

# Text Analysis

## Introduction to Voyant

Dr. Sierra Eckert

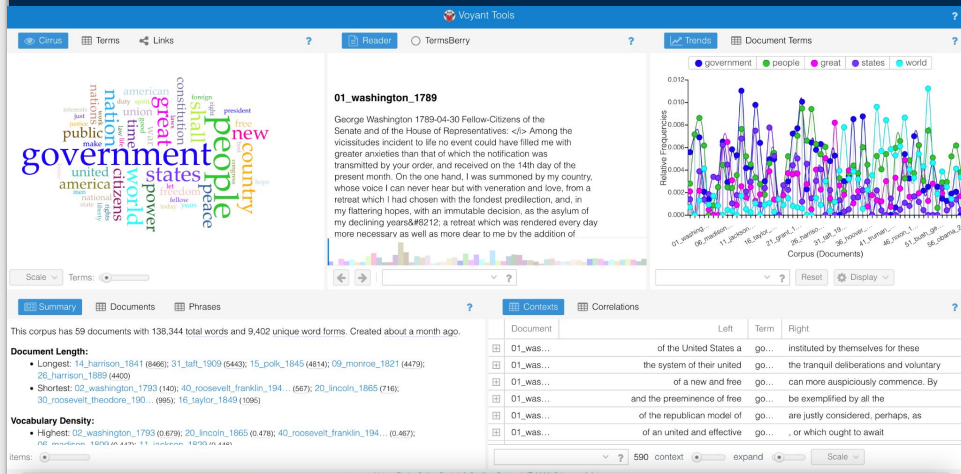
Princeton University

[sceckert@princeton.edu](mailto:sceckert@princeton.edu)

An off-the-shelf alternative  
for exploratory data analysis:

# Voyant - a basic dashboard for text analysis

1. Topic modeling browser

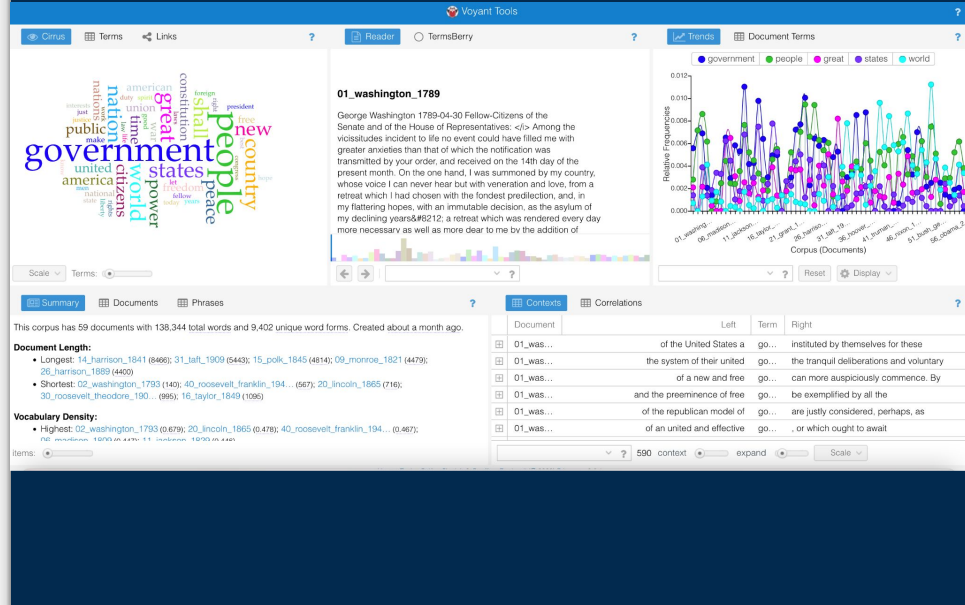
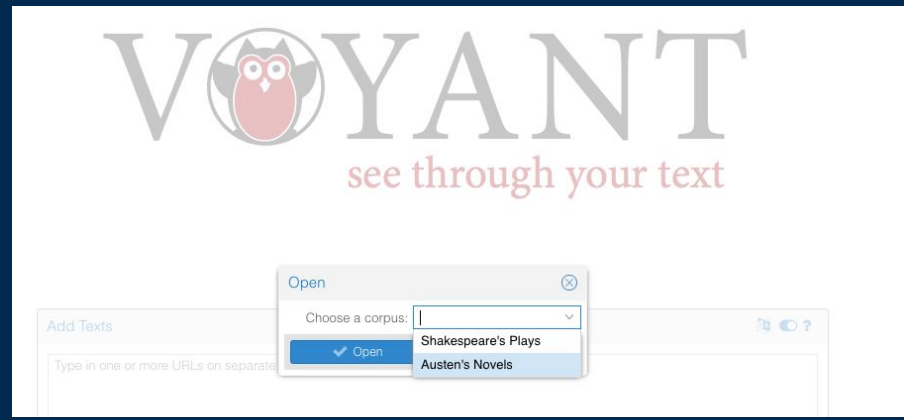


