

# An Introduction to Textual Analysis With Voyant Tools

November 4, 2020

Devon Mordell, Educational Developer  
MacPherson Institute, McMaster University

**Do More With Digital Scholarship Series**



Image: [Mhsheikhholeslami](#) (CC 4.0 BY-SA)

# Session Recording and Privacy

*This session is being recorded with the intention of being shared publicly via the web for future audiences.*

*In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.*

*Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.*

# Code of Conduct

*The Sherman Centre and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.*

*As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.*

*Please refer to our code of conduct webpage for more information:*

[scds.ca/events/code-of-conduct/](https://scds.ca/events/code-of-conduct/)



**We're here to help!**

Use **TH:** [query] in chat to  
let the facilitators know  
that you're having  
technical issues

**By the end of  
this workshop...**

***You'll be able to:***

- Describe what kinds of analytical tasks Voyant can do
- Explain how Voyant processes uploaded texts
- Use a Voyant Tool to make observations about a text
- Create a customized dashboard in Voyant

*Voyant tools is a  
web-based reading  
and analysis  
environment for  
digital texts*

## **Good for visualizing:**

- how frequently words appear in a text
- which words appear together (i.e. are co-located)
- where words appear in a text

## **Works best with:**

- [large text corpora]
- born-digital texts\*

# Distant reading



In contrast to close reading...



...using computational methods  
to analyze a collection of texts



# Exploratory Data Analysis

## Visualization → pattern recognition:

- surface data collection or cleaning errors
- highlight data anomalies
- test underlying assumptions
- refine research questions

# Data sources

## From:

- Websites
- Social media APIs
- Documents (scanned or born-digital)
- Text collections - [Project Gutenberg](#) and [DH text collections](#)

# Assembling the corpus

## “Corpus”

A collection (body) of documents

Consider:

- what should be included or omitted?
- what pre-processing tasks must be performed?
- what errors might affect your analysis?
- how should you arrange your documents?



# Pre-processing your data

## Voyant will:

- guess the file format
- ignore punctuation and symbols
- parse & index corpus text based on specified delimiters, or **tokenize** text
- apply **stopwords** dictionary to omit common terms unlikely to be significant (e.g. “the,” “and”)

# Pre-processing your data

## Voyant will not:

- correct misspelled or merged words (via OCR)
- cluster spelling variants
- translate HTML entities or Unicode characters

But errors can be easier to discover in Voyant...

# On data provenance

I.e. where the data came from and  
what kinds of transformations were  
performed on it

As with any other research,  
**documentation** → **trustworthiness**





# Canadian Federal Election Party Platforms

**The  
dataset**



## Merged words, split words, unrecognized characters – oh my!

we're making life easier for parents  
who want to build a better life for their  
kids, we're doing more to protect our  
environment, and we're keeping our  
economy strong and growing in an  
increasingly difficult world.



Here is just some of what we've been able to do in the last four years:

Introduced and increased  
the Canada Child Benefit to help families keep up with the cost of living  
Raised taxes on the wealthiest 1% and cut taxes for the middle class  
Building more affordable housing with Canada's first-ever National Housing Strategy  
More communities, and are on track to eliminate all of them by 2021  
Lowered the small business tax rate to help small businesses grow and create more jobs  
Moving implementing National Pharmacare  
Moving forward with a ban on single-use plastics  
Strengthened the Canada Pension Plan to give Canadians  
a secure retirement  
This election, we all have a choice to make. We can keep moving forward  
and build on the progress we've made, or  
we can go back to the  
hurtful cuts of the Conservative years.

# Getting started with Voyant

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

A screenshot of the "Add Texts" interface in Voyant Tools. The interface has a light grey header with the title "Add Texts" and three small icons (a magnifying glass, a globe, and a question mark). Below the header is a large text input field with a light blue border and a placeholder text that reads "Type in one or more URLs on separate lines or paste in a full text." At the bottom of the interface, there are three buttons: "Open" (with a folder icon), "Upload" (with a document icon), and "Reveal" (a blue button with a white checkmark icon).

*Voyant Tools is a web-based reading and analysis environment for digital texts.*

## Alternatively: Voyant server

- Download Voyant Server and run locally
  - Java app – download Java if not already installed
- Mac Users:
  - may have to open via terminal
    - `java -jar VoyantServer.jar`

# Adding texts to Voyant

## To add texts to the environment:

1. Upload documents (.txt, .doc, .csv, .html, .xml, .pdf, .zip, .json...)
2. Paste URLs (must link to a readable file format)
3. Paste plain text
4. Choose from existing 😞

## Before you “reveal”

→ consider your options...

