Computational Text Analysis for Content Analysis

Developed by **Jonathan Sullivan, Colleen Nugent, Vaishali Kushwaha** for
POLS 1160: International Relations
Carl Cilke
Fall 2020

Northeastern University
NULab for Texts, Maps, and Networks

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

Demo Outline

1. Introduction and Text Preparation
2. Word Counter
3. Word Tree
4. Voyant
5. Conclusion

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

Introduction and Text Preparation

Northeastern University
NULab for Texts, Maps, and Networks

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

Demo Objectives

- Introduction to computational text analysis
- Understand how to build a small textual corpus and prepare the text for analysis
- Understand how web-based computational text analysis programs work, such as in their behind-the-scenes data preparation
- Understand how to interpret the results from your text analysis

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

Computational Text Analysis

Computational text analysis refers to an array of methods that can be used to "read" texts with a computer. This form of analysis can range from basic word frequency counts to more advanced techniques like machine learning.

Text analysis is often used on a **corpus**, or a collection of multiple texts, and provides a glimpse into patterns across the texts. Some people also perform text analysis on larger individual documents, like novels or autobiographies.

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

Why Computational Text Analysis?

Computational text analysis can help us analyze very large amounts of data and discover **patterns** in texts.

Particular disciplines care **deeply** about the language used and how this language may reach intended audiences. Text analysis provides another method for approaching these discourses.

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

How Computational Text Analysis?

Computational text analysis can be conducted using **web-based tools or coding languages like Python and R**.

Many web-based tools are open-access and usually free as long as you have access to a computer and the internet. These tools have limits on how much data they can handle but are usually useful for small projects. Some of these include Voyant, Data Basic.io tools, and Jason Davies' Word Tree tool.

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

Our Text

Our text is a plain text (.txt file) of *COVID-19: Measures affecting trade in goods* compiled by the WTO Secretariat, as of 7 Oct 2020. This list and the information is an informal situation report and an attempt to provide transparency with respect to trade and trade-related measures taken in the context of the COVID-19 crisis.

Our text uses only the information in the column named 'Measure' , but for all the Member/Observer country listed in the data. You may also choose to analyze trade-related COVID measures of a specific country or region, and accordingly limit your text to include measures specific to the selected country/region.

Note that the data preparation is incredibly important for text analysis; always be thoughtful about what you specifically want to analyze.

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

Preparing Our Text

1. Navigate to WTO's COVID-19: Measures affecting trade in goods https://www.wto.org/english/tratop_e/covid19_e/trade_related_goods_measure_e.htm
2. Copy and paste the text into a **plain text editor** (on Macs: Text Edit; on Windows: Notepad)
   a. Mac users, you will need to make your Text Edit into a plain text editor. Open Text Edit, go to Preferences, and make sure "plain text" is selected
3. 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!

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

Word Counter

Northeastern University
NULab for Texts, Maps, and Networks

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

Word Counter

- https://databasic.io/en/wordcounter/
- A user-friendly basic 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
- Can be run with and without **stopwords**

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

Word Counter Examples

TOP WORDS

Word	Frequency
90	314
19	304
covid	246
pandemic	203
00	201
due	197
temporary	176
export	174
hs	158
10	140
certain	103

Shows the top words in the text.

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

BIGRAMS

bigram	Frequency
covid 19	246
to the	233
the covid	215
19 pandemic	203
due to	196
pandemic temporary	90
temporary export	68
on certain	65
6307 90	63
elimination of	59
temporary elimination	57

TRIGRAMS

trigram	Frequency
the covid 19	215
covid 19 pandemic	202
to the covid	198
due to the	196
19 pandemic temporary	90
temporary elimination of	57
personal protective equipment	50
elimination of import	47
of import tariffs	45
temporary export ban	44

It is interesting how many of the trigrams have covid, temporary or elimination!

