Text Analysis Final Project - Topic Modeling: The Federalist Papers

DS 745 - Chris Austin

University of Wisconsin – Data Science

Table of Contents

[Dataset Description 1](#_Toc531261810)

[Problem Description 1](#_Toc531261811)

[Describe Methodology and Identify Relevant Variables 1](#_Toc531261812)

[Text Cleaning 1](#_Toc531261813)

[Data Frame and Corpus Generation 2](#_Toc531261814)

[Statistical Methods, Important Variables/Structures, and Usage of Results 3](#_Toc531261815)

[Visually Describe Textual Data/Findings Along with Discussion 3](#_Toc531261816)

[Consideration of - and Findings from - a Three-Topic Structure 3](#_Toc531261817)

[Three-Topic Evaluation 4](#_Toc531261818)

[Consideration of - and Findings from - a Four-Topic Structure 5](#_Toc531261819)

[A Comparison of 3-Topic and 4-Topic Mappings 6](#_Toc531261820)

[Four-Topic Evaluation 6](#_Toc531261821)

[Effectiveness of the Four-Topic Structure 7](#_Toc531261822)

[Discussion of Relevant Graphic Output 7](#_Toc531261823)

[Topic 1 8](#_Toc531261824)

[Topic 2 8](#_Toc531261825)

[Topic 3 9](#_Toc531261826)

[Topic 4 9](#_Toc531261827)

[Theme Differentiation 10](#_Toc531261828)

[Possibilities for Further Refinement 10](#_Toc531261829)

[References 11](#_Toc531261830)

# Dataset Description

The dataset under consideration for this project is the compilation of 85 essays written between October 1787 and April 1788 on the formation of the Constitution of the United States, best-known as the Federalist Papers. Authored by then-future fourth President James Madison, first Secretary of the Treasury Alexander Hamilton, and first Supreme Court Chief Justice John Jay, they lay out the philosophical basis for Federal separation of powers and their appropriate boundaries, as well as the roles of state governments, the rights of citizenry, and philosophical guidelines for relations with foreign powers. Twelve of the papers’ authorships are disputed, and have been noted as such in the dataset.

All signed with the *nom-de-plume* “Publius”, the authors envisioned their work as replicating the intentions of Publius Valerius Publicola. Publicola (an affectation granted by the citizens of Rome, meaning “friend of the people”) was a Roman patriot, general, and statesman who lived in the sixth century B.C.E. and who, according to Plutarch’s *Lives*, saved the early Roman republic several times from tyranny and military subjugation.

The dataset is available as a .txt file at the Project Gutenberg website (found in the References section) and can be accessed in several different formats.

# Problem Description

The problem approached in this project is the summarization of the papers into several concise topics using Linear Discriminant Analysis (LDA). The initial goal is to pick out ten words in each of three topic groups, and to tag some sort of basic theme to them. Of the themes considered in the Federalist Papers, which stand out? At that point, can a fourth additional topic/theme be distinguished after that, and does it note a ‘better’ set of topic descriptors as compared to three?

Then, once the number of topics to be considered are finalized, are the papers that most exemplify those topics reasonably representative of that topic? The conclusions of this paper include a brief discussion regarding each specific topic and the five papers with the highest tendency towards it.

# Describe Methodology and Identify Relevant Variables

## Text Cleaning

The first step in the process was to load the text file into R and to immediately remove the metadata related to the Gutenberg Project as well as legal disclaimers relating to usage in the public domain. The line location of the first chapter was noted, as well as the end of the last paper. Text between those two locations were kept and the rest discarded.

Next, a list of chapter boundaries was created. Each chapter starts with a capitalized ‘FEDERALIST’ along with the paper number, making the bookmarking of these rather simple. Complicating the issue, however, was the inclusion of two versions of Federalist 70. One was programmatically removed to avoid repetition in some of the word counts and to avoid impacts upon topic representation.

After another visual inspection of the text, additional metadata required deletion. Each of the papers noted the governing publication for which they were intended (the Independent Journal, Daily Advertiser, New York Packet, and McClean’s). A subsequent target of deletion was the identical preamble (“To the People of the state of New York”) present in each.

The placing of two hyphens without intervening spaces in over 30 locations gave rise to combinations of non-sensical words. These were replaced with spaces so later processes could better parse the words.

## Data Frame and Corpus Generation

The text file was thus in a state that allowed for the creation of a list object, giving rise to a data frame object whose rows correspond to each of the Federalist papers. This data frame was the origin of the text corpus from which the analysis was conducted, and the authors were added as one of the columns of the data frame, accompanying the paper’s text, paper number, and a pre-formatted label in another column.

The words in the corpus were transformed to lowercase and purged of whitespace, numbers, and punctuation. Common English words were removed that lent no weight to the analysis. In each of the papers the author’s last name leads off the paper and is thus removed. Their closing signature as “Publius” is likewise removed.