An off-the-shelf alternative  
for exploratory data analysis:

## Voyant - a basic dashboard for text analysis

1. Topic modeling browser

If you're interested, I'm  
happy to share a simple  
tutorial I developed using  
today's corpus!







# Voyant Tools

In the upper center square is the **TEXT of your CORPUS**.

This gives you the full text that you're analyzing, listed in order that they are labeled (in this case, by date). If you hover over a word, it will tell you how many times it appears in the collection.

The screenshot displays the Voyant Tools interface. On the left is a word cloud with prominent words like 'government', 'people', 'states', 'world', and 'citizens'. In the center, a red box highlights the 'Reader' tab, which shows the text of document '01\_washington\_1789'. On the right, a 'Trends' chart shows relative frequencies of words across various documents. Below the word cloud, there are controls for 'Scale' and 'Terms'. At the bottom, a 'Summary' section provides corpus statistics, and a 'Contexts' table shows word usage in different documents.

**01\_washington\_1789**

George Washington 1789-04-30 Fellow-Citizens of the Senate and of the House of Representatives: </i> Among the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order, and received on the 14th day of the present month. On the one hand, I was summoned by my country, whose voice I can never hear but with veneration and love, from a retreat which I had chosen with the fondest predilection, and, in my flattering hopes, with an immutable decision, as the asylum of my declining years&#8212;a retreat which was rendered every day more necessary as well as more dear to me by the addition of

Document

Document	Left	Term	Right
01_was...	of the United States a	go...	instituted by themselves for these
01_was...	the system of their united	go...	the tranquil deliberations and voluntary
01_was...	of a new and free	go...	can more auspiciously commence. By
01_was...	and the preeminence of free	go...	be exemplified by all the
01_was...	of the republican model of	go...	are justly considered, perhaps, as
01_was...	of an united and effective	go...	, or which ought to await

590 context expand Scale



In the bottom left are STATISTICS about your corpus.

Scrolling down in the “Summary” view gives a longer list of some of most distinctive words in each text, the average document and sentence length. “Phrases” allows you to sort by short phrases.

# Voyant Tools

The screenshot displays the Voyant Tools interface. On the left is a word cloud for the corpus. The center shows a document titled "01\_washington\_1789" with its text. On the right is a line chart showing relative frequencies of terms across documents. The bottom left contains a summary table with the following data:

This corpus has 59 documents with 138,344 total words and 9,402 unique word forms. Created about a month ago.

**Document Length:**

- Longest: 14\_harrison\_1841 (8466); 31\_taft\_1909 (5443); 15\_polk\_1845 (4814); 09\_monroe\_1821 (4479); 26\_harrison\_1889 (4400)
- Shortest: 02\_washington\_1793 (140); 40\_roosevelt\_franklin\_194... (567); 20\_lincoln\_1865 (716); 30\_roosevelt\_theodore\_190... (995); 16\_taylor\_1849 (1095)