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

Word Tree

Northeastern University
NULab for Texts, Maps, and Networks

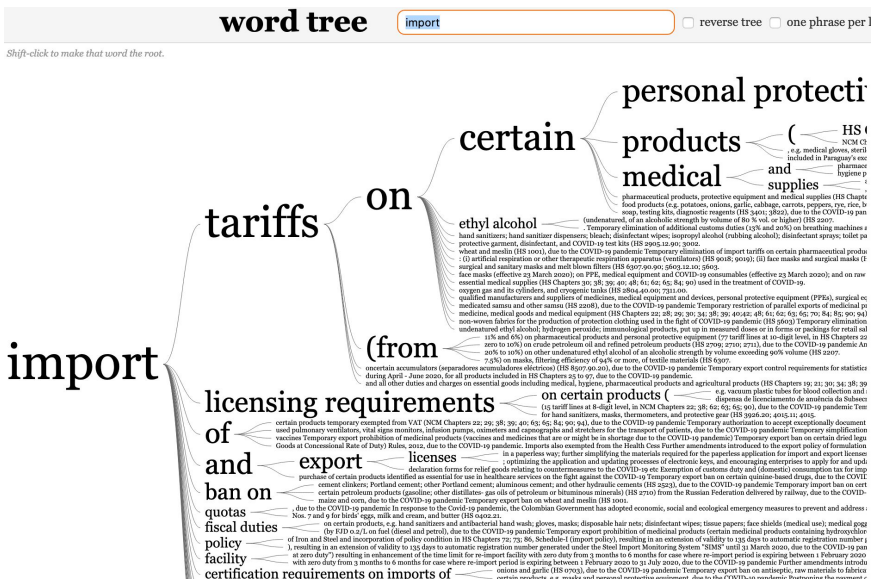
Feel free to ask questions at any point during the presentation!

Word Trees

- 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: fewer than 1 million words should work

