

# Computational Text Analysis for Content Analysis

---

Taught By Sara Morrell, Zhen Guo, and Mel Williams  
Digital Integration Teaching Initiative (DITI)

POLS 3482 East Asian Politics

Daniel Aldrich

Spring 2026

# Workshop Agenda

- Introduction to key terms and concepts in computational text analysis (CTA).
- Discussion of CTA's applications and uses in research.
- Introduction to web-based text analysis tools.
  - Word Counter, Word Trees, Voyant, Lexos

Slides: <https://bit.ly/sp26-aldrich-pols3482-text-analysis>

Corpus: <https://bit.ly/sp26-aldrich-pols3482-text-analysis-data>

# What is Computational Text Analysis?

# Computational Text Analysis

Computational text analysis refers to the **array of methods used to “read” texts with a computer**. It is similar to statistical analysis, but the data is texts (words) instead of numbers.

Text analysis:

- Involves a computer drawing out patterns in a text, and a researcher interpreting those patterns.
- Includes methods such as word count frequency, keywords in context, computational modeling (with machine learning), and sentiment analysis.
- Is conducted using web-based tools or coding languages like Python and R.

# Why Computational Text Analysis?

Computational text analysis can help us **analyze** very large amounts of data, **identify keywords**, and **discover patterns** in texts. Using text analysis, researchers may find surprising results that they would not have discovered from traditional methods alone.

For example: "[Gendered Language in Teacher Reviews](#)" by Ben Schmidt shows stark differences in the ways that male and female professors are reviewed on "Rate My Professor."

# Gendered Language

## Gendered Language in Teacher Reviews

I've had trouble keeping this site up continuously during COVID. As of March 2021, I'm now trying a new strategy to cache common queries on the server even when the underlying database is down. If you find that many searches don't change the results, that's why.

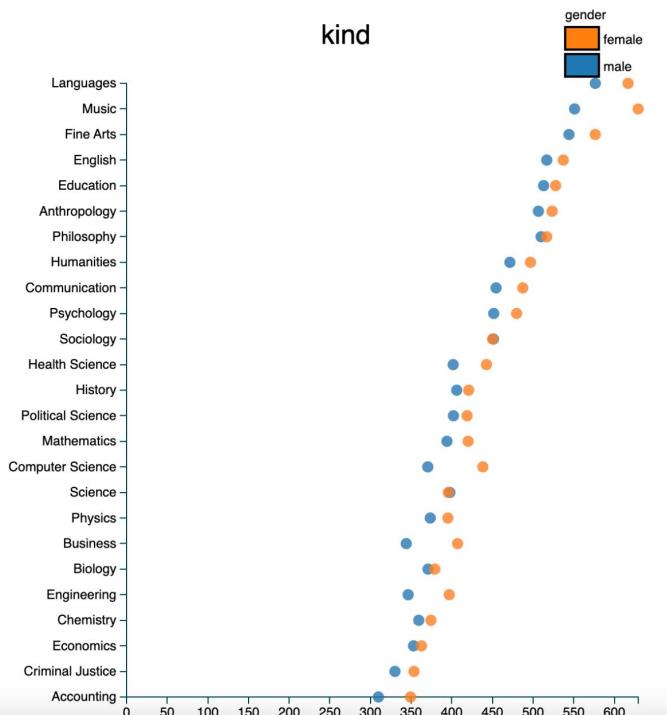
This interactive chart lets you explore the words used to describe male and female teachers in about 14 million reviews from RateMyProfessor.com.

Not all words have gender splits, but a surprising number do. Even things like pronouns are used quite differently by gender.

Search term(s) (case-insensitive):  
use commas to aggregate multiple terms

kind

All ratings   Only positive   Only negative



Go to  
[bit.ly/schmidt-gender](https://bit.ly/schmidt-gender)  
and try a few queries.  
For example:

- Smart
- Ditz
- Unprofessional
- Nice

—How do you think Schmidt determined gender for this tool?

Feel free to ask questions at any point during the presentation!

# Language used in U.S. News



Word Cloud of U.S. TV News on “Japan”.  
Terms like “Alliance” and “Security”  
appear frequently with “Japan” in U.S.  
TV news coverage 2009- October 2024.

- Go to the [Television Explorer](#). Search “Japan”, “Korea”, and “China”.

- What do you notice about the TV coverage of these terms? What is surprising?
  - How do you think political values affect language?
  - How might this language shape policies?

*Feel free to ask questions at any point during the presentation!*

# Key Terms (1/2)

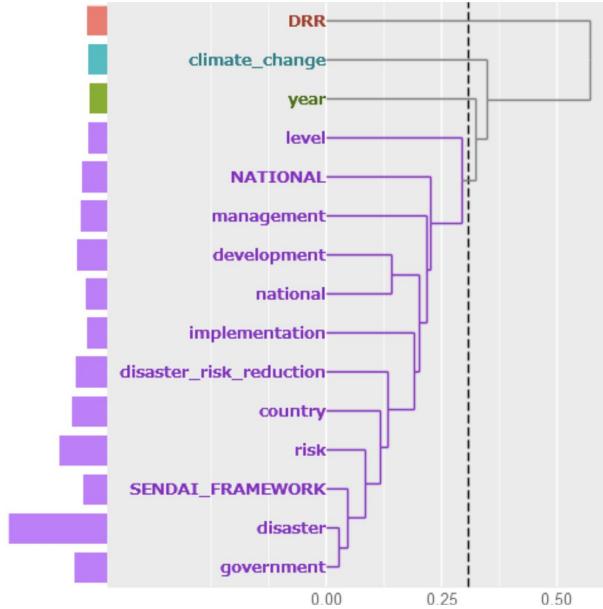
- **Corpus (plural-corpora):** A collection of texts used for analysis and research purposes.
- **Stop words:** Words that appear frequently in a language, like pronouns, prepositions, and basic verbs. These are often removed for computational analysis. Some English stop words include: a, the, she, he, I, me, us, of, is, would, could, should, etc.
- **Word Count Frequency:** Counting the total times a word appears in a text/corpus or the percentage of how often it appears.