**Vocabulary Density:**

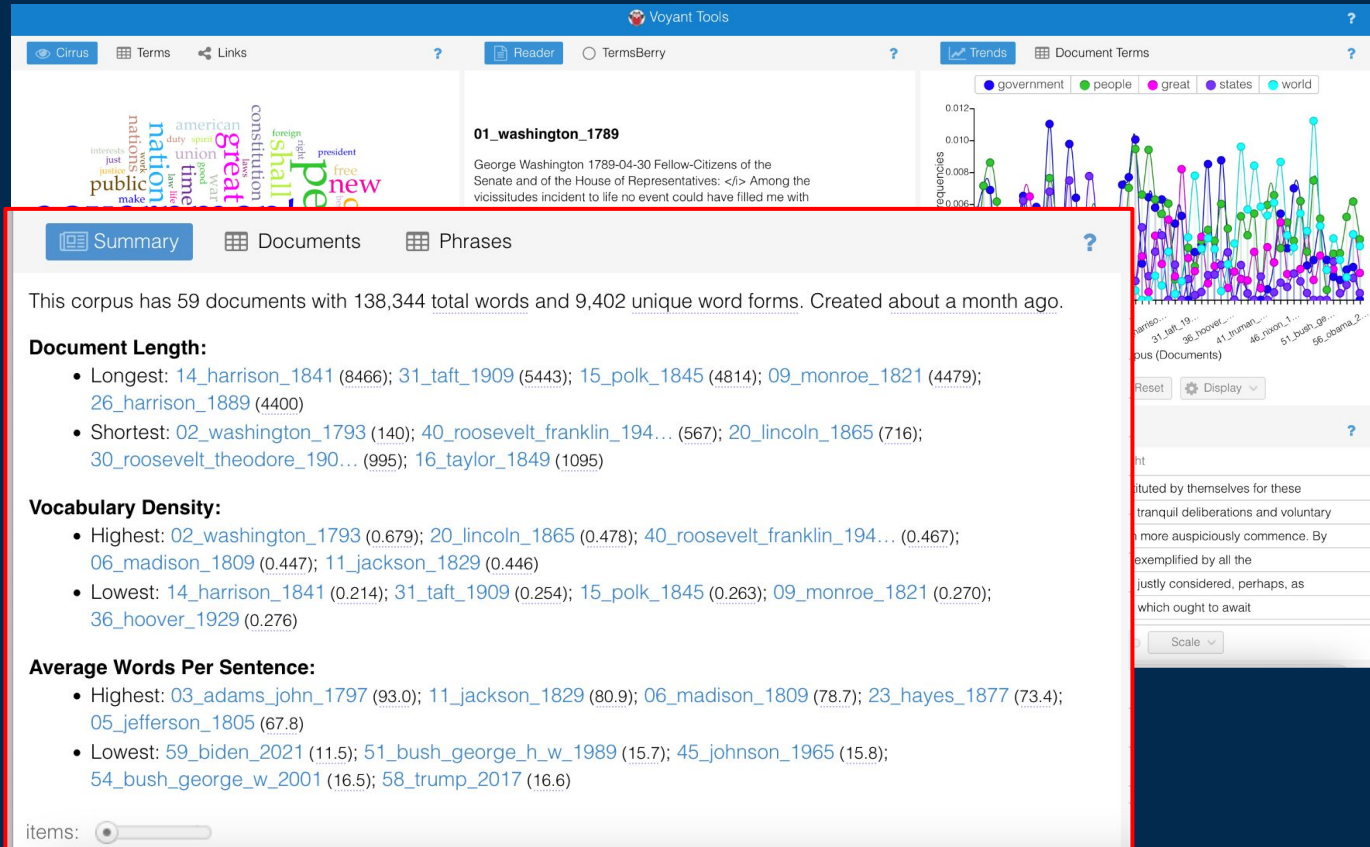
- Highest: 02\_washington\_1793 (0.679); 20\_lincoln\_1865 (0.478); 40\_roosevelt\_franklin\_194... (0.467); 06\_madison\_1800 (0.437); 11\_monroe\_1820 (0.435)

Document	Left	Term	Right
01_was...	of the United States a	go...	instituted by themselves for these
01_was...	the system of their united	go...	the tranquil deliberations and voluntary
01_was...	of a new and free	go...	can more auspiciously commence. By
01_was...	and the preeminence of free	go...	be exemplified by all the
01_was...	of the republican model of	go...	are justly considered, perhaps, as
01_was...	of an united and effective	go...	, or which ought to await



With this same “Summary” box, you’ll have descriptive statistics on the text or corpus (collection of texts) that you’re working with. These include “Document Length”, “Vocabulary Density”, “Average Words Per Sentence” You’ll also see “Most Frequent Words” in the corpus and Most Distinctive Words in each document. If you want to know what exactly these are measuring, click on the question mark in this box’s upper right corner.

# Voyant Tools



The screenshot shows the Voyant Tools interface. At the top, there's a navigation bar with "Voyant Tools" and a search icon. Below that, there are tabs for "Cirrus", "Terms", "Links", "Reader", "TermsBerry", "Trends", and "Document Terms". The main area is divided into three sections: a word cloud on the left, a document preview in the center, and a trends graph on the right. The word cloud contains words like "nation", "great", "constitution", "public", "time", "union", "duty", "spirit", "foreign", "shall", "president", "new", "free", "wait", "love", "make", "interest", "nations", "public", "time", "union", "duty", "spirit", "foreign", "shall", "president", "new", "free", "wait", "love", "make". The document preview shows "01\_washington\_1789" with a snippet of text: "George Washington 1789-04-30 Fellow-Citizens of the Senate and of the House of Representatives: <i> Among the vicissitudes incident to life no event could have filled me with". The trends graph shows the frequency of words over time, with a legend for "government", "people", "great", "states", and "world".