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

Word Tree Examples

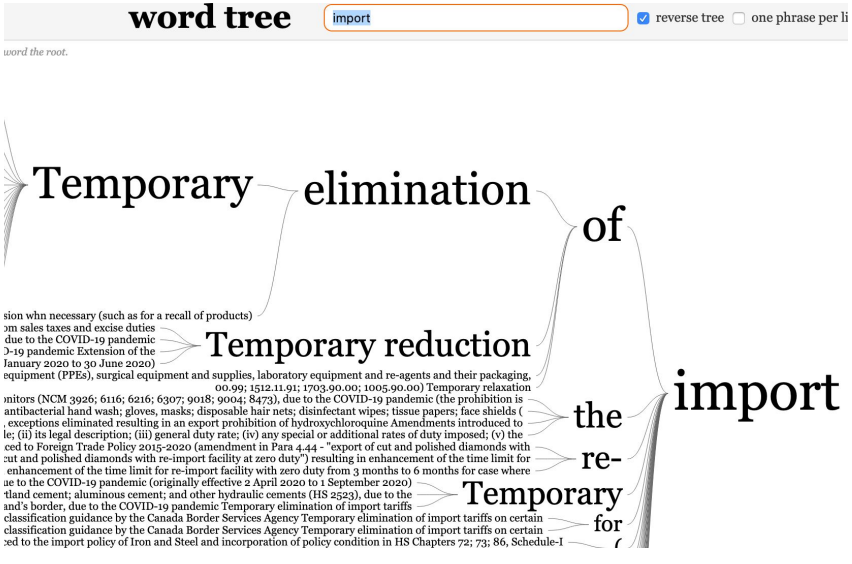


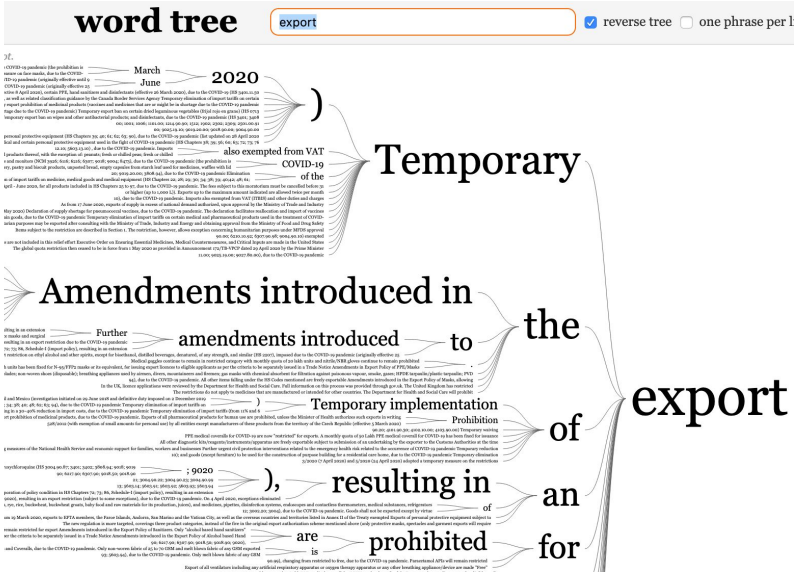
Compare the impact of COVID measures on import and export policies across member countries.

Interesting to know that import tariffs and export bans are only on 'certain' goods: PPE, face maks, medical equipments etc.

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

Word Tree: Reverse Trees Examples





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.

Interesting to know that both import and export measures are ‘temporary’.

Northeastern University
NULab for Texts, Maps, and Networks

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

Voyant

Northeastern University
NULab for Texts, Maps, and Networks

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

Voyant

- https://voyant-tools.org/
- Powerful web-based text analysis platform
- Voyant makes it possible to perform analysis 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!

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

VOYANT

see through your text

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

Alternatively, insert URLs or full text into the textbox.