To further prepare the text for topic analysis, the words were ‘stemmed’ by taking them down to their root. For example, ‘divide’, ‘divided’, and ‘dividing’ are shortened to ‘divid’. This keeps words with different endings from being counted as separate components of potentially different themes. Once cleaned, one of the smaller papers (Federalist 13) looks like this:

"advantag union respect economi govern connect subject revenu may proprieti consid economi money save one object may use appli anoth will much less drawn pocket peopl state unit one govern will one nation civil list support divid sever confederaci will mani differ nation civil list provid princip depart coextens necessari govern whole entir separ state thirteen unconnect sovereignti project extravag replet danger mani advoc idea men specul upon dismember empir seem general turn toward three confederaci one consist four northern anoth four middl third five southern state littl probabl greater number accord distribut confederaci compris extent territori larger kingdom great britain wellinform man will suppos affair confederaci can proper regul govern less comprehens organ institut propos convent dimens state attain certain magnitud requir energi govern form administr requisit one much greater extent idea admit precis demonstr rule can measur momentum civil power necessari govern given number individu consid island britain near commensur suppos confederaci contain eight million peopl reflect upon degre author requir direct passion larg societi public good shall see reason doubt like portion power suffici perform task societi far numer civil power proper organ exert capabl diffus forc great extent can manner reproduc everi part great empir judici arrang subordin institut supposit confederaci state like divid requir govern less comprehens one propos will strengthen anoth supposit probabl present us three confederaci altern general union attend care geograph commerci consider conjunct habit prejudic differ state shall led conclud case disunion will natur leagu two govern four eastern state caus form link nation sympathi connect may certainti expect unit new york situat never unwis enough oppos feebl unsupport flank weight confederaci obvious reason facilit access new jersey small state think frontier opposit still power combin appear obstacl admiss even pennsylvania strong induc join northern leagu activ foreign commerc basi navig true polici coincid opinion disposit citizen southern state various circumst may think much interest encourag navig may prefer system give unlimit scope nation carrier well purchas commod pennsylvania may choos confound interest connect advers polici must event frontier may deem consist safeti expos side turn toward weaker power southern rather toward stronger power northern confederaci give fairest chanc avoid flander america whatev may determin pennsylvania northern confederaci includ new jersey likelihood one confederaci south state noth can evid thirteen state will abl support nation govern better one half one third number less whole reflect must great weight obviat object propos plan found principl expens object howev come take nearer view will appear everi light stand mistaken ground addit consider plural civil list take view number person must necessarili employ guard inland communic differ confederaci illicit trade time will infal spring necess revenu also take view militari establish shown unavoid result jealousi conflict sever nation state divid shall clear discov separ less injuri economi tranquil commerc revenu liberti everi part"

A Document Term matrix was then created from the data frame. It results in a very sparsely populated matrix that contains the number of times a word is used in each document:

Docs abolit abort abound abraham abreg (...)

4 0 0 1 0 0 ...

9 1 0 0 0 0 ...

10 1 0 0 0 0 ...

18 1 0 0 0 0 ...

19 0 1 0 0 1 ...

22 0 1 0 0 0 ...

38 1 0 0 0 0 ...

43 0 0 1 0 0 ...

64 0 0 1 0 0 ...

73 0 0 0 1 0 ...

81 1 0 0 0 0 ...

83 3 0 0 0 0 ...

This allows for summation of word totals that can be used as part of the LDA routines but are calculated here manually for illustrative purposes.

abolit abort abound abraham abreg (...)

8 2 3 1 1 ...

## Statistical Methods, Important Variables/Structures, and Usage of Results

LDA is used here to reduce the number of unique words and their counts into groupings (topics) and from their attributes (the words themselves and their summed usage). The topics become defined by their frequency of usage together when word counts are summarized at a multi-paper topical level. Topics are programmatically refined by adding or removing attributes (the word and its topic-level sum) that maximize topics’ differentiation.

As an analogy, consider several similar varieties of flower: they look generally the same in a meadow from a distance. Upon closer inspection, one can find groups of similarities that cause distinct classification into separate species. Species designations are carved out of attributes that ‘cluster’ together but are distinct from other clusters. Here, words and the sum of their usage across topics are the attributes being considered.

Each of the 85 papers are then categorized into the number of topics desired (first three, then four). Each topic has the top ten words extracted that, when viewed, help to identify a theme that can be applied to the topic. Words are not necessarily unique to a theme, they can occasionally repeat across themes.

In the next section, we will analyze the top five papers in each topic (once we have settled on three or four) to see how well the subject matter matches up with the summarized theme. To determine a topic’s “top five papers”, a table of probabilities is available as part of the LDA function that signifies the chance that the document should be attached to any given topic.

In the closing section, we will also present a graphical representation of words’ relationship to the topic at hand using the LDAvis package in R.

