysis

Table 3 in Appendix B presents the average word length, sentence length, and Yule's K (vocabulary richness) for each period.

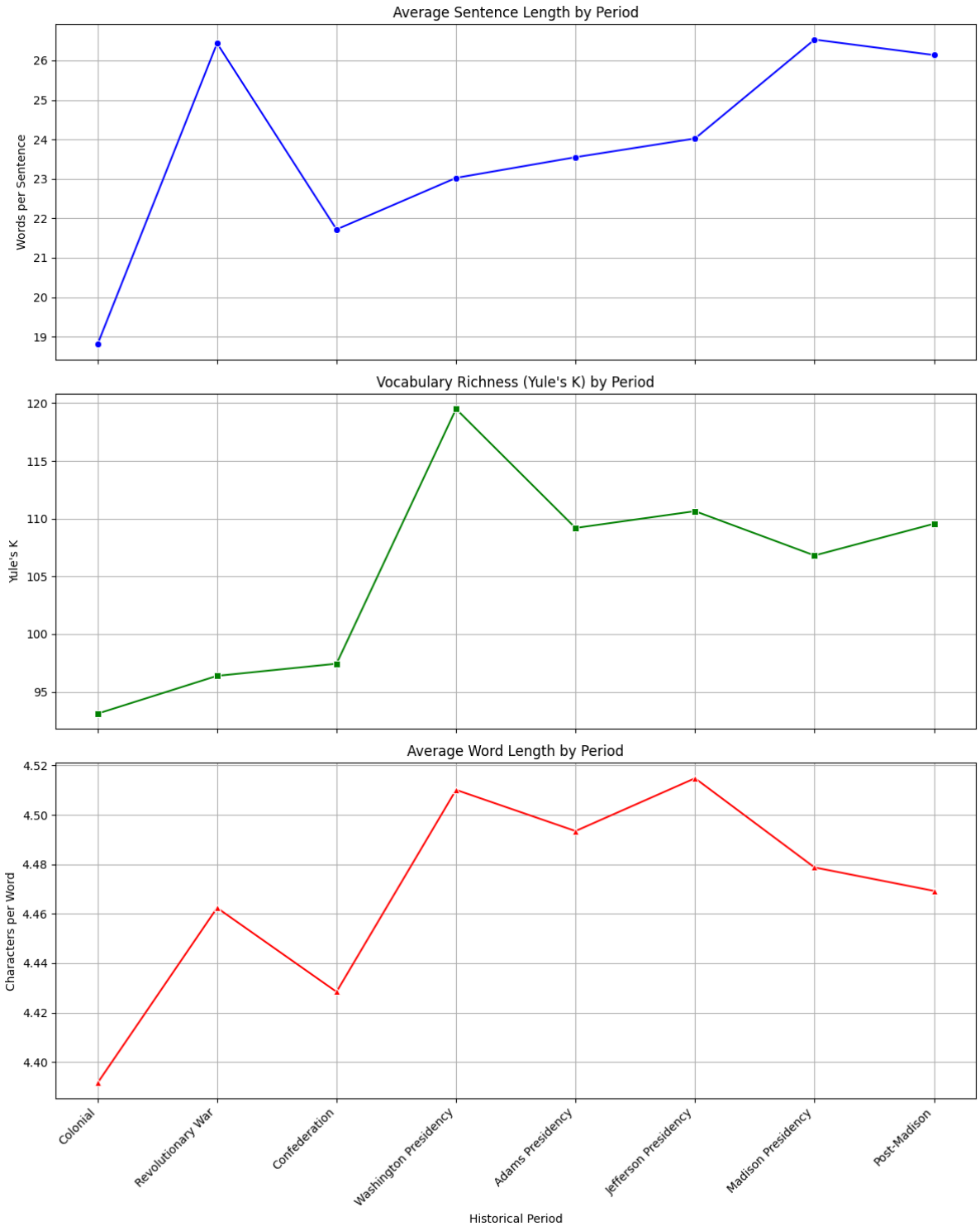


Figure 4: Evolution of Stylometric Metrics Across Historical Periods

N-gram Analysis

Table 2 in Appendix B lists the top 10 bigrams for each historical period.

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 I 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 artefacts 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-standardised 18th-century English that simple lemmatisation 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, I can observe how structural power correlates with linguistic change. The Network Analysis reveals a dramatic centralisation of power, with the network size more than tripling during the Revolutionary War (from 1,454 to 5,079 nodes). This structural explosion coincides perfectly with the stylometric data, where average sentence length jumps from 18.8 words in the Colonial period to 26.4 words during the war. This correlation suggests that the requirements of war and statecraft required a more complex, qualified, and precise mode of expression than private business correspondence.