Click here for help and advanced options

Northeastern University
NULab for Texts, Maps, and Networks

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

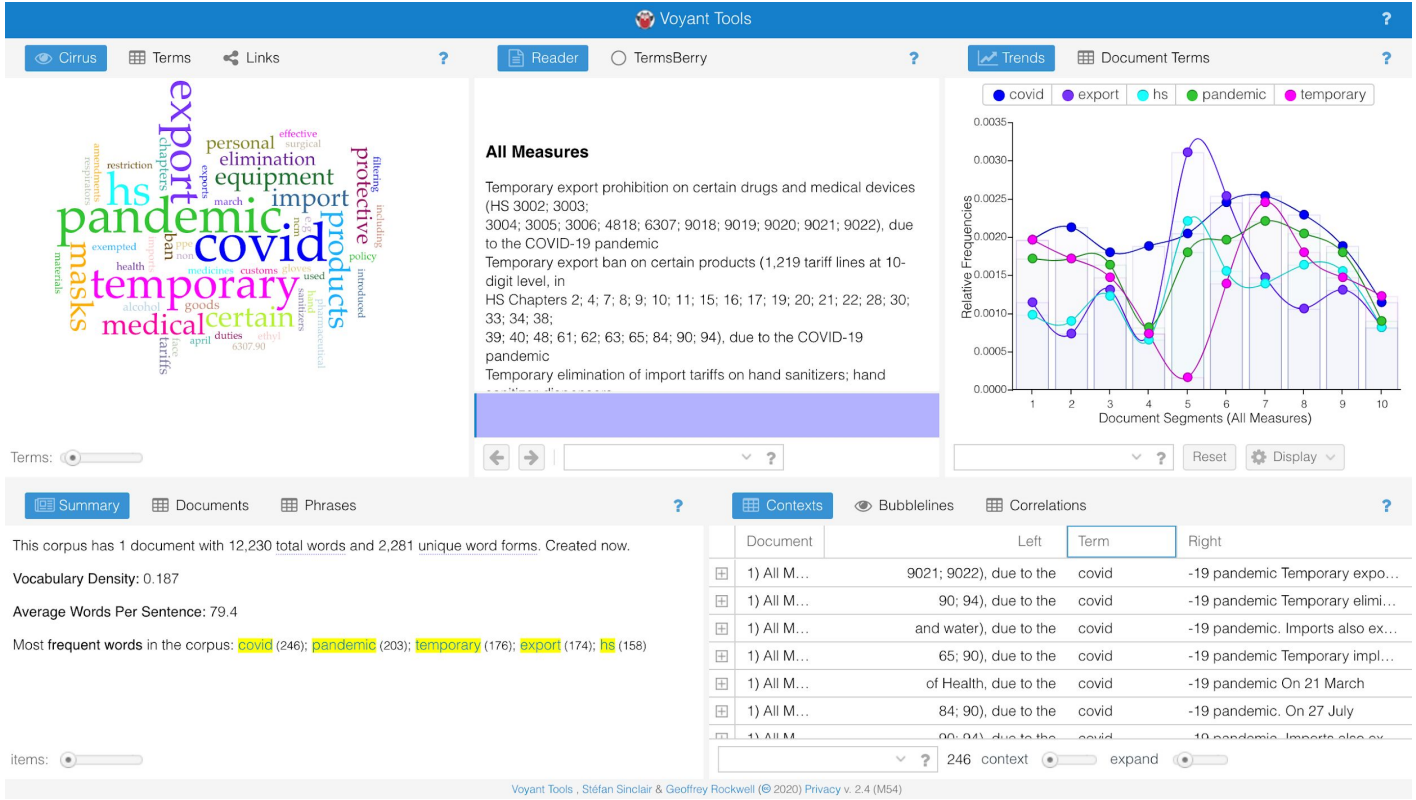
Voyant: Understanding the Dashboard

Results:

From the WTO's COVID-19 measure compilation you can see the default results page with multiple panes:

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

These boxes can all be changed!

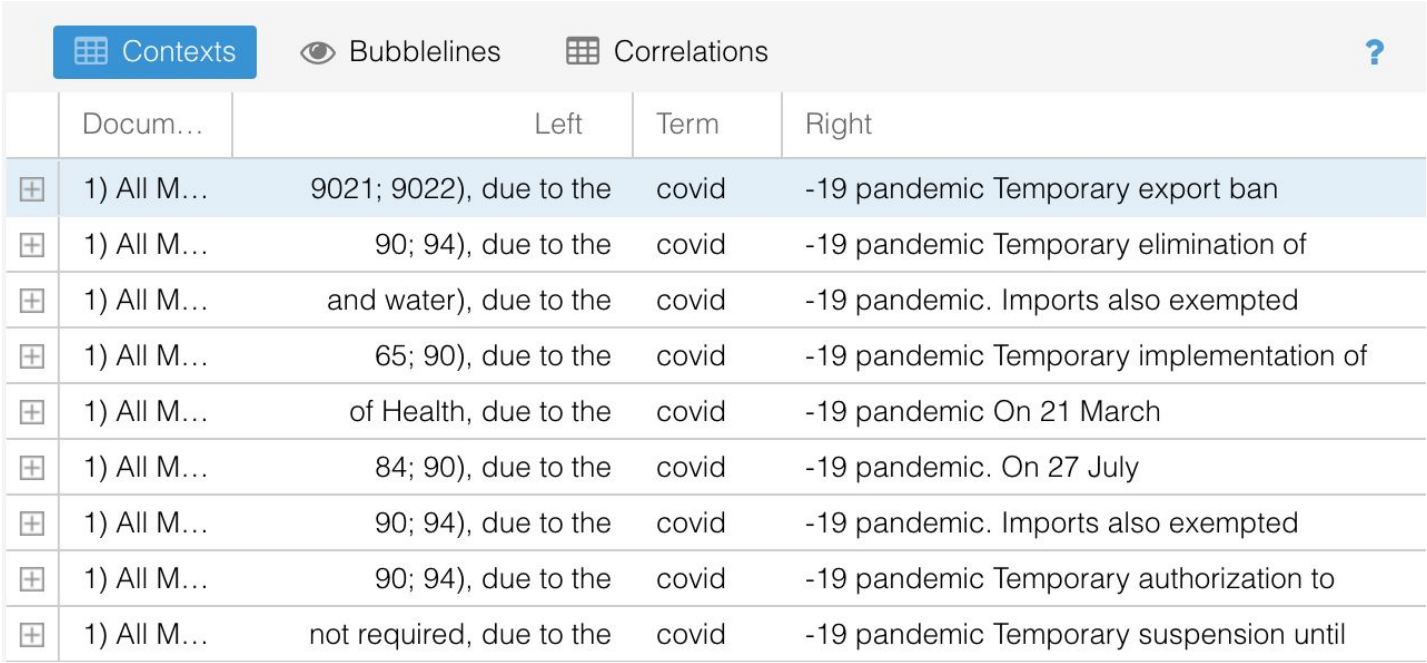


Northeastern University
NULab for Texts, Maps, and Networks

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

Voyant: Contexts (concordances)