The image shows a software interface with an 'Options' dialog box open. The dialog box has a title bar with a close button (X) and contains the following elements:

- Input Format:** A dropdown menu set to 'Auto-Detect (recommended)'.
- Text:** A paragraph of text: "For more information on the advanced options below, see the documentation on [creating a corpus](#)."
- Collapsible Sections:** A list of sections with expand/collapse arrows: Corpus, Text, XML, HTML, JSON, Tables, Processing, and Access Manager.
- Processing Section:** Expanded, showing:
  - Text Language:** A dropdown menu set to 'Auto-detect'.
  - Tokenization:** A dropdown menu with 'Automatic (highly recommended)' selected. The menu is open, showing options: 'Automatic (highly recommended)', 'Simple Word Boundaries', and 'Whitespace Only'.

In the background, there is a window titled 'Add Texts' with a text input field and 'Open' and 'Upload' buttons. To the right, a 'Reveal' button with a checkmark is visible. A large purple arrow points from the text 'Open “options”' to the 'Options' dialog box.

# First Things First...

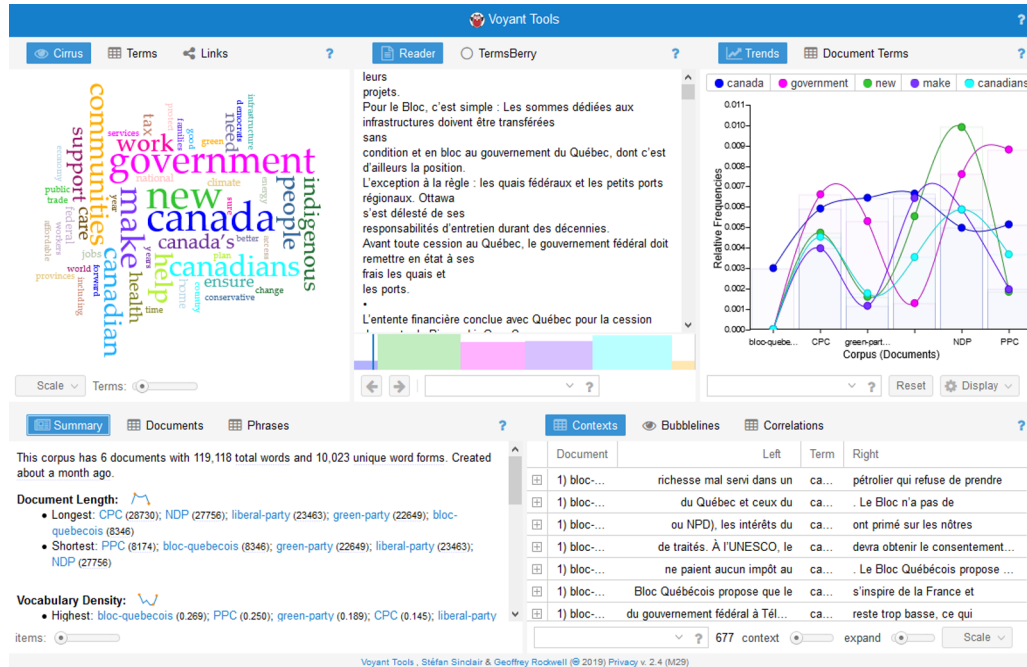
You can return to your corpus and continue to work on it later, but:

**ensure that you bookmark the URL**

→ it is the only way to access your corpus again

# Dashboard

→ the whole enchilada...









# Contexts Tool

→ active word in the contexts in which it appears (re: sense)

Contexts Bubblelines Correlations ?