# Key Terms (2/2)

- **nGram:** A continuous sequence of  $n$  items in a text. A bigram (or 2 continuous words) could be ‘United States,’ while a trigram (3 words) could be ‘yes we can.’
- **Sentiment Analysis:** Measuring the sentiment of a text based on a scale such as negative/positive or happy/sad. Each word has a particular weight to determine where on the scale it falls, and these weights are calculated to determine a text’s overall sentiment.

# Examples from Practice

# Word Frequency and Clusters



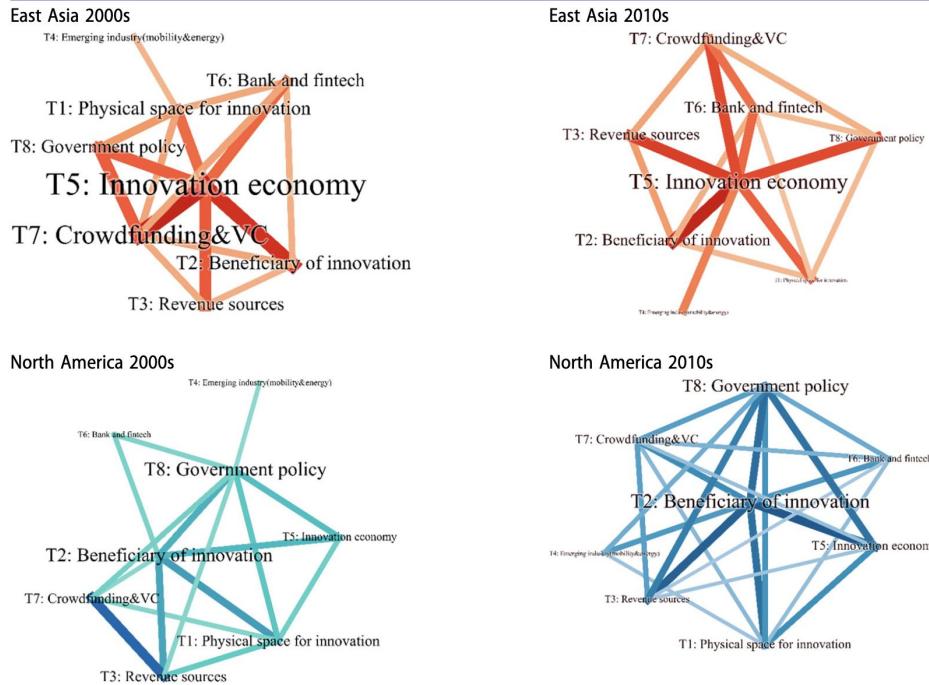
The bars on the left-hand side indicate the TF of each word [21].

**Fig. 1.** Result of the hierarchical cluster analysis.

Sasaki, D. (2019). Analysis of the attitude within asia-pacific countries towards disaster risk reduction: Text mining of the official statements of 2018 Asian ministerial conference on disaster risk reduction. Journal of Disaster Research, 14(8), 1024-1029.

# Comparative Topic Modeling

**Table 11.** Inter-topic linkages.



Re Lee, K., Hyun Kim, J., Jang, J., Yoon, J., Nan, D., Kim, Y., & Kim, B. (2023). [News big data analysis of international start-up innovation discourses through topic modelling and network analysis: comparing East Asia and North America](#). Asian Journal of Technology Innovation, 31(3), 581-603.

# Text Preparation

# Corpus Building

## Questions to consider as you begin your research:

- What are my research questions and why am I creating a corpus?
- What am I asking my corpus to do?
- What text(s) should form my corpus to answer my research questions?
- How should I organize my corpus to streamline my research processes and save time?

For more information, see our [Corpus Building Handout](#).

# Preparing Your Text

1. Choose the texts or text selections that you would like to include.
2. Create a folder on your computer or cloud storage where you will store your corpus. Give it a clearly descriptive name, without spaces or special characters.
3. If you are using a text that isn't already plain text, then copy and paste your text into a **plain text editor** (on Macs: Text Edit; on Windows: Notepad)
  - a. Mac users, you may need to make your Text Edit into a 'plain text' editor. Open Text Edit, go to Preferences, and make sure "plain text" is selected
4. Save the text as a plain text file (with a .txt extension). Always make sure to name your files so you know what is in them!
5. Repeat steps above for each text in the corpus.

# Our Text

We will use a set of plain text news articles from Asahi Shimbun pertaining to diplomacy published between 11/19/25 and 11/26/2025:

- [Japan fires back at 'unsubstantiated' Chinese letter to U.N.](#)
- [Kihara urges caution on 'misleading' Taiwan remarks](#)
- [China says trade cooperation with Japan 'severely damaged' by Taiwan comments](#)
- [China postpones trilateral culture ministers' meeting with South Korea and Japan, Seoul says](#)
- [China reimposes ban on Japanese seafood amid 'Taiwan' row](#)

# Sample Corpus

The sample .txt files are available on:

<https://bit.ly/sp26-aldrich-pols3482-text-analysis-data>

- You can download the files individual or click the Download all in the upper right corner
- If you download all, a zipped folder will download containing the files
  - On Mac: Double click the folder to unzip it
  - On PC: Right click the folder and select Extract all

# Word Counter and Word Tree

# Word Counter

- <https://databasic.io/en/wordcounter/>
- A user-friendly **word counting tool**
- It allows you to count words, bigrams, and trigrams in plain text files and to download spreadsheets with your results
- The max file upload is 10MB
- The default is to lowercase all words and remove stopwords, but you can control these options

# Word Counter Examples (1/2)

Word Counter will show you a word cloud, which can give you a sense of the **most used words in a document**. Words used more often are bigger, and ones used less often are smaller.