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

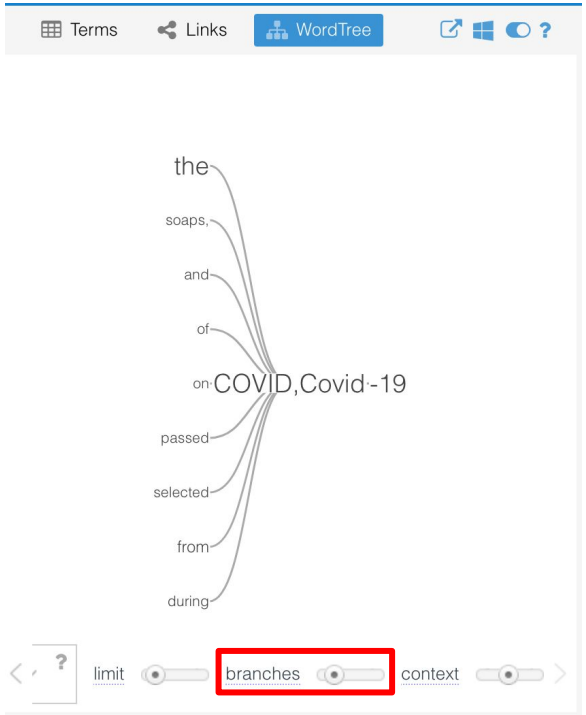


Docum…	Left	Term	Right
1) All M…	9021; 9022), due to the	covid	-19 pandemic Temporary export ban
1) All M…	90; 94), due to the	covid	-19 pandemic Temporary elimination of
1) All M…	and water), due to the	covid	-19 pandemic. Imports also exempted
1) All M…	65; 90), due to the	covid	-19 pandemic Temporary implementation of
1) All M…	of Health, due to the	covid	-19 pandemic On 21 March
1) All M…	84; 90), due to the	covid	-19 pandemic. On 27 July
1) All M…	90; 94), due to the	covid	-19 pandemic. Imports also exempted
1) All M…	90; 94), due to the	covid	-19 pandemic Temporary authorization to
1) All M…	not required, due to the	covid	-19 pandemic Temporary suspension until

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

Voyant: Changing displayed results

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.

Terms:

Northeastern University
NULab for Texts, Maps, and Networks

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

Conclusion

Northeastern University
NULab for Texts, Maps, and Networks

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

Your Turn!

Using the data from *WTO's COVID-19: Measures affecting trade in goods,* begin practicing web-browser text analysis

- Decide to analyze measures for all the countries or compare measures across countries/regions. (Follow your faculty and assignment guidelines)
- Follow the "Preparing Your Text" steps to get your .txt file.
- Explore different WordCount and WordTree features.
- Explore different Voyant features.
- Analyze and save the results for your assignment

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

Post-Exploration Discussion

- What interesting or surprising results came up?
- How might you interpret those results based on what you know about your field or learnt in this course?
- What do you find challenging or exciting about these tools?
- And if you have queries, email us or book an appointment

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

Thank you!

If you have any questions, contact us at **nulab.info@gmail.com**

Developed by Jonathan Sullivan, Colleen Nugent and Vaishali Kushwaha
Delivered by Vaishali Kushwaha and Adam Tomasi
Digital Integration Teaching Initiative
DITI Research Fellows

These slides are available at **https://bit.ly/diti-fall2020-cilke**

Schedule an appointment with us! **http://bit.ly/diti-office-hours**

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