# Visually Describe Textual Data/Findings Along with Discussion

## Consideration of - and Findings from - a Three-Topic Structure

The top ten words in each of the three topics generated – and the number of papers tagged to that topic - were arranged as follows:

|  |
| --- |
| Topic 3.1 Topic 3.2 Topic 3.3 |
| "state" "nation" "will" |
| "power" "union" "govern" |
| "constitut" "state" "peopl" |
| "author" "must" "may" |
| "execut" "general" "state" |
| "case" "power" "interest" |
| "law" "can" "must" |
| "may" "foreign" "repres" |
| "legisl" "upon" "number" |
| "shall" "war" "differ" |
| "Paper Count" "Paper Count" "Paper Count" |
| "21" "24" "40" |

Putting one potential theme to each of these topics, the words in topic 3.1 look to focus on the law and the power of the state to execute them. Topic 3.2 is plainly about the role of the government in foreign relations and conflict. Topic 3.3 looks to be focused on the government’s role in handling differing interests. To summarize these ten-word topics to simple phrases, this is but one possible interpretation:

Topic 3.1: A Nation of Laws

Topic 3.2: Foreign Relations and External Conflict

Topic 3.3: Structure Amidst Differing Agendas

### Three-Topic Evaluation

To examine that interpretation, let us look at the top five papers assigned by topic probability as computed by LDA for the three topics selected. Each paper’s heading assists us in this assessment:

**Five highest probabilities in topic 3.1**

Paper pTopic1 pTopic2 pTopic3 N TopicDesc

Fed 82:Hamilton 0.782 0.084 0.135 1 A Nation of Laws

Fed 47:Madison 0.751 0.066 0.183 1 A Nation of Laws

Fed 32:Hamilton 0.733 0.131 0.136 1 A Nation of Laws

Fed 81:Hamilton 0.710 0.073 0.217 1 A Nation of Laws

Fed 83:Hamilton 0.661 0.098 0.242 1 A Nation of Laws

Federalist 82: The Judiciary Continued

Federalist 47: The Particular Structure of the New Government and the Distribution of Power Among Its Different Parts

Federalist 32: The Same Subject Continued (Concerning the General Power of Taxation)

Federalist 81: The Judiciary Continued, and the Distribution of the Judicial Authority

Federalist 83: The Judiciary Continued in Relation to Trial by Jury

This seems appropriate. Three of the five papers are directly related to the Judiciary branch. In regards to Federalist 32, there were few areas in which the founders were more eager to place legal boundaries on the government than in regards to the subject of taxation. To those of us in the modern day, the Boston Tea Party and the Stamp Act are but historical reference. To the founders, however, the prospect of lax controls on taxation likely elicited strong emotional reactions.

Federalist 47 is another matter. Its theme is a justification of the checks and balances accorded to the each of the three branches of government in order to curtail the excesses of the other two, rather than a totally separate relationship where no balances exist:

One of the principal objections inculcated by the more respectable adversaries to the Constitution, is its supposed violation of the political maxim, that the legislative, executive, and judiciary departments ought to be separate and distinct. In the structure of the federal government, no regard, it is said, seems to have been paid to this essential precaution in favor of liberty.

The several departments of power are distributed and blended in such a manner as at once to destroy all symmetry and beauty of form, and to expose some of the essential parts of the edifice to the danger of being crushed by the disproportionate weight of other parts.

**Five highest probabilities in topic 3.2**

Paper pTopic1 pTopic2 pTopic3 N TopicDesc

Fed 18:Madison 0.185 0.614 0.202 2 Foreign Relations and External Conflict

Fed 8:Hamilton 0.128 0.597 0.275 2 Foreign Relations and External Conflict

Fed 11:Hamilton 0.118 0.597 0.284 2 Foreign Relations and External Conflict

Fed 6:Hamilton 0.111 0.586 0.303 2 Foreign Relations and External Conflict

Fed 19:Madison 0.212 0.574 0.214 2 Foreign Relations and External Conflict

Federalist 18: The Same Subject Continued (The Insufficiency of the Present Confederation to Preserve the Union)

Federalist 11: The Utility of the Union in Respect to Commercial Relations and a Navy

Federalist 8: The Consequences of Hostilities Between the States

Federalist 6: Concerning Dangers from Dissensions Between the States

Federalist 19: The Same Subject Continued (The Insufficiency of the Present Confederation to Preserve the Union)

It seems that the summary theme “Foreign Relations and External Conflict” for the words in topic 3.2 might not be fully descriptive; there are several (inarguably prescient) sections regarding the handling of internal military conflicts. For the evaluation performed with four topics, it will be changed to “Internal / External Relations and Conflict”. Otherwise, this theme seems to be in-line with the papers selected.

**Five highest probabilities in topic 3.3**

Paper pTopic1 pTopic2 pTopic3 N TopicDesc

Fed 10:Madison 0.123 0.143 0.734 3 Structure Amidst Differing Agendas

Fed 56:Disputed 0.206 0.099 0.695 3 Structure Amidst Differing Agendas

Fed 58:Disputed 0.208 0.123 0.669 3 Structure Amidst Differing Agendas

Fed 57:Disputed 0.223 0.125 0.652 3 Structure Amidst Differing Agendas

Fed 55:Disputed 0.188 0.166 0.646 3 Structure Amidst Differing Agendas

Federalist 10: The Same Subject Continued (The Union as a Safeguard Against Domestic Faction and Insurrection)

Federalist 56: The Same Subject Continued (The Total Number of the House of Representatives)

Federalist 58: Objection That The Number of Members Will Not Be Augmented as the Progress of Population Demands Considered

Federalist 57: The Alleged Tendency of the New Plan to Elevate the Few at the Expense of the Many Considered in Connection with Representation

Federalist 55: The Total Number of the House of Representatives

Federalist 10 is the very definition of the subject of differing agendas: it is a discussion on the interrelated effects of opinionated faction upon liberty. The remaining papers discuss the structure of the House of Representatives and constitute an organized four-paper defense (as seen in the first sentence of each) of its construction in the face of detractors’ arguments.

