Intro to Computational Text Analysis

By Adam Tomasi and Milan Skobic
Digital Integration Teaching Initiative (DITI)

Advanced Writing for the Sciences - ENGW 3307

Cecelia Musselman

Spring 2021

Northeastern University
NULab for Texts, Maps, and Networks

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

Workshop Agenda

- Introduction to definitions and key terms in computational text analysis (What)
- Discussion on its applications and uses in research (Why)
- Demonstration of web-based text analysis tools (How)
  - Word Counter, Word Trees, Voyant, Lexos

Slides, handouts, and data available at
http://bit.ly/diti-spring2021-musselman2

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

Workshop Objectives

- Understand best practices for collecting and storing textual data when performing basic computational text 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 explorations

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

Introduction

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

Computational Text Analysis

Text analysis is making inferences based on textual data. Computational text analysis refers to an array of methods that can be used to “read” texts with a computer. It is similar to statistical analysis, but the data are texts.

- It involves a computer drawing out patterns in a text, and a researcher interpreting those patterns.
- It includes methods such as word count frequency, nGrams, and sentiment analysis.

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.

We might discover patterns that a close reading of each document individually would not have uncovered.

Computational methods also let us more efficiently analyze large corpora, saving the time and effort demanded by reading hundreds of text files.

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

Key Terms

- **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.
- **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.

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

Text Preparation

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

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?

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

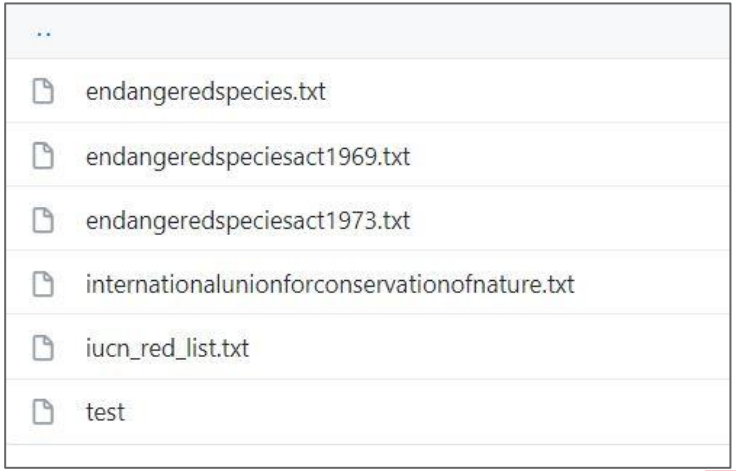
Preparing Your Text

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. Copy and paste the 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'. 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.

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

Our Text

We will be using five Wikipedia entries related to endangered species and institutional efforts to protect them. These .txt files are available at the GitHub link shared earlier.



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

Exploratory Tool: Word Counter

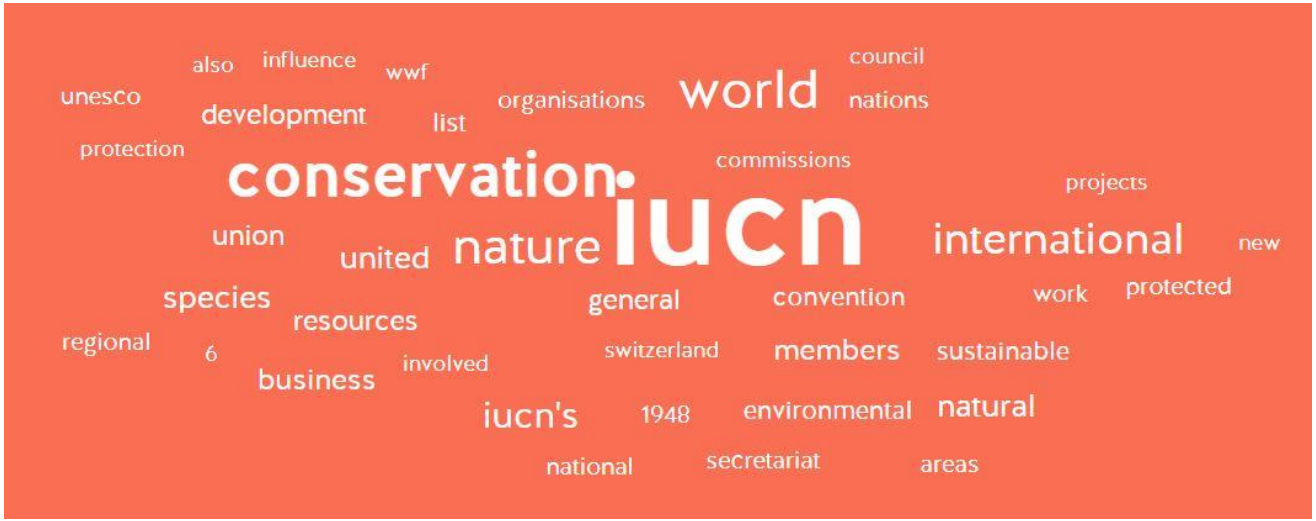
Feel free to ask questions at any point during the presentation!