**Summary** Documents Phrases ?

This corpus has 59 documents with 138,344 total words and 9,402 unique word forms. Created about a month ago.

**Document Length:**

- Longest: [14\\_harrison\\_1841](#) (8466); [31\\_taft\\_1909](#) (5443); [15\\_polk\\_1845](#) (4814); [09\\_monroe\\_1821](#) (4479); [26\\_harrison\\_1889](#) (4400)
- Shortest: [02\\_washington\\_1793](#) (140); [40\\_roosevelt\\_franklin\\_194...](#) (567); [20\\_lincoln\\_1865](#) (716); [30\\_roosevelt\\_theodore\\_190...](#) (995); [16\\_taylor\\_1849](#) (1095)

**Vocabulary Density:**

- Highest: [02\\_washington\\_1793](#) (0.679); [20\\_lincoln\\_1865](#) (0.478); [40\\_roosevelt\\_franklin\\_194...](#) (0.467); [06\\_madison\\_1809](#) (0.447); [11\\_jackson\\_1829](#) (0.446)
- Lowest: [14\\_harrison\\_1841](#) (0.214); [31\\_taft\\_1909](#) (0.254); [15\\_polk\\_1845](#) (0.263); [09\\_monroe\\_1821](#) (0.270); [36\\_hoover\\_1929](#) (0.276)

**Average Words Per Sentence:**

- Highest: [03\\_adams\\_john\\_1797](#) (93.0); [11\\_jackson\\_1829](#) (80.9); [06\\_madison\\_1809](#) (78.7); [23\\_hayes\\_1877](#) (73.4); [05\\_jefferson\\_1805](#) (67.8)
- Lowest: [59\\_biden\\_2021](#) (11.5); [51\\_bush\\_george\\_h\\_w\\_1989](#) (15.7); [45\\_johnson\\_1965](#) (15.8); [54\\_bush\\_george\\_w\\_2001](#) (16.5); [58\\_trump\\_2017](#) (16.6)

items:

This box also allows you to control the global filters for the toolset.  
 Hover over the upper right corner of this box and click on the “options” toggle

# Voyant Tools

The screenshot displays the Voyant Tools interface with the following components:

- Word Cloud:** A word cloud on the left side with prominent words like "government", "people", "states", "country", "new", "shall", "great", "united", "citizens", "world", "power", "freedom", "peace", "constitution", "foreign", "president", "free", "nation", "public", "american", "citizens", "world", "power", "freedom", "peace", "constitution", "foreign", "president", "free", "nation", "public", "american".
- Document Text:** A text box titled "01\_washington\_1789" containing the opening of George Washington's inaugural address: "George Washington 1789-04-30 Fellow-Citizens of the Senate and of the House of Representatives: </i> Among the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order, and received on the 14th day of the present month. On the one hand, I was summoned by my country, whose voice I can never hear but with veneration and love, from a retreat which I had chosen with the fondest predilection, and, in my flattering hopes, with an immutable decision, as the asylum of my declining years&#8212;a retreat which was rendered every day more necessary as well as more dear to me by the addition of".
- Trends Chart:** A line chart titled "Trends" showing "Relative Frequencies" on the y-axis (0.000 to 0.012) and "Corpus (Documents)" on the x-axis. The chart tracks the frequency of terms like "government", "people", "great", "states", and "world" across various documents.
- Summary Panel:** A panel at the bottom left providing corpus statistics: "This corpus has 59 documents with 138,344 total words and 9,402 unique word forms. Created about a month ago." It includes sections for "Document Length" (listing longest and shortest documents) and "Vocabulary Density" (listing highest density documents).
- Contexts Panel:** A table at the bottom right showing "Contexts" for the term "01\_was...". The table has columns for "Left", "Term", and "Right". A red box highlights the "options" icon in the top right corner of this panel.

In the options box, click on “None.” Then click Confirm. What happened?

What you just removed was a “stop words” list.

Click on options again, and click on “auto-detect.” Then click on the Edit list button. What do you notice about the words?

When would you want to filter certain words out? When *wouldn't* you want to remove them? What are the implications?