## Consideration of - and Findings from - a Four-Topic Structure

While the specification of three topics looks to generate a successful parsing of papers into recognizable themes, is there room for a fourth topic? What re-shuffling of topics, papers and word arrangements will occur and will they seem to be more appropriate than a three-topic scenario?

With a fourth topic added, their overall ordering is different. To more easily compare them and see which ones correspond to each other between scenarios, they are to be displayed together along with a color key. One note: some words are shared between topics in the three-topic format, and some are newly-introduced – shown in black - in the four-topic scheme.

### A Comparison of 3-Topic and 4-Topic Mappings

|  |
| --- |
| Topic 3.1 Topic 3.2 Topic 3.3 |
| "state" "nation" "will" |
| "power" "union" "govern" |
| "constitut" "state" "peopl" |
| "author" "must" "may" |
| "execut" "general" "state" |
| "case" "power" "interest" |
| "law" "can" "must" |
| "may" "foreign" "repres" |
| "legisl" "upon" "number" |
| "shall" "war" "differ" |
| "Paper Count" "Paper Count" "Paper Count" |
| "21" "24" "40" |

Given this mapping, topic 3.1 becomes topic 4.3. Six of the ten words cross over to the other topic.

Topic 3.2 becomes topic 4.4. The new words here (specifically ‘danger’ and ‘situation’) here seem to bolster the theme of conflict amongst two nations (‘country’ and ‘confederacy’).

Topic 3.3 becomes topic 4.2. One new word is introduced here (‘object’), which complements the general theme of differing agendas.

|  |
| --- |
| Topic 4.1 Topic 4.2 Topic 4.3 Topic 4.4 |
| "power" "will" "state" "nation" |
| "execut" "govern" "power" "union" |
| "legisl" "peopl" "constitut" "war" |
| "bodi" "state" "law" "danger" |
| "senat" "may" "author" "confederaci" |
| "constitut" "must" "govern" "countri" |
| "might" "repres" "case" "foreign" |
| "appoint" "interest" "may" "great" |
| "one" "differ" "court" "time" |
| "offic" "object" "nation" "situat" |
| "Paper Count" "Paper Count" "Paper Count" "Paper Count" |
| "18" "30" "18" "19" |

This leaves topic 4.1 as the new topic introduced. We will look at the five papers most likely by probability to embody the theme and look at the papers’ composition to ensure that this is an appropriate (and potentially an improved) topic mapping.

### Four-Topic Evaluation

Topic 4.1: Powers and Limitations of the Executive Branch

Topic 4.2: Structure Amidst Differing Agendas

Topic 4.3: A Nation of Laws

Topic 4.4: Internal/External Relations and Conflict

**Five highest probabilities in topic 1**

Paper pTopic1 pTopic2 pTopic3 pTopic4 N TopicDesc

Fed 47:Madison 0.615 0.110 0.220 0.055 1 Powers and Limitations of the Executive Branch

Fed 76:Hamilton 0.551 0.298 0.117 0.034 1 Powers and Limitations of the Executive Branch

Fed 77:Hamilton 0.516 0.241 0.155 0.088 1 Powers and Limitations of the Executive Branch

Fed 73:Hamilton 0.485 0.220 0.177 0.118 1 Powers and Limitations of the Executive Branch

Fed 67:Hamilton 0.480 0.097 0.334 0.090 1 Powers and Limitations of the Executive Branch

Federalist 47: The Particular Structure of the New Government and the Distribution of Power Among Its Different Parts

Federalist 76: The Appointing Power of the Executive

Federalist 77: The Appointing Power Continued and Other Powers of the Executive Considered

Federalist 73: The Provision For The Support of the Executive, and the Veto Power

Federalist 67: The Executive Department

New topic 1 looks to be focused on the duties of the Executive Branch. A topic theme of “Powers and Limitations of the Executive Branch” effectively sums up the subject matter. Federalist 47 was previously in the “A Nation of Laws” theme, and as noted below, was supplanted from that group for good reason.

**Five highest probabilities in topic 4.2**

Paper pTopic1 pTopic2 pTopic3 pTopic4 N TopicDesc

Fed 10:Madison 0.122 0.674 0.097 0.108 2 Structure Amidst Differing Agendas

Fed 56:Disputed 0.105 0.629 0.183 0.083 2 Structure Amidst Differing Agendas

Fed 55:Disputed 0.151 0.604 0.152 0.092 2 Structure Amidst Differing Agendas

Fed 54:Disputed 0.109 0.591 0.253 0.048 2 Structure Amidst Differing Agendas

Fed 58:Disputed 0.140 0.589 0.170 0.101 2 Structure Amidst Differing Agendas

The papers in topic 2 remain the same as in the mapped three-topic structure (topic 3.3).