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
- Default is lowercase all words and apply stopwords
- It can be run with and without stopwords

Word Counter Examples

This is a "**word cloud**". It is helpful to get a sense of the **most used words in a document**. Words used more often are bigger, and ones used less often are smaller.



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

Word Counter Examples

TOP WORDS

Word	Frequency
iucn	119
conservation	53
world	42
nature	40
international	31
iucn's	24
natural	19
species	18
united	18
business	17

Shows the top words in the text.

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

BIGRAMS

bigram	Frequency
the world	25
the iucn	25
of the	25
and the	23
of nature	18
in the	16
world conservation	14
with the	14
for the	14
natural resources	12

TRIGRAMS

trigram	Frequency
international union for	8
the united nations	8
conservation of nature	7
the world conservation	7
of natural resources	6
the iucn red	6
iucn red list	6
red list of	6
and the world	6
the protection of	6

The United Nations is prominent as a trigram, as is “IUCN Red List” -- already these suggest interesting threads for analysis!

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

Exploratory Tool: Word Trees

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

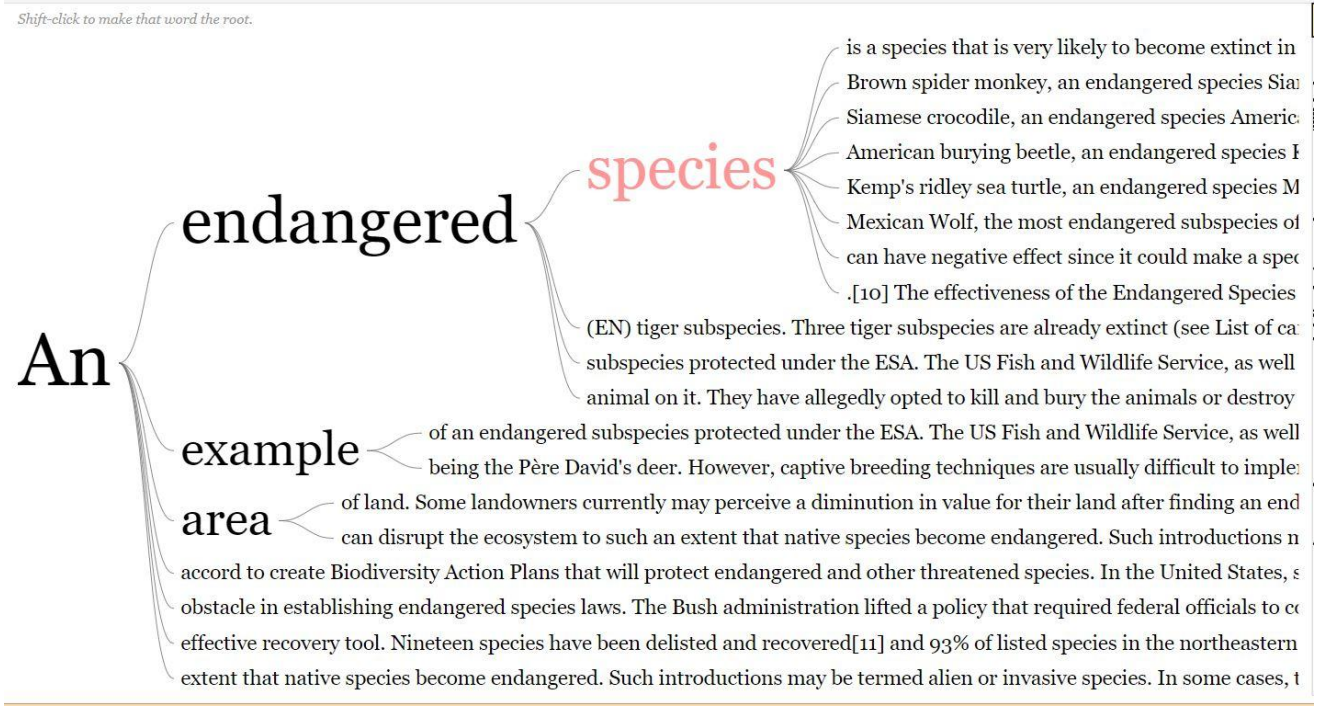
Word Trees

- 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 Example

"Endangered" and "species" are common roots, but you can see that the Wikipedia entry is using case studies ("example") and addressing land questions ("area").

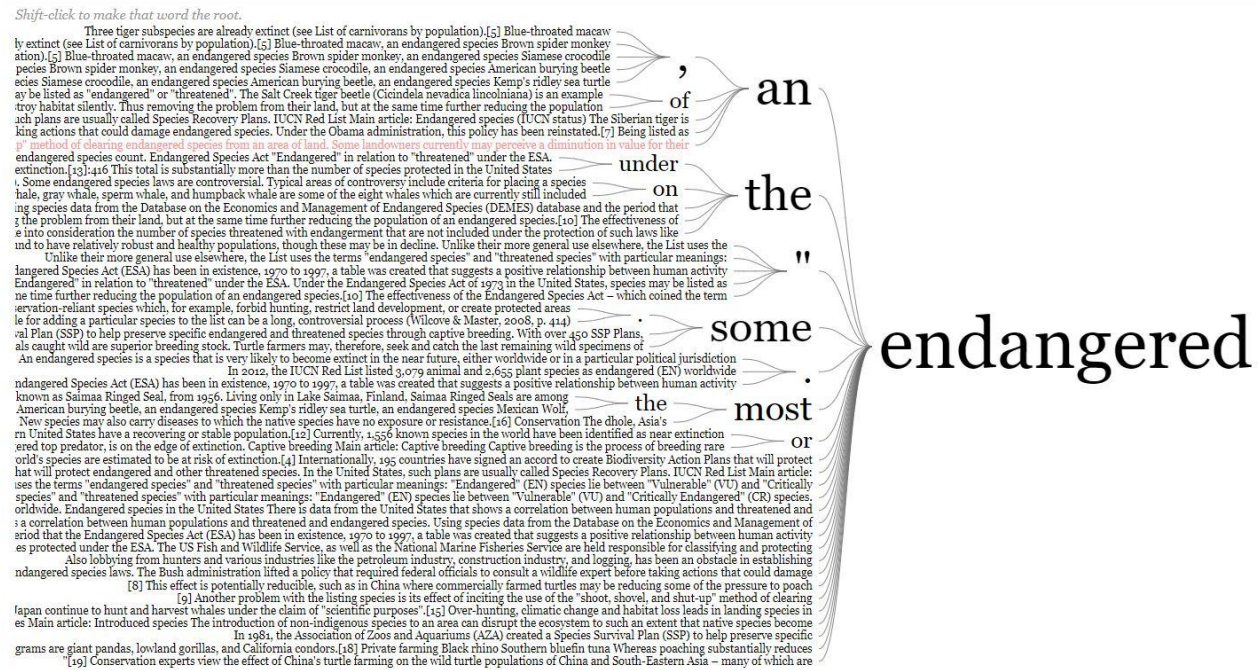


Feel free to ask questions at any point during the presentation!