Kihara urges caution on 'misleading' Taiwan remarks

*Feel free to ask questions at any point during the presentation!*

# Word Counter Examples (2/2)

Shows the top words in the text.

Stopwords aren't removed for the bigrams and trigrams because they need context.

TOP WORDS	
Word	Frequency
takaichi	9
remarks	8
japanese	7
prime	6
government	6
china	6

BIGRAMS	
bigram?	Frequency
group of	5
of 20	5
at the	5
takaichi s	5
in the	5
prime minister	4

TRIGRAMS	
trigram?	Frequency
group of 20	5
the group of	4
of 20 summit	3
on nov 21	3
a mutually beneficial	3
mutually beneficial relationship	3

Kihara urges caution on 'misleading' Taiwan remarks

# Word Tree

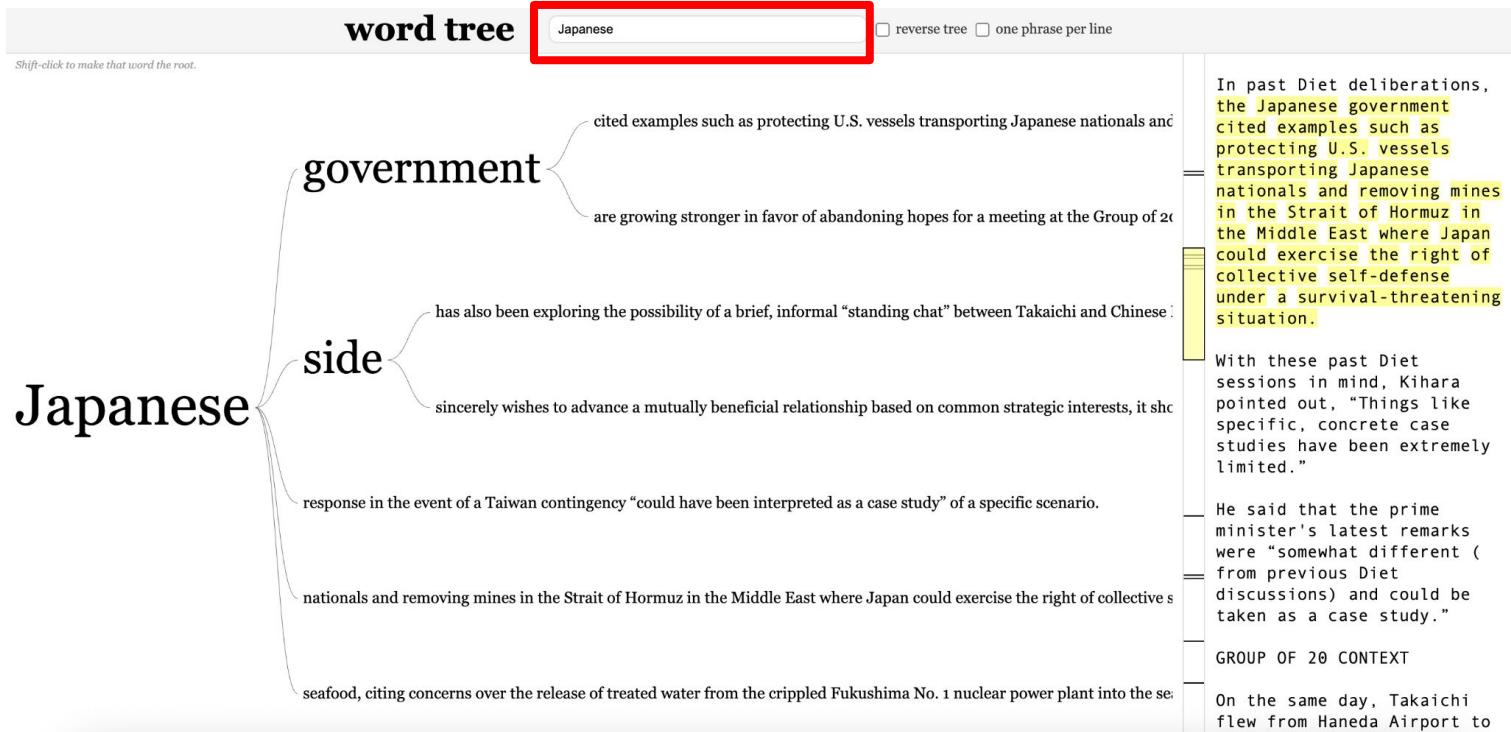
- <https://www.jasondavies.com/wordtree/>
- A word tree **depicts multiple parallel sequences of words.**
- This is a good way to see patterns in word usage, based on words that appear before and after a term or terms of interest.
- There are some restrictions in size with this tool: fewer than 1 million words should work.
- Upload your text, enter a keyword or phrase to search, then try reversing the tree.
- It's often useful to search frequent terms identified by WordCounter

# Word Tree Example

Text source:

Kihara urges  
caution on  
'misleading'  
Taiwan  
remarks

What other  
terms should  
we try?



# Word Tree: Reverse Trees

It is worth reversing the tree to see the words that often precede it. To do this click “reverse tree” next to the search bar.

Shift-click to make that word the root.

word tree   reverse tree  one phrase per line

The diagram shows the word 'Japanese' at the top right, with lines connecting it to several other words: 'the', 'comment', 'nship', 'ent', 'appropriate', 're', and 'are'. These words are connected to a main vertical line that descends through several paragraphs of text. The text discusses Japanese government actions and statements regarding Taiwan and U.S. vessels in the Middle East. The 'reverse tree' feature highlights the words 'Japanese', 'the', and 'comment' in yellow, indicating they are the most frequent precursors to the search term in the document.