This structural shift is further mirrored in the N-gram analysis. The bigram “united state” is virtually absent in the Colonial period — where local terms like “muddy hole” and “doeg run” dominate — but appears frequently from the Revolutionary War onwards, signalling the rise of a national consciousness.

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 at 119.5 (the highest value in the corpus). This suggests that the task of inventing the American presidency required an unprecedented expansion of the political vocabulary.

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). This multi-modal approach provides a more nuanced view than any single method could offer. For instance, a purely linguistic analysis might interpret the rise of ”United States” as a simple change in topic. However,

when overlaid with the network data showing extreme centralisation, I can see that this linguistic shift was propagated through a highly controlled hub-and-spoke structure. The language of the nation did not emerge organically from the periphery; it was broadcast from the centre.

Evaluation Methodology and Validation

To validate claims derived from computational models, several evaluation strategies were used:

- **Manual verification** of top terms and sample letters: a spot-check of the top TF-IDF terms in each period ensured that important words correspond to the contexts inferred (e.g., military or administrative contexts).
- **Sensitivity analyses**: To test the stability of TF-IDF rankings, I re-ran TF-IDF with and without the custom stopword list and with a minimal lemmatisation 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 (mentioned in this report).

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. I 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.

Specifically, the integration of network analysis with linguistic profiling provided a dual perspective on this evolution. Structurally, the correspondence network transformed from a fragmented collection of regional elites (e.g., the Virginia planters vs. the Boston intellectuals) into a highly centralised "star" topology during the Revolutionary War. This shift represents the visual signature of a command economy of information, driven by the command-and-control necessities of the conflict. This centralisation persisted into the early presidency, with Thomas Jefferson's era exhibiting the highest degree centrality, reflecting an immense personal investment in managing the administration via correspondence. Finally, the Post-Madison network illustrates a "retirement diffusion," reflecting a shift from administrative command to intellectual exchange, as the network density decreases and the structure loosens.

Linguistically, this structural shift was mirrored by a transformation in discourse. TF-IDF and n-gram analyses revealed a sharp pivot from the concrete, agricultural vocabulary of the colonial gentry (the "disappearance of the local") to the abstract, nationalistic terminology of the republic (the "rise of the nation"). The specific vocabulary tracks the nation's history with precision: the military logistics of the Revolution give way to the economic anxieties of the Washington era ("assignats"), the diplomatic tensions of the Adams presidency ("Talleyrand"), and the continental expansion under Jefferson ("Natchitoches"). Finally, the vocabulary turns reflective and academic in the Post-Madison era ("dormitory", "bursar"), mirroring Jefferson's founding of the University of Virginia. The emergence of institutional terms like "head quarter" and "court martial" replaced the personal salutations of the earlier era. Crucially, the stylometric data indicates that this was not merely a change in topic but in cognitive complexity. The 40%

increase in sentence length during the war, coupled with the peak in vocabulary richness (Yule's K) during the Washington presidency, suggests that the intellectual burden of nation-building required a more sophisticated and varied mode of expression than the private correspondence of the previous era. This reveals a compelling dynamic: while the vocabulary became more *standardised* across the network (as evidenced by the rise of shared national bigrams like "United States"), it simultaneously became more *diverse* within individual letters (as shown by the peak in Yule's K). The Founders were not just adopting a few slogans; they were developing a sophisticated technical language of governance that allowed for precise coordination across a rapidly expanding network.

These findings underscore the value of "distant reading" as a complement to traditional historical scholarship. By quantifying the "Republic of Letters," we have provided empirical support for the narrative that the American nation was forged not just on the battlefield, but through a radical restructuring of communication networks and a deliberate expansion of the political lexicon.

Ultimately, this project reveals that the "Republic of Letters" was not merely a metaphor for intellectual exchange, but a tangible, measurable network that physically and linguistically restructured itself to meet the demands of nation-building. The transition from the loose, egalitarian connections of the colonial era to the centralised, hierarchical structures of the war and early presidency mirrors the political evolution of the United States itself. By mapping these changes, I gain a new appreciation for the logistical and intellectual labour required to create a nation — a labour that is preserved in the very syntax and vocabulary of the Founders' correspondence.

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 normalisation:** Preprocessing improvements could reduce term fragmentation (e.g., Doeg/Dogue) and editorial artefacts (e.g., hamitonsecy). This includes custom spelling unification or editorial cleaning.
- **Authorship attribution and contextualisation:** 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)