**Five highest probabilities in topic 4.3**

Paper pTopic1 pTopic2 pTopic3 pTopic4 N TopicDesc

Fed 82:Hamilton 0.106 0.077 0.779 0.038 3 A Nation of Laws

Fed 32:Hamilton 0.092 0.102 0.718 0.089 3 A Nation of Laws

Fed 83:Hamilton 0.121 0.150 0.663 0.066 3 A Nation of Laws

Fed 80:Hamilton 0.088 0.171 0.653 0.089 3 A Nation of Laws

Fed 81:Hamilton 0.205 0.119 0.644 0.032 3 A Nation of Laws

Federalist 47 falls off the prior list, but a stronger paper (80) is added. It is more in line with the judiciary subject matter of papers 81-83.

Federalist 80: The Powers of the Judiciary

**Five highest probabilities in topic 4.4**

Paper pTopic1 pTopic2 pTopic3 pTopic4 N TopicDesc

Fed 8:Hamilton 0.088 0.229 0.119 0.565 4 Internal/External Relations and Conflict

Fed 11:Hamilton 0.100 0.235 0.124 0.541 4 Internal/External Relations and Conflict

Fed 18:Madison 0.186 0.174 0.099 0.540 4 Internal/External Relations and Conflict

Fed 5:Jay 0.145 0.245 0.083 0.527 4 Internal/External Relations and Conflict

Fed 4:Jay 0.113 0.244 0.126 0.518 4 Internal/External Relations and Conflict

The two papers from John Jay that were added are even more appropriate to the topic of conflict:

Federalist 4: The Same Subject Continued (Concerning Dangers From Foreign Force and Influence)

Federalist 5: The Same Subject Continued (Concerning Dangers From Foreign Force and Influence)

## Effectiveness of the Four-Topic Structure

Evaluating the effectiveness of a three-topic versus a four-topic structure, the search for a fourth appropriate topic was successful and indeed preferential, given consideration of their top five papers by probability:

1. A new topic whose top-five papers had a strong theme related to the Executive Branch.
2. One topic whose top-five papers was unchanged from the three-topic structure.
3. Two topics switched out prior top-five papers for ones that strengthened their topical themes.

# Discussion of Relevant Graphic Output

Providing the LDA object resulting from the four-topic analysis, output from the LDAvis package provides some additional insights. However, much like topic numbers were rearranged between processing a three-topic versus a four-topic structure, LDAvis does not keep the topic numbers consistent.

We can determine which they are by keeping the lambda parameter of LDAvis set to 1. Lambda – when set to 0 - lets us incrementally weight a word’s relevance in only the topic highlighted, or - when set to 1 – weights its relevance to overall usage. At that setting, the top ten words presented when one of the topics is chosen will comprise one of the word lists presented earlier.

## Topic 1

|  |
| --- |
| Topic 4.2 |
| "will" |
| "govern" |
| "peopl" |
| "state" |
| "may" |
| "must" |
| "repres" |
| "interest" |
| "differ" |
| "object" |

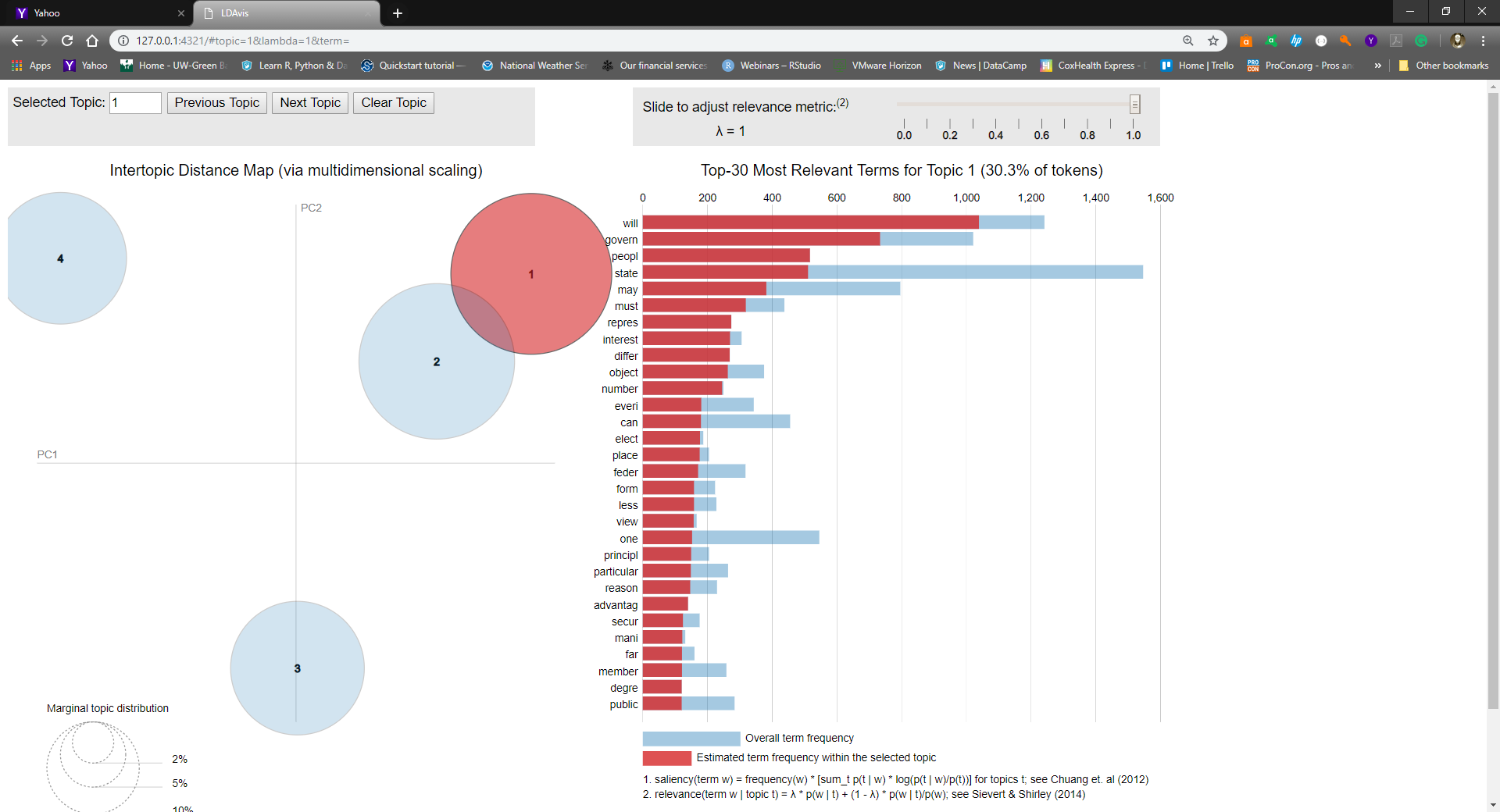
LDAvis topic 1 corresponds to LDA topic 4.2, whose theme is “Structure Amidst Differing Agendas”.