S. forces would rush to defend Taiwan in the event of a crisis, were appropriate. In past Diet deliberations, ie is important precisely because there are pending issues," according to a close aide to the prime minister. comment that her thinking on promoting a mutually beneficial relationship has not changed, Mao said, "If nship based on common strategic interests, it should retract the erroneous remarks and fulfill its promises ent must be extremely careful in the future to avoid a misunderstanding. He noted that Takaichi's recent Diet testimony on a appropriate. In past Diet deliberations, the Japanese government cited examples such as protecting U.S. vessels transporting re are no plans (for a meeting)." In addition, China is increasing economic pressure, such as by effectively halting imports of

was responding to a reporter 's question on whether Takaichi's remarks, which assumed U.S. forces would rush to defend Taiwan in the event of a crisis, were appropriate.

In past Diet deliberations, the Japanese government cited examples such as protecting U.S. vessels transporting Japanese nationals and removing mines in the Strait of Hormuz in the Middle East where Japan could exercise the right of collective self-defense under a survival-threatening situation.

With these past Diet sessions in mind, Kihara pointed out, "Things like specific, concrete case studies have been extremely limited."

He said that the prime minister's latest remarks were "somewhat different ( from previous Diet

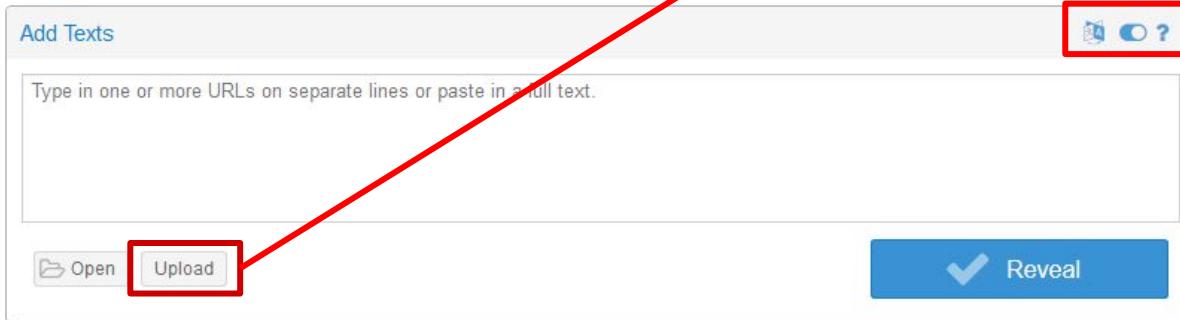
# Voyant

# Voyant Introduction

Voyant makes it possible to **perform analyses on one or multiple files in many ways**, including word counts, nGrams (n=number of words), word frequency distributions, word trends across documents, and concordances. It also makes nice visualizations!

<https://voyant-tools.org/>

# Voyant: Upload



Click on Upload and navigate to the folder with the text documents you wish to analyze.

Alternatively, insert URLs or full text into the textbox.

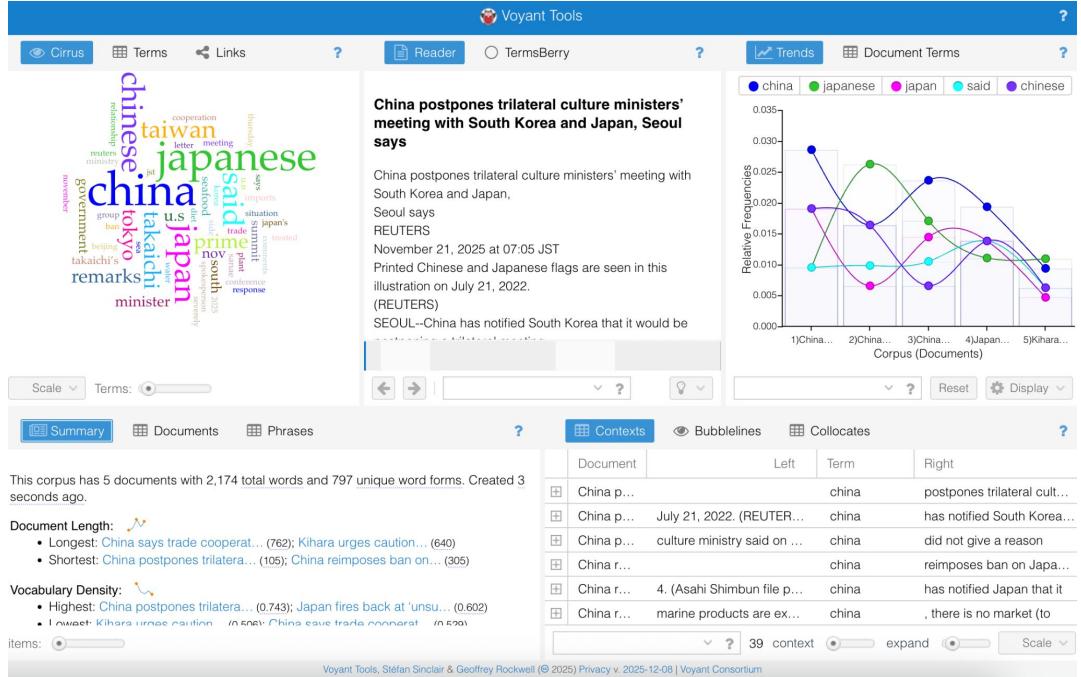
Click here for help and advanced options

# Voyant: Corpus Dashboard

## Results:

After you upload your corpus, you will see the default results page with multiple panes:

- A word cloud
  - Reader section
  - Trends
  - Document Summary
  - Word Contexts