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.

Here the Wikipedia entry makes comparisons such as delineating which species are “the most endangered.”



Northeastern University
NULab for Texts, Maps, and Networks

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

Voyant

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

Voyant

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/

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

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

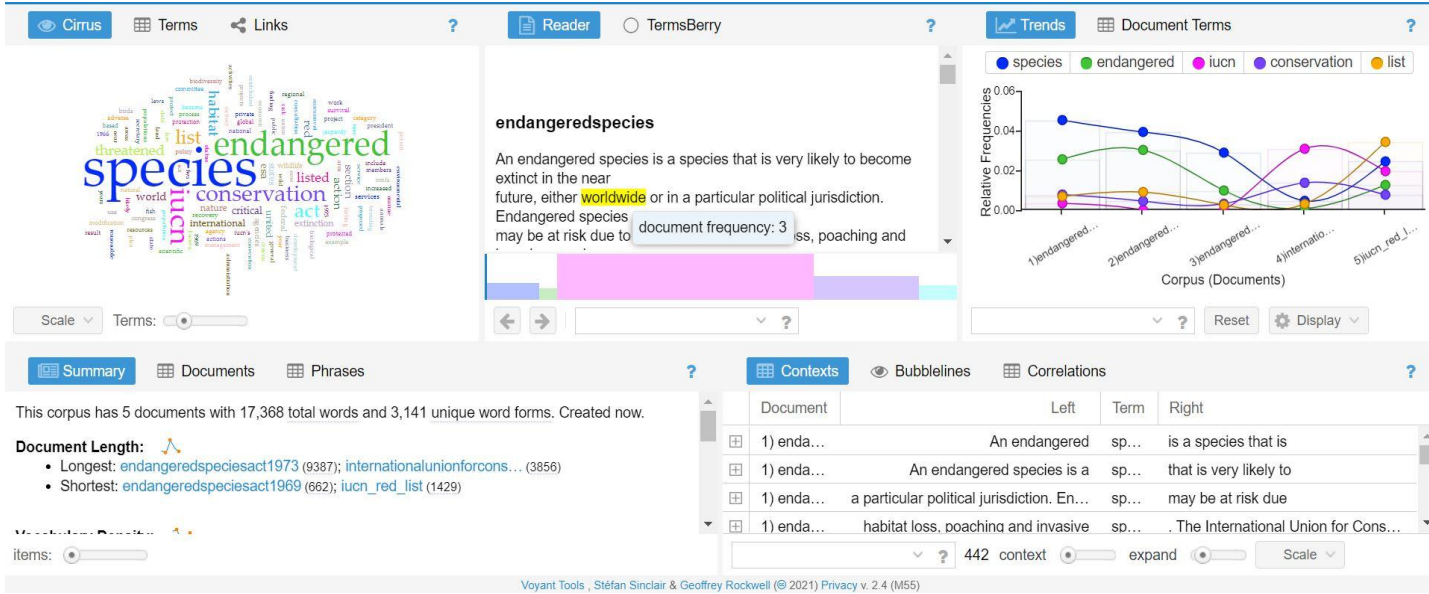
Voyant: Understanding the Dashboard

Results:

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!



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 “boston” appears in the text and the contexts in which it appears.

Contexts

Document	Left	Term	Right
1) enda…	due to factors such as	habitat	loss, poaching and invasive species
1) enda…	such as captive breeding and	habitat	restoration. Conservation status Main article
1) enda…	bury the animals or destroy	habitat	silently. Thus removing the problem
1) enda…	Over-hunting, climatic change and	habitat	loss leads in landing species
3) enda…	attention as unregulated hunting and	habitat	loss contributed to a steady
3) enda…	pairs remained.[10] Loss of	habitat	, shooting, and DDT poisoning contributed
3) enda…	species do not have enough	habitat	for long-term survival. These
3) enda…	the next few decades without	habitat	restoration.[21] Along with other
3) enda…	climate change, land use change,	habitat	loss, invasive species, and overexploitation
3) enda…	Endangered Species Act by reducing	habitat	protections for at-risk species
3) enda…	Section 4). If determinable, critical	habitat	must be designated for listed
3) enda…	species' existence or destroy critical	habitat	(Section 7). Any import, export
3) enda…	Section 4 also requires critical	habitat	designation and recovery plans for
3) enda…	modification, or curtailment of its	habitat	or range. 2. An over
3) enda…	in the provision on critical	habitat	designation.[45] The 1978 amendment