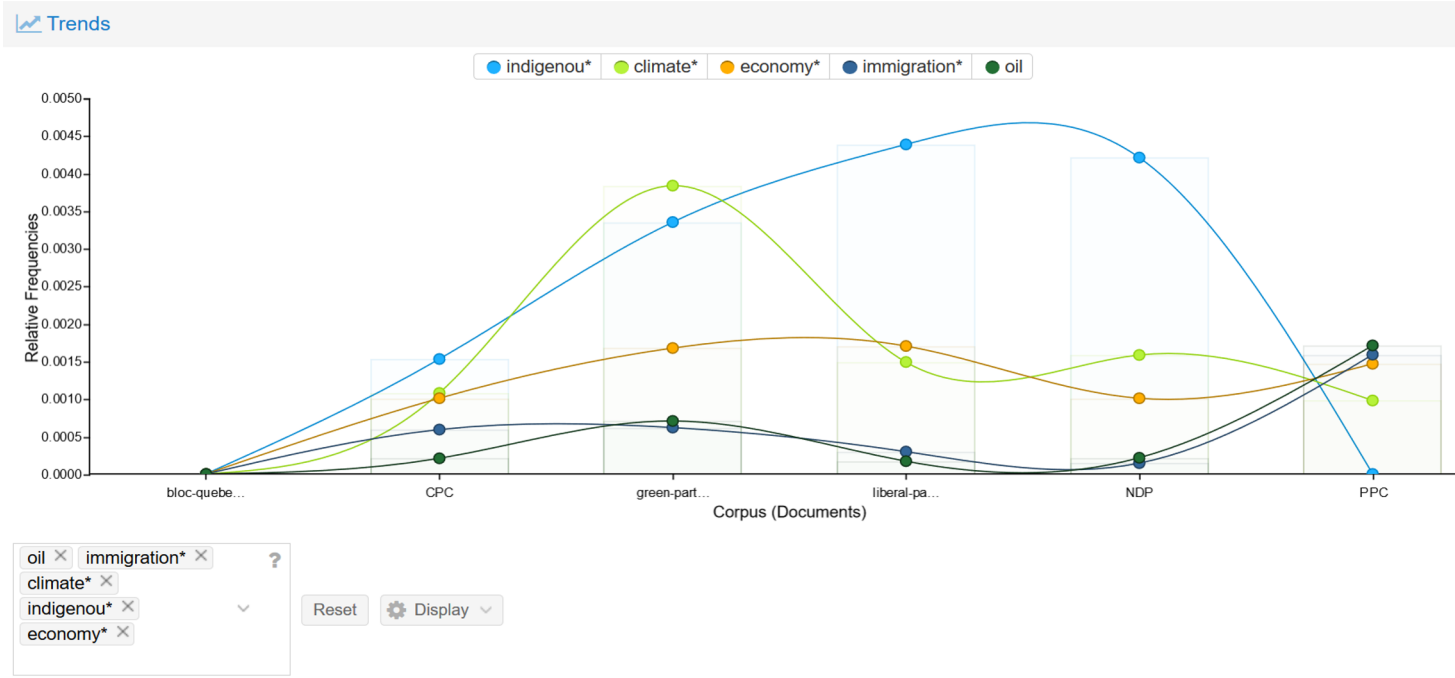
Document	Left	Term	Right
1) CPC	Innovation to Fight Climate Change	green	Technology, Not Taxes A Cleaner
1) CPC	savings for maternity leave, transit,	green	home renovations, and your kids
1) CPC	in Your Pocket ♦ Introduce the	green	Public Transit Tax Credit To
1) CPC	a 20-year period. The	green	Public Transit Tax Credit will
1) CPC	end of the workday. The	green	Public Transit Tax Credit is
1) CPC	hundreds of dollars with the	green	Public Transit Tax Credit. A
1) CPC	today's labour market. Introduce the	green	Home Renovation Tax Credit To
1) CPC	help cover the cost of	green	home renovations between \$1,000 and
1) CPC	your home's energy efficiency. These	green	home improvements will save you
1) CPC	Money in Your Pocket ♦ The	green	Home Renovation Tax Credit will

green\* x ? 220 context expand Scale v



# Trends Tool

→ compare relative frequencies of multiple terms across multiple documents



# Modifying a corpus

Documents

	Title	Words	Types	Ratio	Words/Sentence
1	CPC	28,7...	4,157	14%	25.1
2	NDP	27,7...	3,457	12%	27.9
3	PPC	8,174	2,043	25%	22.0
4	green-party	22,6...	4,277	19%	31.2
5	liberal-party	23,4...	2,973	13%	33.7