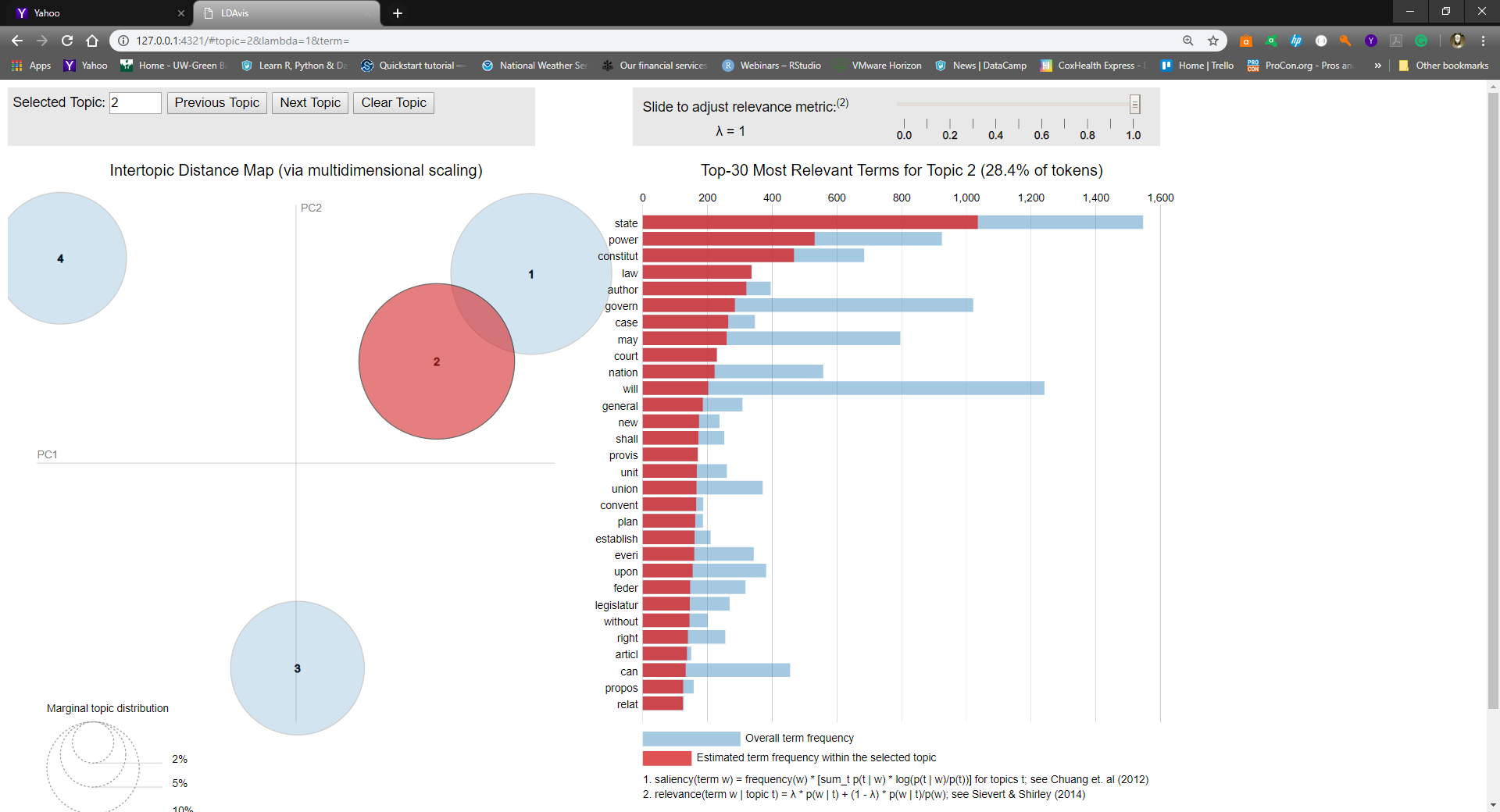
Figure – LDAvis Topic 1, Lambda = 1

The graphic to the left shows us the usage of words inside a specific topic in relation to their usage in all topics. Of note here are not just relative usage of the top two words in the list (‘will’ and ‘govern’), used at a high frequency compared to other topics, but additional words that are used almost exclusively within this topic.

‘People’ and ‘interest’ along with ‘differ’ and ‘number’ show exclusive or near-exclusive use in papers whose highest probability