## Voyant Tools

The screenshot shows the Voyant Tools interface. On the left is a word cloud for the document '01\_washington\_1789'. The central panel displays the text of the document: "George Washington 1789-04-30 Fellow-Citizens of the Senate and of the House of Representatives: <-> Among the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order, and received on the 14th day of the present month. On the one hand, I was summoned by my country, whose voice I can never hear but with veneration and love, from a". On the right is a line graph showing relative frequencies of words across documents. A red box highlights the 'Options' dialog box, which is open to the 'Stopwords' section. The 'Stopwords' dropdown is set to 'Auto-detect', and the 'Categories' dropdown is set to 'None'. The 'Apply globally' checkbox is checked. The 'Confirm' button is highlighted in blue.

For more about stopwords—their history and their role in computational analysis today— see this article by Daniel Rosenberg, “Stop, Words.” *Representations* 127, no. 1 (August 1, 2014): 83–92.

<https://doi.org/10.1525/rep.2014.127.1.83>.

In the bottom right is a CONCORDANCE.

This gives you the context of words in your corpus as they appear in each document.

Try typing in “government” and sorting by the words that appear on the left.

Toggle to the “Bubblelines” view. Type in “America,” “government,” “liberty.” What do you notice?

# Voyant Tools

The screenshot displays the Voyant Tools interface. On the left is a word cloud for the term "government". The central pane shows the document "01\_washington\_1789" with its text and a bar chart below. On the right is a "Trends" chart showing relative frequencies of terms across documents. At the bottom, a concordance table is visible, showing the context of the term "government" in various documents.

**Word Cloud Terms:** government, people, country, states, freedom, power, world, citizens, united, america, nation, public, make, duty, spirit, union, great, shall, president, free, new, peace, fellow, today, years, men, national, state, republic, theory, constitution, foreign, with, love, let, free, nation, duty, spirit, union, great, shall, president, free, new, peace, fellow, today, years, men, national, state, republic, theory.

**Document Text:**  
**01\_washington\_1789**  
George Washington 1789-04-30 Fellow-Citizens of the Senate and of the House of Representatives: </i> Among the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order, and received on the 14th day of the present month. On the one hand, I was summoned by my country, whose voice I can never hear but with veneration and love, from a retreat which I had chosen with the fondest predilection, and, in my flattering hopes, with an immutable decision, as the asylum of my declining years&#8212; a retreat which was rendered every day more necessary as well as more dear to me by the addition of

**Concordance Table:**

Document	Left	Term	Right
01_was...	of the United States a	go...	instituted by themselves for these
01_was...	the system of their united	go...	the tranquil deliberations and voluntary
01_was...	of a new and free	go...	can more auspiciously commence. By
01_was...	and the preeminence of free	go...	be exemplified by all the
01_was...	of the republican model of	go...	are justly considered, perhaps, as
01_was...	of an united and effective	go...	, or which ought to await

You can also look for short phrases in context:  
 Try typing in "I think" and sorting by the words that appear on the right.  
 What do you notice?

# Voyant Tools

The screenshot shows the Voyant Tools interface. On the left is a word cloud for the corpus. The top right features a 'Trends' chart with a legend for 'government', 'people', 'great', 'states', and 'world'. The main area is titled '01\_washington\_1789' and includes a document summary with sections for Document Length, Vocabulary Density, and Average Words Per Sentence. At the bottom right, a 'Contexts' table is highlighted with a red border, displaying search results for the phrase 'government is' across various documents.

**Document Summary:**  
 This corpus has 59 documents with 138,344 total words and 9,402 unique word forms. Created about a month ago.

**Document Length:**

- Longest: 14\_harrison\_1841 (8466); 31\_taft\_1909 (5443); 15\_polk\_1845 (4814); 09\_monroe\_1821 (4479); 26\_harrison\_1889 (4400)
- Shortest: 02\_washington\_1793 (140); 40\_roosevelt\_franklin\_194... (567); 20\_lincoln\_1865 (716); 30\_roosevelt\_theodore\_190... (995); 16\_taylor\_1849 (1095)

**Vocabulary Density:**

- Highest: 02\_washington\_1793 (0.679); 20\_lincoln\_1865 (0.478); 40\_roosevelt\_franklin\_194... (0.467); 06\_madison\_1809 (0.447); 11\_jackson\_1829 (0.446)
- Lowest: 14\_harrison\_1841 (0.214); 31\_taft\_1909 (0.254); 15\_polk\_1845 (0.263); 09\_monroe\_1821 (0.270); 36\_hoover\_1929 (0.276)