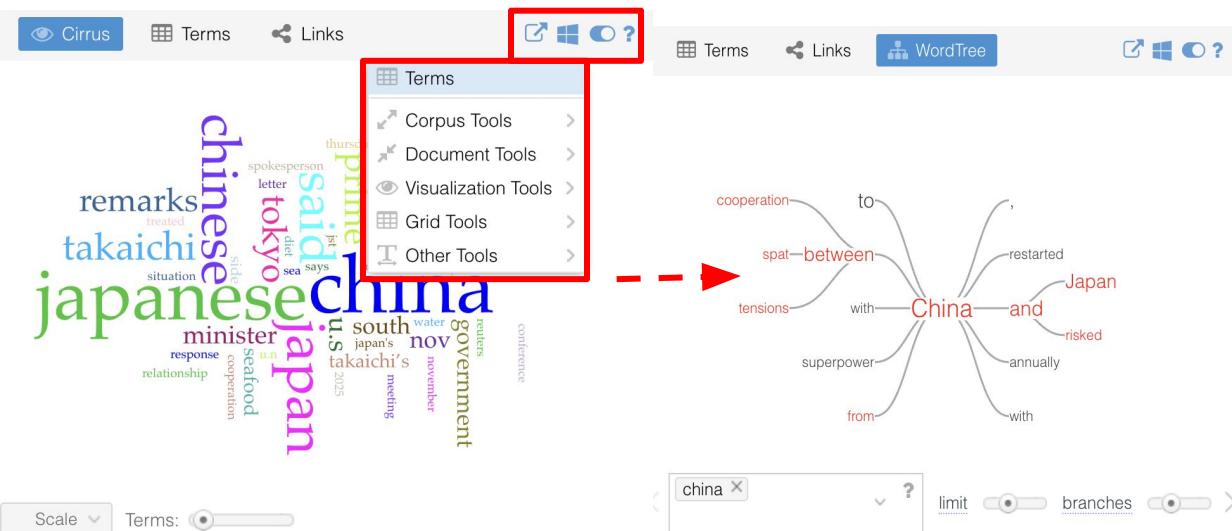


These boxes can all be changed!

*Feel free to ask questions at any point during the presentation!*

# Voyant: Changing displayed results

Hover on the right top corner of a pane and buttons will appear. Select the panes button and choose a new option from the dropdown menu



For our new pane option, we have chosen the WordTree visualization from the 'visualization tools' dropdown sub-menu. You can select the number of "branches" by dragging the scroll button at the bottom and click words to reveal additional branches.

# Voyant: Contexts (concordances)

Contexts, or concordances, show the different contexts around particular search terms. For example, you can see all the times the word “japanese” appears in the text and the contexts in which it appears.

What other  
terms  
should we  
try?

Document	Left	Term	Right
China postpones trilater...	05 JST Printed Chinese and	japanese	flags are seen in this
China reimposes ban o...	China reimposes ban on	japanese	seafood amid 'Taiwan' row THE
China reimposes ban o...	it will suspend imports of	japanese	seafood, citing heightened monitoring over
China reimposes ban o...	plant. Sources related to the	japanese	government confirmed the suspension on
China reimposes ban o...	Taiwan contingency, which would necessit...	japanese	involvement. At a news conference
China reimposes ban o...	spokesperson Mao Ning said, "The	japanese	side has still not provided
China reimposes ban o...	the current circumstances, even if	japanese	marine products are exported to
China reimposes ban o...	responsibility will rest with the	japanese	side," suggesting additional measures. In

japanese ? 33 context   Scale ?

# Voyant: Topics tool

You can view major topics across the corpus or individual documents by hovering over the windows icon and choosing the Topics tool under Corpus or Document tools.

Try changing the number of topics to see how this changes the results.

The screenshot shows the Voyant interface with the 'Topics' tab selected in the top navigation bar. A dropdown menu is open, and the 'Topics' option is highlighted with a red box. Below the dropdown, there is a search bar, a 'Terms' button, a 'Topics' button with a value of '8' (also highlighted with a red box), a 'Run' button, and a 'Toggle diagnostics' button. The main area displays several colored boxes representing different topics, each containing a list of keywords. The topics are: prime tokyo remarks government november meeting group kihara mao power (purple), japan taiwan minister ministry reuters u.n jst response nuclear largest (green), chinese u.s sanae beijing trade sea damaged foreign attack office (pink), south korea export thursday yamazaki united told pacific month 07 (light purple), china japanese said takaichi summit imports spokesperson ban 2025 says (light blue), nov seafood side water relationship news mutually caution past release (orange), takaichi's diet situation treated concrete specific case affairs shimbun strategic (light blue), and letter cooperation japan's china's comments severely trump president ambassador diplomatic (light green).

Feel free to ask questions at any point during the presentation!

# Lexos

# Lexos Summary

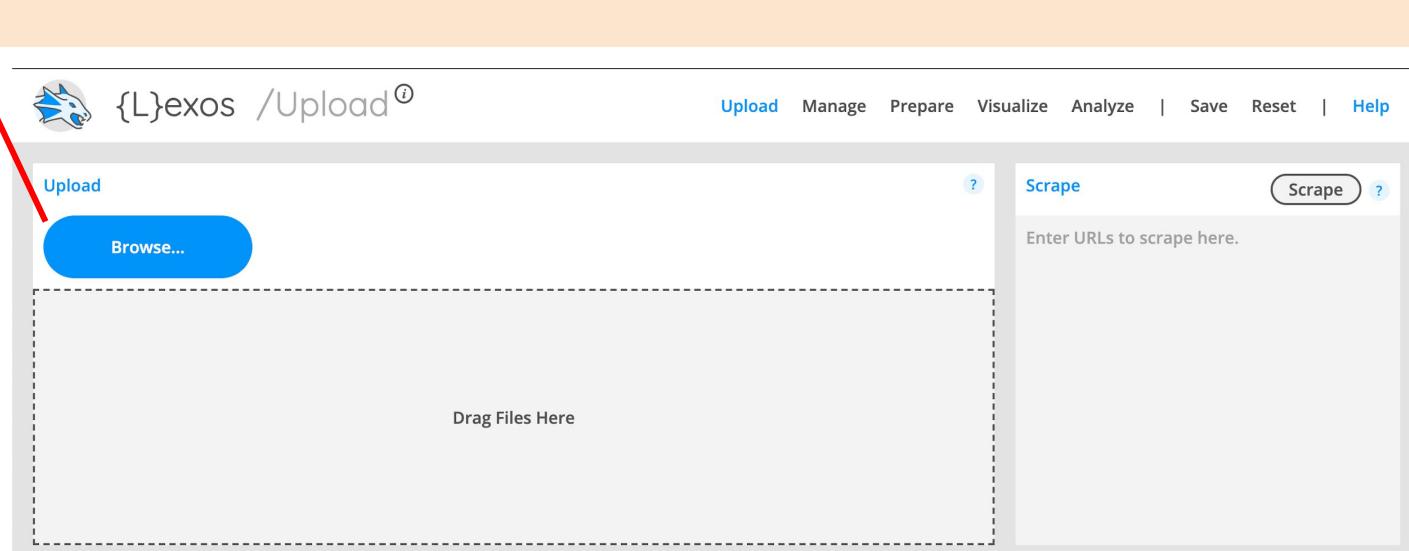
Lexos provides a step-by-step guide for text uploading, preparation, and analysis.