0

+ Add   - Remove   ✓ Keep   ⇅ Reorder   ✕ Cancel



**There's more?  
Much more...**

Choose another tool for this panel area



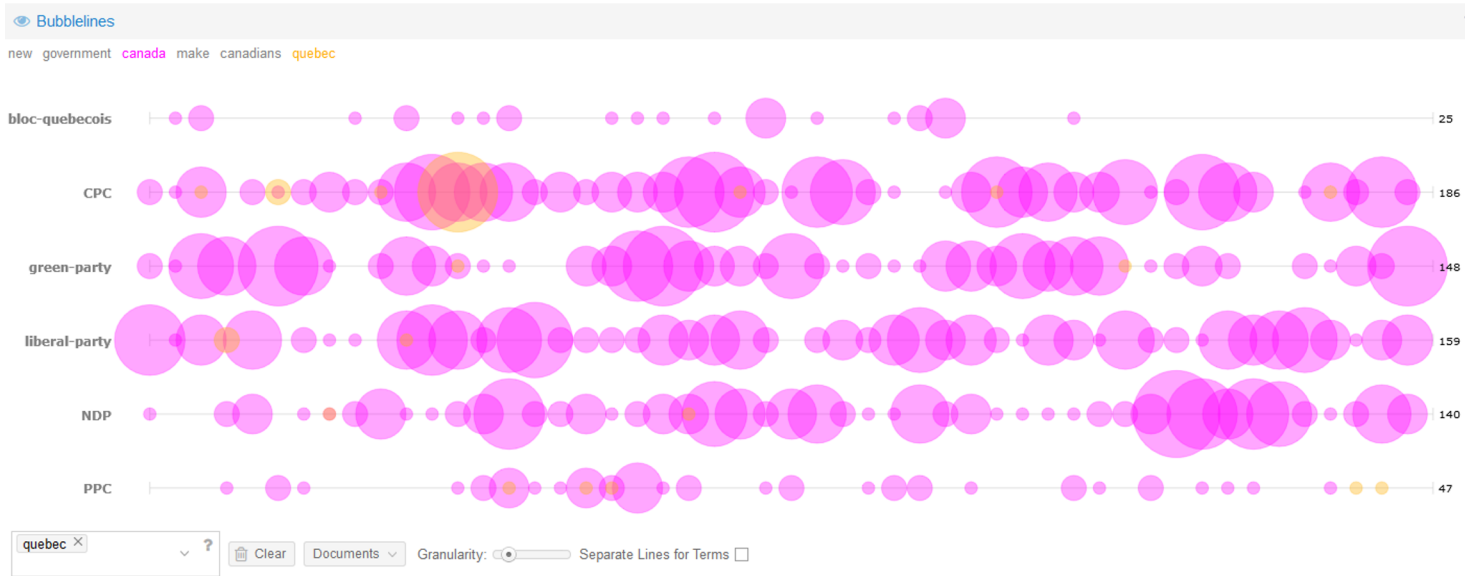
Contextual help

Open tool in own window

Define settings for tool