tags it to topic 1. This finding can lead one to more strongly infer that the theme is an appropriate representation of the topic, if one accepts the premise that exclusive use of key words in selections of literary works is a distinguishing element of a topic.

## Topic 2

|  |
| --- |
| Topic 4.3 |
| "state" |
| "power" |
| "constitut" |
| "law" |
| "author" |
| "govern" |
| "case" |
| "may" |
| "court" |
| "nation" |

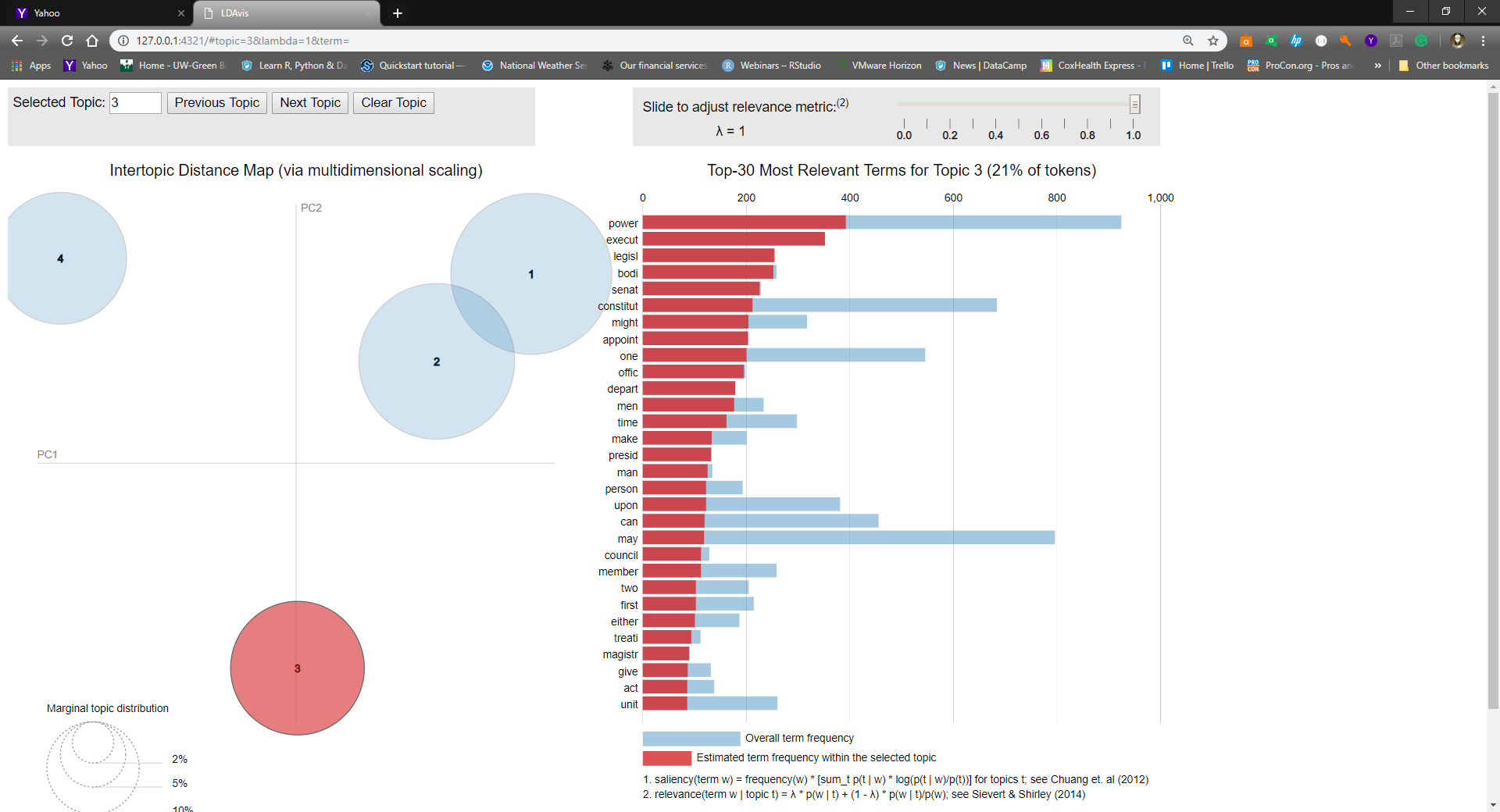
LDAvis topic 2 corresponds to LDA topic 4.3, whose theme is “A Nation of Laws “.