- **Upload:** upload your .txt file
- **Manage:** select the files you want to prepare and analyze
- **Prepare:** prepare your text for analysis
- **Visualize:** create visualizations of patterns across your corpus or in single texts
- **Analyze:** analyze your text

<http://lexos.wheatoncollege.edu/upload>

# Lexos: Upload

Click Browse and select your entire text (or drag file into the “Drag Files Here” area). It can be easy to miss when the upload is done—click “Manage” to double check that the text file is there.



# Lexos: Manage

Make sure the document(s) you want to use are selected  
(blue = selected, gray = not selected)

Active	#	Document	Class	Source	Excerpt	Download	?
	1	China reimposes ban on Japanese seafood amid 'Taiwan' row		China reimposes ban on Japanese seafood amid 'Taiwan' row.txt	China reimposes ban on Japanese seafood amid 'Taiwan' row THE ASAHI SHIMBUN November 19, 2025 at 18:25 JST A Chinese flag flu...ssary procedures, shipments of frozen scallops from Japan to China restarted on Nov. 7—just days before this latest suspension.		
	2	Japan fires back at 'unsubstantiated' Chinese letter to U.N.		Japan fires back at 'unsubstantiated' Chinese letter to U.N..txt	Japan fires back at 'unsubstantiated' Chinese letter to U.N. REUTERS November 26, 2025 at 07:30 JST Japanese Prime Minister S...e to take control of it. The island's government rejects Beijing's claim and says only Taiwan's people can decide their future.		
	3	Kihara urges caution on 'misleading' Taiwan remarks		Kihara urges caution on 'misleading' Taiwan remarks.txt	Kihara urges caution on 'misleading' Taiwan remarks THE ASAHI SHIMBUN November 22, 2025 at 16:57 JST Prime Minister Sanae Tak...ue with the Chinese side now?" (This article was written by Haruka Suzuki, Nobuhiko Tajima, and correspondent Tokuhiko Saito.)		
	4	China says trade cooperation with Japan 'severely damaged' by Taiwan comments		China says trade cooperation with Japan 'severely damaged' by Taiwan comments.txt	China says trade cooperation with Japan 'severely damaged' by Taiwan comments REUTERS November 21, 2025 at 07:10 JST Japanese...et with Takaichi on the side of this weekend's G-20 summit in South Africa, its spokesperson said there were no plans to do so.		
	5	China postpones trilateral culture ministers' meeting with South Korea and Japan, Seoul says		China postpones trilateral culture ministers' meeting with South Korea and Japan, Seoul says.txt	China postpones trilateral culture ministersâ€™ meeting with South Korea and Japan, Seoul says REUTERS November 21, 2025 at 07: ... dispute following a comment by Japan's prime minister about how Tokyo might react to a hypothetical Chinese attack on Taiwan.		

# Lexos: Prepare (Scrub Case and Punctuation)

Lexos demonstrates some more advanced options you have for preparing your corpus. By “scrubbing,” you are transforming the texts in your corpus and making choices that will impact your results. Here are some possibilities:

- **Make Lowercase:** make all your letters lowercase. Even though you know “A” and “a” are the same letter, the computer treats these as two separate characters. Lowercasing removes this distinction.
- **Remove Punctuation:** remove punctuation, which may influence your results.

# Lexos: Prepare (Scrub Words)

You can also stem words and remove certain words. Here are some possibilities:

- **Stop/Keep Words:** remove a list of words. Usually these would be **stopwords**. With WordCounter, you had to use the stopwords list the tool provided—now, you can choose your own.
- **Lemmas:** standardize to the *stem* of word. For example, you can stem all forms of the verb talk: talking, talked, talks, etc. to “talk”

# Lexos: Removing Stopwords

Get a list of English stopwords here:

<https://gist.github.com/sebleier/554280> (there is also a copy on the GitHub page). Copy and paste the stopwords (or upload the .txt file) into the “Stop/Keep Words” box then select “Stop”

The screenshot shows the Lexos interface with the following details:

- Scrubbing Options:** Make Lowercase (checked), Remove Digits (checked), Remove Spaces (unchecked), Remove Tabs (unchecked), Remove Newlines (unchecked).
- Stop/Keep Words:** The "Stop" radio button is selected (highlighted with a red box). The list of words includes: i, me, my, myself, we, our, ours, ourselves.
- Previews:** Preview (highlighted with a blue box), Apply, Download. The preview shows a news article about China reimposing a ban on Japanese seafood.

# Lexos: Applying your Preparations

## BEFORE PREP

Previews

Preview

Apply

Download

[China reimposes ban on Japanese seafood amid 'Taiwan' row](#)

China reimposes ban on Japanese seafood amid 'Taiwan' row

THE ASAHI SHIMBUN November 19, 2025 at 18:25 JST A Chinese flag flu... ssary procedures, shipments of frozen scallops from Japan to China restarted on Nov. 7—just days before this latest suspension.