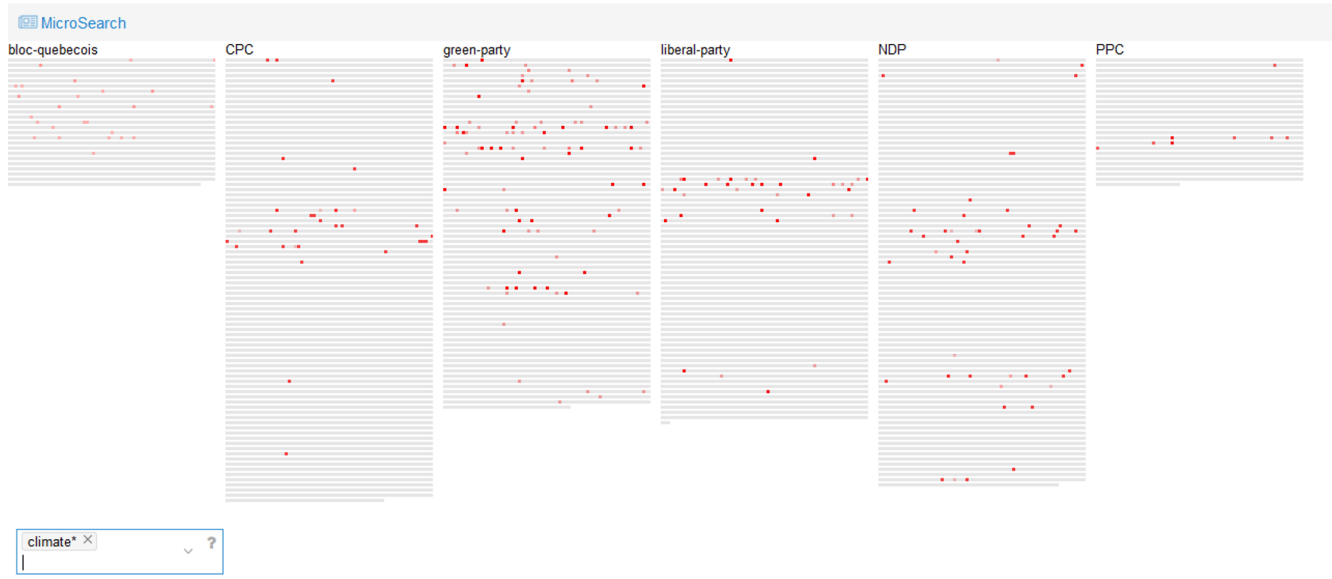
# Bubblelines Tool

→ frequency of term occurrence within segments of equal length



# Microsearch Tool

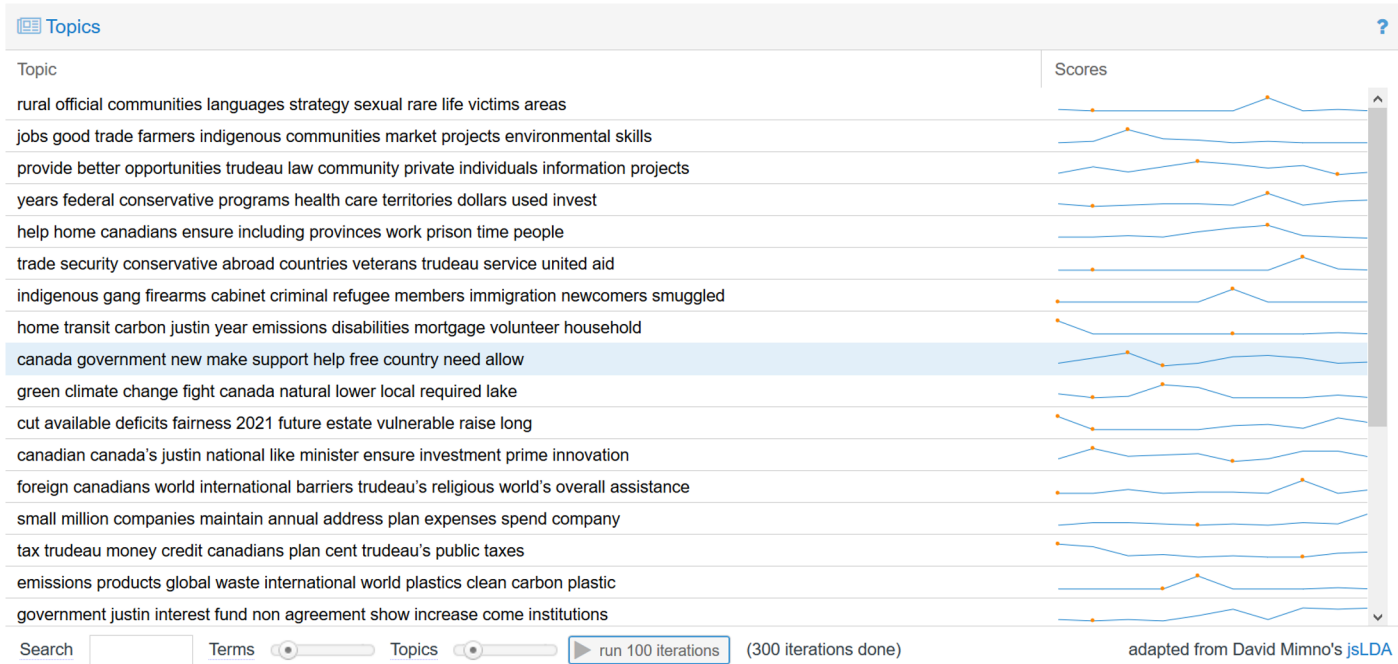
→ relative location of active term in document