**Average Words Per Sentence:**

- Highest: 03\_adams\_john\_1797 (93.0); 11\_jackson\_1829 (80.9); 06\_madison\_1809 (78.7); 23\_hayes\_1877 (73.4); 05\_jefferson\_1805 (67.8)
- Lowest: 59\_biden\_2021 (11.5); 51\_bush\_george\_h\_w\_1989 (15.7); 45\_johnson\_1965 (15.8); 54\_bush\_george\_w\_2001 (16.5); 58\_trump\_2017 (16.6)

**Contexts Table:**

Document	Left	Term	Right ↑
24_garfi...	of the ballot. Bad local	government is	certainly a great evil, which
31_taft_...	mere consideration of economy. Our	government is	able to afford a suitable
19_linc...	other alternative, for continuing the	government is	acquiescence on one side or
11_jack...	invincible. As long as our	government is	administered for the good of
51_bus...	night about which form of	government is	better. We don't have to
11_jack...	expenditure of money by the	government is	but too apt to engender
38_roos...	to carry out their will.	government is	competent when all who compose
04_jeffe...	our Government, but whether our	government is	controlled by the people. January
58_trum...	our Government, but whether our	government is	controlled by the people. January
30_roos...	unchanged. We know that self-	government is	difficult. We know that no
08_mon...	result has shown that our	government is	equal to that, the greatest
41_trum	shall think Democracy maintains that	government is	established for the benefit of

Finally, Voyant will also allow you to download data and visualizations.  
 Hover over the upper right corner of the Concordance view and click on the arrow and box export view

## Voyant Tools

The screenshot displays the Voyant Tools interface for a corpus of 59 documents. The main view shows a word cloud with prominent terms like 'government', 'people', 'nation', 'power', 'citizens', 'constitution', 'states', 'peace', 'new', 'world', 'public', 'great', 'time', 'work', 'united', 'nations', 'rights', 'make', 'country', 'national', 'liberty', 'and', 'justice', 'under', 'the', 'law', 'shall', 'good', 'time', 'work', 'united', 'nations', 'rights', 'make', 'country', 'national', 'liberty', 'and', 'justice', 'under', 'the', 'law'. Below the word cloud are controls for 'Scale' and 'Terms'.

The central panel displays the document '01\_washington\_1789' by George Washington (1789-04-30), titled 'Fellow-Citizens of the Senate and of the House of Representatives: The Address to the'. A bar chart below the title shows the relative frequency of terms across the corpus.

The bottom left panel provides summary statistics:

- Document Length:**
  - Longest: 14\_harrison\_1841 (8466); 31\_taft\_1909 (6443); 15\_polk\_1845 (4814); 09\_monroe\_1821 (4479); 26\_harrison\_1889 (4400)
  - Shortest: 02\_washington\_1793 (140); 40\_roosevelt\_franklin\_194... (567); 20\_lincoln\_1865 (716); 30\_roosevelt\_theodore\_190... (995); 16\_taylor\_1849 (1095)
- Vocabulary Density:**
  - Highest: 02\_washington\_1793 (0.679); 20\_lincoln\_1865 (0.478); 40\_roosevelt\_franklin\_194... (0.467); 06\_madison\_1809 (0.447); 11\_jackson\_1829 (0.446)
  - Lowest: 14\_harrison\_1841 (0.214); 31\_taft\_1909 (0.254); 15\_polk\_1845 (0.263); 09\_monroe\_1821 (0.270); 36\_hoover\_1929 (0.276)
- Average Words Per Sentence:**
  - Highest: 03\_adams\_john\_1797 (93.0); 11\_jackson\_1829 (80.9); 06\_madison\_1809 (78.7); 23\_hayes\_1877 (73.4); 05\_jefferson\_1805 (67.8)
  - Lowest: 59\_biden\_2021 (11.5); 51\_bush\_george\_h\_w\_1989 (15.7); 45\_johnson\_1965 (15.8); 54\_bush\_george\_w\_2001 (16.5); 58\_trump\_2017 (16.6)

The bottom right panel shows a concordance table for the term 'government is'. The table has columns for Document, Left, Term, and Right. A tooltip is visible over the 'Right' column header, stating: 'Export a URL, an embeddable tool, data or a bibliographic reference.'