## AFTER PREP

Previews

Preview

Apply

Download

[China reimposes ban on Japanese seafood amid 'Taiwan' row](#)

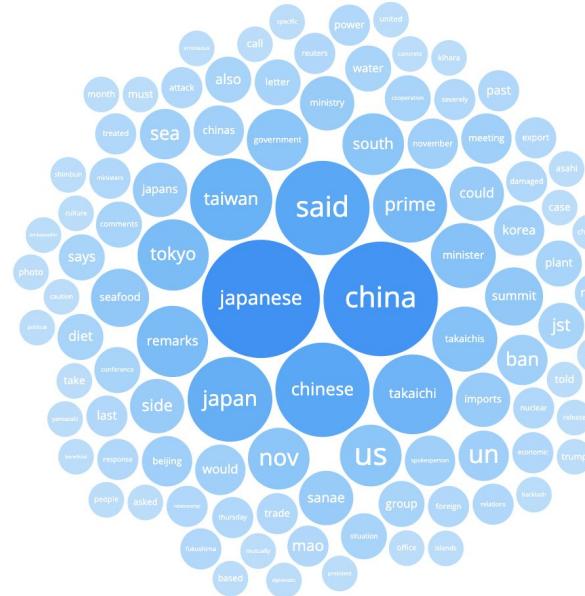
china reimposes ban japanese seafood amid taiwan row asahi shimbun november jst chinese flag flutters wind front great hall p.... except prefectures completing necessary procedures shipments frozen scallops japan china restarted nov days latest suspension

Once you have made decisions about your preparations, click “**Apply**” and wait a few minutes. Because the program is going through each document and completing all the processes you selected, it needs some time. Then, you will see the final results of your preparation! You can also **download** your new corpus and use it with other tools.

# Lexos: Visualize



**Word Cloud:** visualize a word cloud across the entire text/corpus.



Bubbleviz: visualize word counts through bubbles across the entire text/corpus.

# Lexos: Visualize > Multicloud

Font: Open Sans

### Term Count:

80

Color: ( Default

**Generate**

A word cloud visualization showing the frequency of various terms in news articles. The most prominent term is "China", followed by "Chinese", "power", and "plant". Other significant terms include "related", "prime", "sea", "side", "said", "nov", "mao", "japan", "row", "prohibited", "provided", "promised", "refused", "accept", "water", "wind", "people", and "procedures". The size of each word corresponds to its frequency in the dataset.

Japan fires back at 'unsubstantiated' claims from China's state media, which said US officials told the country it must stop its recent moves in the South China Sea. The US has also told China it must stop its recent moves in the South China Sea.

Kihara urges caution on 'misleading' S1 L1 mao us

China says trade cooperation with Ja

**N** Northeastern University  
**NULab for Digital Humanities and  
Computational Social Science**

*Feel free to ask questions at any point  
during the presentation!*

# Lexos: Rolling Window

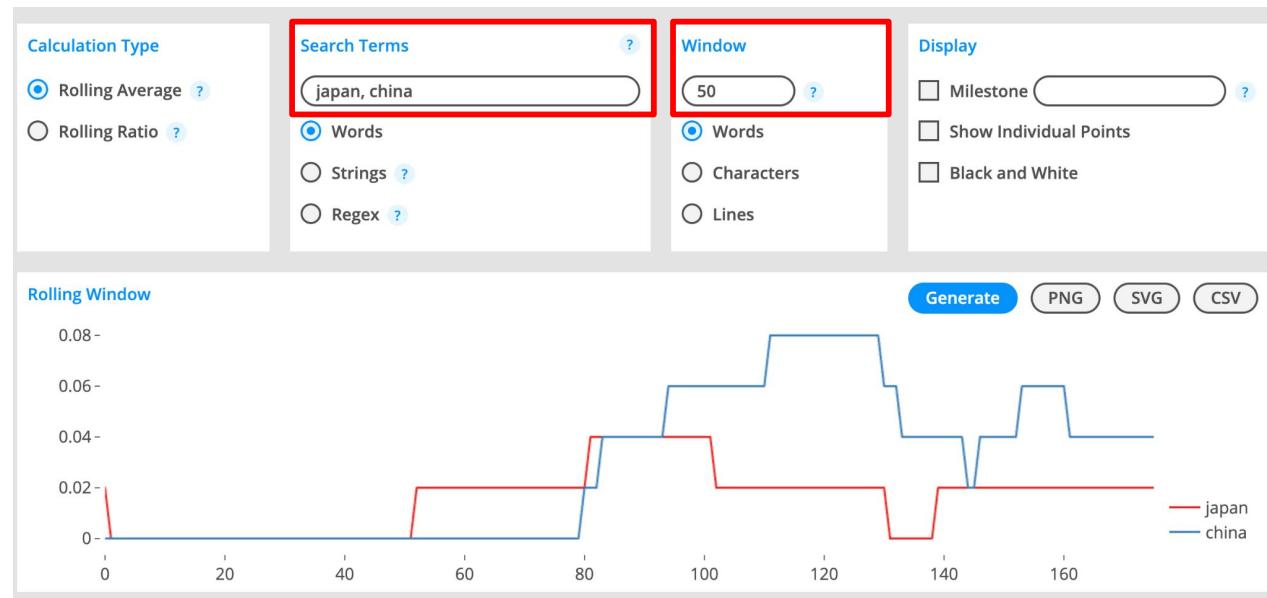
Rolling windows allow you to look at word trends across **one** document. To use a rolling window, first select a single text in the "Manage" screen, then:

1. Go to “Visualize-> Rolling Window” and type in a search term you want to visualize. You can also search multiple terms by clicking “String” and separating words with a comma.
2. Choose a Window size (the number of words each “window” contains). For shorter documents, it’s good to have a number like 300/500. For larger documents, you may want to make your window larger. Play around with the window size until you get a visualization that makes sense.
3. Click “Generate”

# Lexos: Rolling Window Results

Searching for the words ‘japan’ and ‘china’ with a window of 50 (since this is a small document), we can get an idea of how these terms work together in the article.