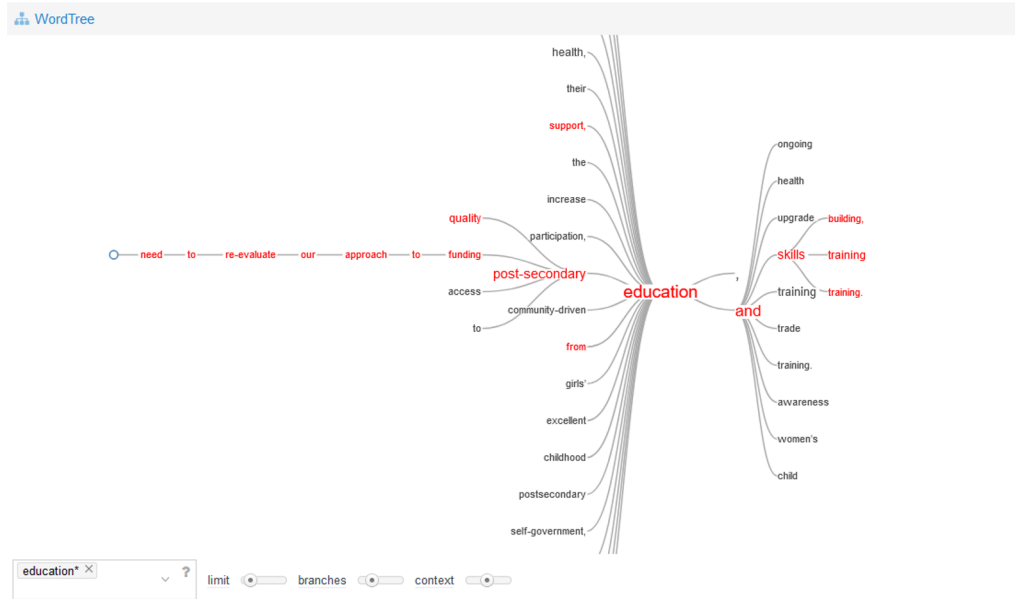
# Topics Tool

→ term clusters based on co-occurrence in the document



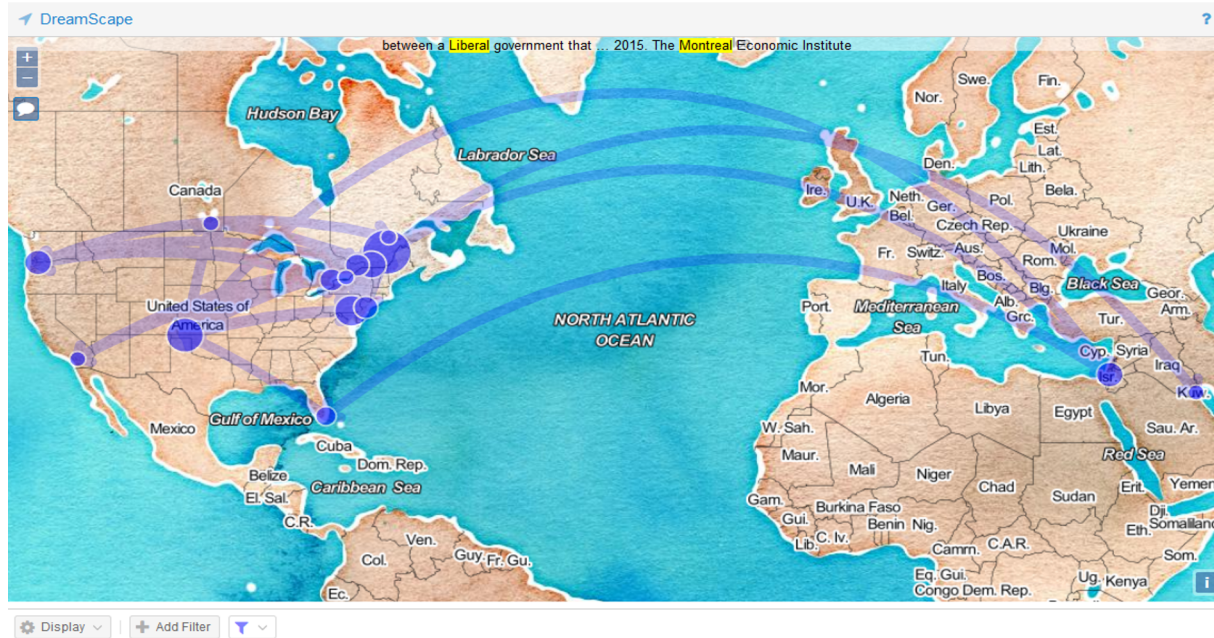
# Word Tree Tool

→ concordance again, but more dynamic in nature



# Dreamscape Tool

→ geolocation of cities in texts



# Use in teaching

Ask learners to analyze primary and/or secondary sources for their research topic to identify key themes and disciplinary language.

# Use in teaching

Ask learners to compare trends across two texts (or an author's entire oeuvre, etc.)

# Use in teaching

Ask learners to do a close reading of a text they have analyzed with Voyant (or vice versa) and to contrast their respective findings.

Use in  
teaching

Ask learners to upload their own writing and observe any overused words or other patterns.

# Use in teaching

Ask learners to create a digital exhibit to display the results of their analysis with embedded Voyant tools.