Words heavily weighted in this topic versus others include ‘law’, ‘court’, ‘author(ity)’, and ‘case’, again bolstering the argument that the theme is a good representation of the topic.

Figure 2 - LDAvis Topic 2, Lambda = 1

## Topic 3

|  |
| --- |
| Topic 4.1 |
| "power" |
| "execut" |
| "legisl" |
| "bodi" |
| "senat" |
| "constitut" |
| "might" |
| "appoint" |
| "one" |
| "offic" |

LDAvis topic 3 corresponds to LDA topic 4.1, whose theme is “Powers and Limitations of the Executive Branch”.

Again, words found exclusively – or nearly so – in papers with their highest probability in this topic are ‘execut(ive)’, ‘legisl(ative)’, ‘bodi(es)’, ‘senat(e)’, ‘appoint’, ‘offic(ial)’, and further down, ‘presid(ent)’. Once again, this seems to bolster the thematic statement assigned to the topic.

Figure 3 -LDAvis Topic 3, Lambda = 1

## Topic 4

|  |
| --- |
| Topic 4.4 |
| "nation" |
| "union" |
| "war" |
| "danger" |
| "confederaci" |
| "countri" |
| "foreign" |
| "great" |
| "time" |
| "situat" |

LDAvis topic 4 corresponds to LDA topic 4.4, whose theme is “Internal/External Relations and Conflict “.

Words with very high or nearly exclusive relative term frequency in papers with their highest probability in this topic are ‘war’, ‘confederaci(es)’, ‘countri(es)’, ‘foreign’, ‘situation’, and outside of the top ten, ‘peac(e)’, ‘america’, and ‘ami(cable)’.

Again, if one accepts the premise of near-exclusive use of specific words in selections of literary works as evidence of an author invoking a topical angle, this looks to confirm one’s suspicions.

Figure 4 - LDAvis Topic 4, Lambda = 1

## Theme Differentiation

The LDAvis diagram also shows a bit of thematic overlap between topic 1 (“Structure Amidst Differing Agendas”) and topic 2 (“ A Nation of Laws”) as measured on the axes of the first two principal components resulting from Linear Discriminant Analysis.

LDA topics 3 (“Powers and Limitations of the Executive Branch”) and 4 (“Internal/External Relations and Conflict“) are well-differentiated on the two axes. Subjectively, this makes sense especially when considering a more similar topical relationship between 1 and 2.

Figure 5 - Topic differentiation per LDA

# Possibilities for Further Refinement

* Nothing precludes the search for the appropriateness of a fifth topic to the analysis. Such an endeavor could be handled in the manner of the process undertaken for transitioning between three and four topics.
* Many large numbers are not written out numerically and are instead spelled out fully. They could be removed if doing so is done in the same spirit as the command to remove numeric values from the corpus. This would reduce the number of occurrences of ‘one’, seen midway through LDAvis topic 4 in Figure 4.

For example, from Federalist 12… :

…Hitherto, I believe, it may safely be asserted, that these duties have not upon an average exceeded in any State three per cent. In France they are estimated to be about fifteen per cent., and in Britain they exceed this proportion. …Upon a ratio to the importation into this State, the whole quantity imported into the United States may be estimated at four millions of gallons; which, at a shilling per gallon, would produce two hundred thousand pounds.

…and, from Federalist 56:

The number of inhabitants in the two kingdoms of England and Scotland cannot be stated at less than eight millions. The representatives of these eight millions in the House of Commons amount to five hundred and fifty-eight. Of this number, one ninth are elected by three hundred and sixty-four persons, and one half, by five thousand seven hundred and twenty-three persons.

# References

History of Alexander Hamiltion –

https://en.wikipedia.org/wiki/Alexander\_Hamilton

History of James Madison –

https://en.wikipedia.org/wiki/James\_Madison

History of John Jay –

<https://en.wikipedia.org/wiki/John_Jay>

Origin of 'Publius' *nom-de-plume* –

https://meynercenter.lafayette.edu/publius-journal/

Parameter settings for R LDAvis package -

https://cran.r-project.org/web/packages/LDAvis/LDAvis.pdf

Parameter settings for R topicmodels package–

https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf

Project Gutenberg text file –

http://www.gutenberg.org/cache/epub/18/pg18.txt

(other formats at - http://www.gutenberg.org/ebooks/18)

Wikipedia background –

https://en.wikipedia.org/wiki/The\_Federalist\_Papers

The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization -

http://pages.cs.wisc.edu/~gfung/federalist.pdf