Document	Left	Term	Right
24_garfi...	of the ballot. Bad local	government is	certainly
31_taft...	mere consideration of economy. Our	government is	able to afford a suitable
19_inc...	other alternative, for continuing the	government is	acquiescence on one side or
11_jack...	invincible. As long as our	government is	administered for the good of
51_bus...	night about which form of	government is	better. We don't have to
11_jack...	expenditure of money by the	government is	but too apt to engender
38_roos...	to carry out their will.	government is	competent when all who compose
04_jeffe...	our Government, but whether our	government is	controlled by the people. January
58_trum...	our Government, but whether our	government is	controlled by the people. January
30_roos...	unchanged. We know that self-	government is	difficult. We know that no
08_mon...	result has shown that our	government is	equal to that, the greatest
41_trum...	shall think. Democracy maintains that	government is	established for the benefit of

At the bottom of the concordance view, there are controls for 'Items' (set to 32), a search box containing 'government is', and buttons for 'expand', 'Scale', and 'Display'.

Finally, Voyant will also allow you to download data and visualizations.

Hover over the upper right corner of the Concordance view and click on the arrow and box export view. You should see a pop up menu.

## Voyant Tools

The screenshot displays the Voyant Tools interface. On the left, a word cloud for the term 'government' is shown with related terms like 'citizens', 'constitution', 'states', 'time', 'public', 'great', 'people', 'nation', 'power', 'make', 'country', 'united', 'shall', 'new', 'world', 'union', 'citizens', 'power', 'nation', 'make', 'country', 'united', 'shall', 'new', 'world', 'union'. The main area shows a document viewer for '01\_washington\_1789' with a concordance view. An 'Export' menu is open, offering options: 'a URL for this view (tools and data)', 'Export View (Tools and Data)', 'Export Current Data', 'export current data as HTML', 'export current data as tab separated values (text)', 'export all available data in JSON', and 'export all available data as tab separated values (text)'. The 'export current data as tab separated values (text)' option is selected. Below the menu, a concordance table is visible with columns for 'Left', 'Term', and 'Right'. The table shows various contexts for the term 'government is'.

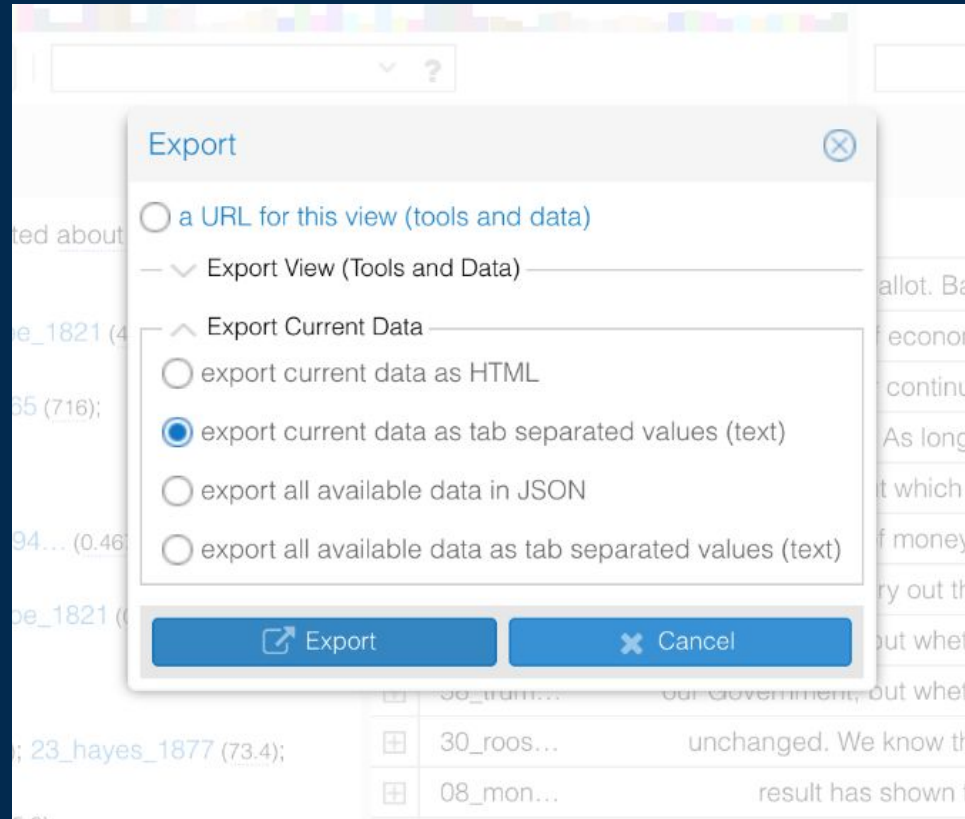
Left	Term	Right
alot. Bad local	government is	certainly a great evil, which
economy. Our	government is	able to afford a suitable
continuing the	government is	acquiescence on one side or
As long as our	government is	administered for the good of
which form of	government is	better. We don't have to
money by the	government is	but too apt to engender
ry out their will.	government is	competent when all who compose
but whether our	government is	controlled by the people. January
30_roos...	government is	difficult. We know that no
08_mon...	government is	equal to that, the greatest
41 trum...	government is	established for the benefit of

## Voyant Tools

Finally, Voyant will also allow you to download data and visualizations.