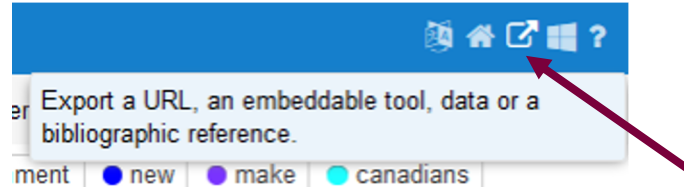


## Customizing your dashboard

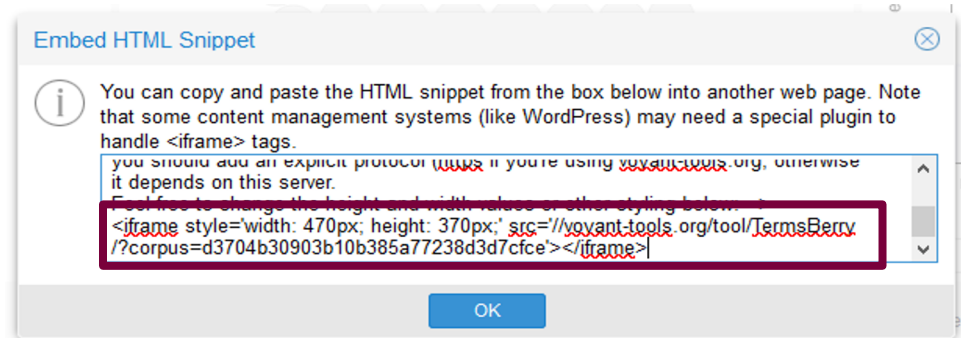
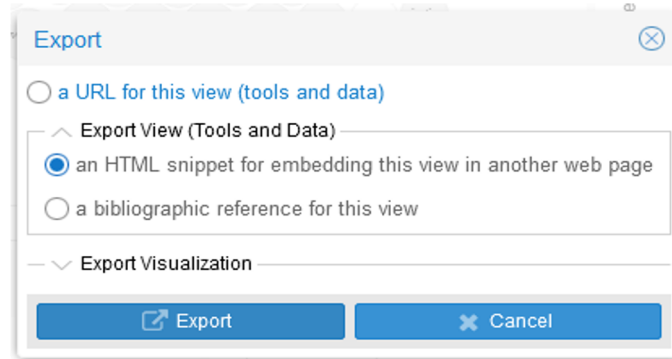
1. Arrange the dashboard with tools
2. Export URL for the entire dashboard (from main Voyant navigation bar)

E.g.

[corpusURL]&panels=cirrus,microsearch,trends,bubblelines,topics



# Embedding Tools



# Embedding Tools

Source

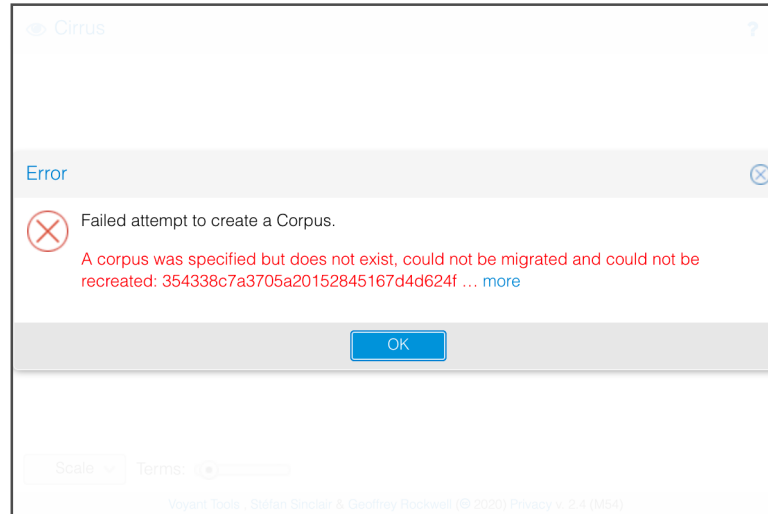
```
<p><iframe src="//vovant-tools.org/tool/Cirrus/?stopList=keywords-a02d31f3a084ca331ac4f8b5cc9d91ab&amp;
stopList=%amp;visible=55&corpus=66c1ae56ced5afb12142391cb76cac6c" style="width: 597px; height:
379px;"></iframe></p>
```

Paste copied code into a text editor & modify URL to an absolute link (add "https:" to start of URL).

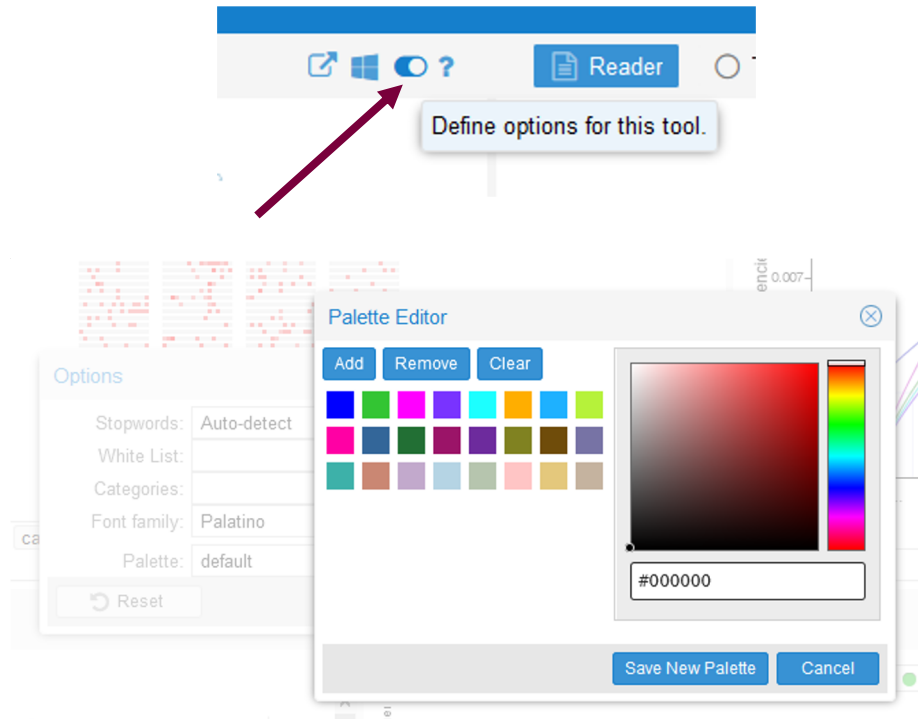
When embedding, consider permanence...