What other search terms should we try?



# Lexos: Analyze > Top Words

The top words tool lets you compare word usage between individual documents and your corpus as a whole. If you want to make more specific comparisons, you can also assign “classes” to subsets of tools with the “Manage” screen.

- Words with high positive scores are **used more often** in each document, relative to the rest of the corpus.
- Words with high negative scores are **used less often**.

Hit the “Generate” button to see the top words for your texts.

# Lexos: Analyze > Top words

## Top Words

Document "China reimposes ban on Japanese seafood amid 'Taiwan' row" Compared To The Corpus

power 2.5054

treated 2.5054

demanding 2.331

monitoring 2.331

related 2.331

Document "Japan fires back at 'unsubstantiated' Chinese letter to U.N." Compared To The Corpus

letter 3.1839

call 2.463

trump 2.463

un 2.2849

armed 2.0097

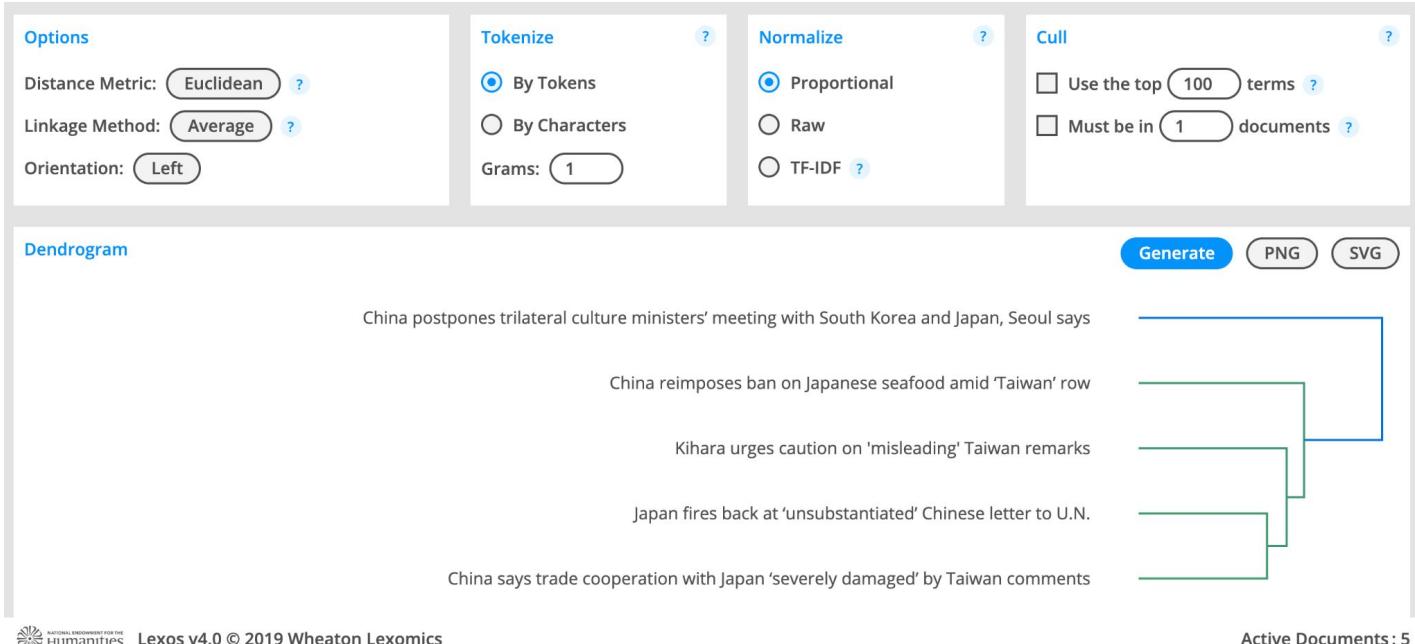
# Lexos: Analyze > Dendrogram

The dendrogram demonstrates similarity between the different documents. Dendograms require at least two documents to compare. Dendograms show:

- Similarities between texts
  - The greater the distance between texts, the less similar they are
  - The smaller the distance between texts, the more similar they are

# Lexos: Dendrogram

The dendrogram demonstrates similarity between the different documents.



Lexos v4.0 © 2019 Wheaton Lexomics

# Lexos: Save or Reset Your Results

Lexos allows you to **save** your results as a Lexos file. If you do this, you can re-upload the Lexos file any time to access your cleaned-up corpus as well as the different analyses you've done. You can also download modified text files from the “Manage” page, which you can use with other tools if you would like.

You can also save individual visualizations as images (PNGs).

Finally, if you want to start over, you can “Reset” your Lexos dashboard.

# Your Turn!

Use the sample or other texts and begin practicing web-browser text analysis.

- Word Counter: <https://databasic.io/en/wordcounter/>
- Word Tree: <https://www.jasondavies.com/wordtree/>
- Voyant: <https://voyant-tools.org/>
- Lexos: <http://lexos.wheatoncollege.edu/upload>

## Discussion Prompts

- What interesting or surprising results came up? What limitations are you observing?
- What kinds of texts would you be curious about comparing?
- Which features do you think will be useful in your future work?

# Thank you!

—Developed by Dipa Desai, Vaishali Kushwaha, Garrett Morrow, Sara Morrell, Ayah Aboelela, Zhen Guo, and Mel Williams

- For more information on the DITI, please see: <https://bit.ly/diti-about>
- Schedule an appointment with us! <https://bit.ly/diti-meeting>
- If you have any questions, contact us at: [nulab.info@gmail.com](mailto:nulab.info@gmail.com)
- We'd love your feedback! Please fill out a short survey here:  
<https://bit.ly/diti-feedback>