Hover over the upper right corner of the Concordance view and click on the arrow and box export view

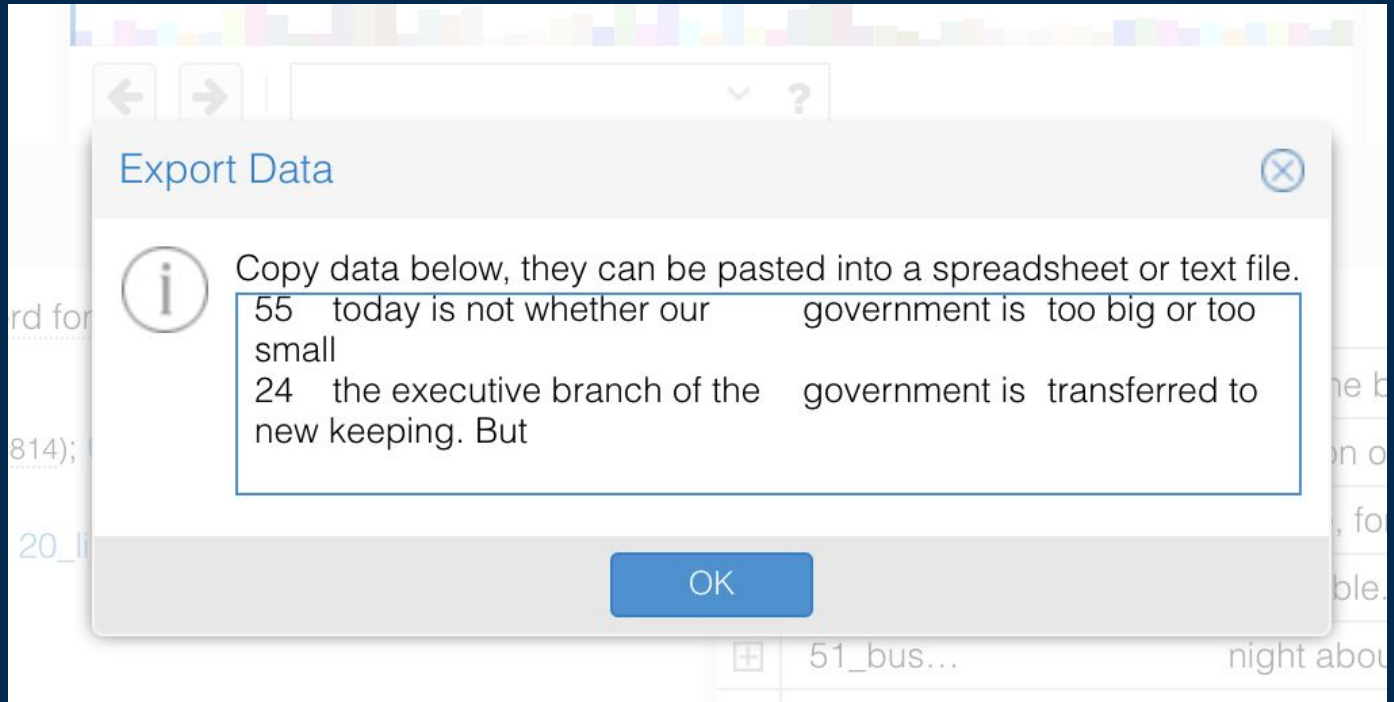
You should see a pop up menu. Click on the third option to “Export Current Data” and select tab sep. values





Exporting the current data as tab separated values (text) will give you a second popup window with data formatted in TSV that can be copied into a spreadsheet (like Excel or Google Sheets) or a simple text editor

## Voyant Tools



# Voyant Tools

Take a minute to play around some of the features. Toggle the amount of words in the CONCORDANCE, or the “items” in the STATISTICS box.

**Brainstorm** a few questions that you could explore with this kind of interface.

What kind of questions could you ask?

What kind of questions could you not ask?