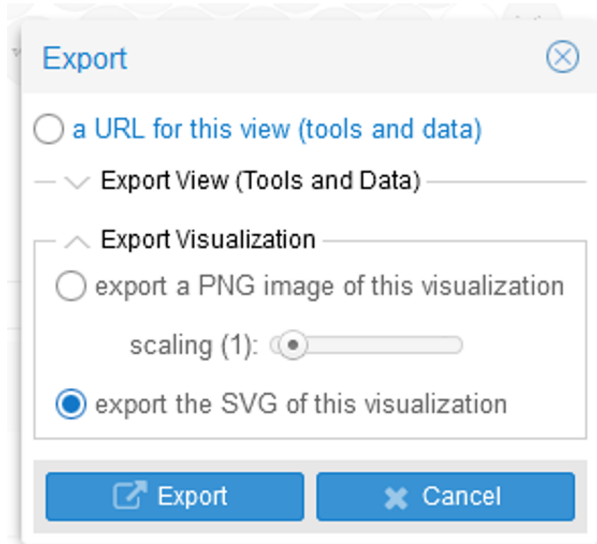


#MAGA - NOVEMBER 6, 2017



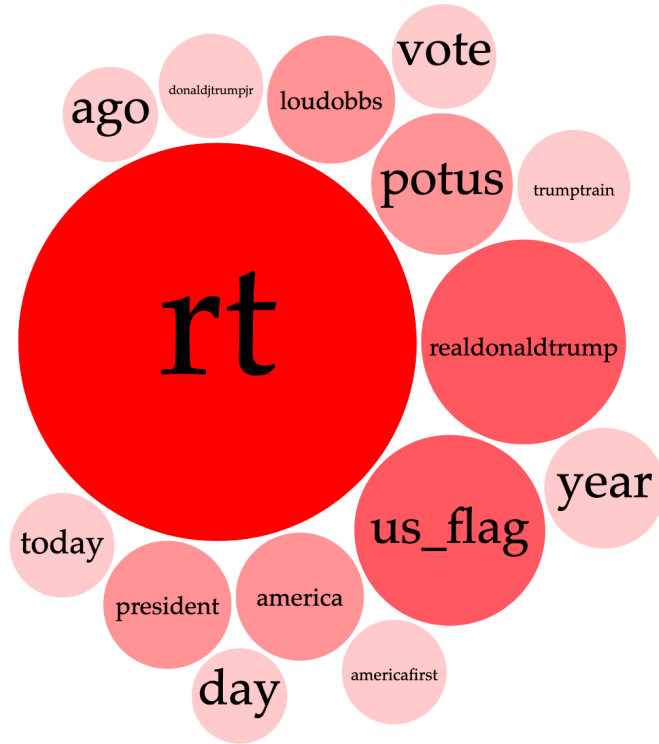
# Customizing Visualizations





Using **Inkscape**,  
any tool with an SVG  
output can be edited





## Other tools...

Jigsaw - named entity recognition, topic modelling

MALLET - topic modelling

Natural Language Toolkit (NLTK) in Python -  
sentiment analysis & more



## References & resources

- Christie, Alex. n.d. “Voyant Tools.” Pedagogy Toolkit for English. Accessed October 29, 2020. <http://pedagogy-toolkit.org/tools/VoyantTools.html>.
- “Examples Gallery | Voyant Tools Documentation.” n.d. Accessed October 29, 2020. <http://digihum.mcgill.ca/voyant/about/examples-gallery/>.
- Sinclair, Stefan, and Geoffrey Rockwell. 2012. “Teaching Computer-Assisted Text Analysis: Approaches to Learning New Methodologies.” In *Digital Humanities Pedagogy: Practices, Principles and Politics*, edited by Brett D. Hirsch (ed.). Open Book Publishers. <https://doi.org/10.11647/OBP.0024>.

# Thank you for coming!

Feel free to get in touch with questions!

[mordelld@mcmaster.ca](mailto:mordelld@mcmaster.ca)