habitat ? 68 context expand Scale

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

Lexos

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

Lexos

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

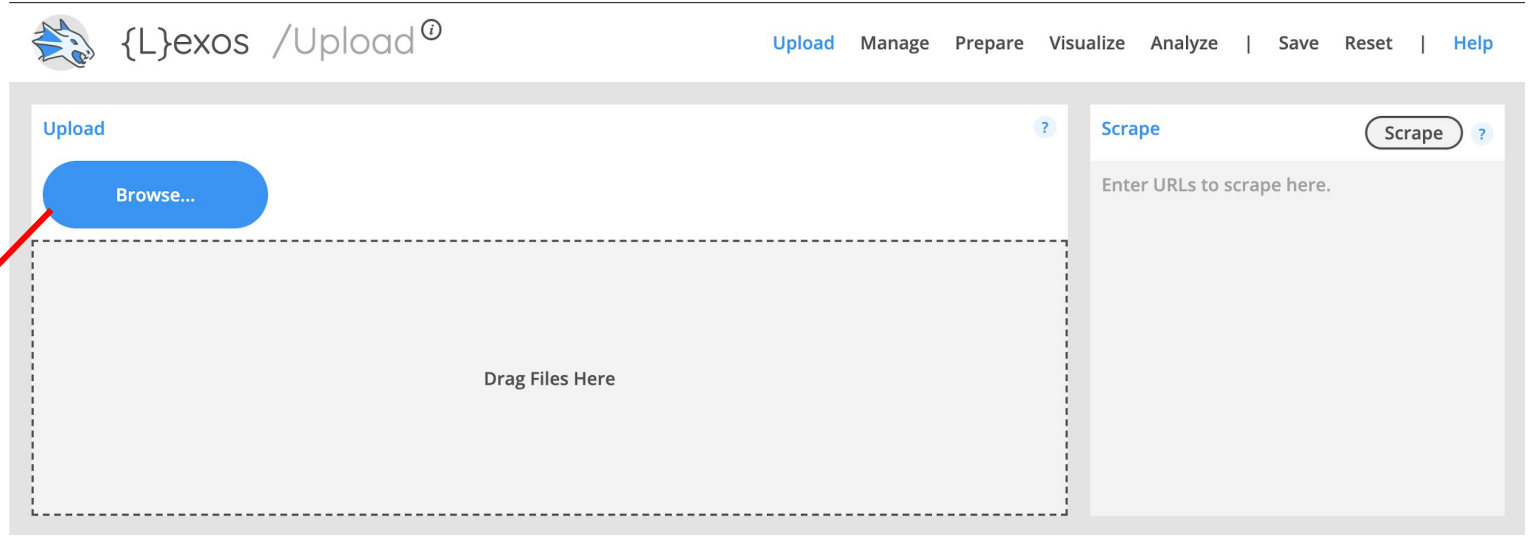
http://lexos.wheatoncollege.edu/upload

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

Lexos: Upload

Click Browse and select your entire text (or drag file into the "Drag Files Here" area)

You will not get a super visible notification when the upload is done - click "Manage" to double check that the text file is there.



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

Lexos: Manage

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

Active	#	Document	Class	Source	Excerpt
	1	endangeredspecies		endangeredspecies.txt	An endangered species is a species that is very likely to become extinct in the near future, either worldwide or in a particula... ...a to breed in landlocked tanks, raising the possibility that fish farming may be able to save the species from overfishing.[22]
	2	endangeredspeciesact1969		endangeredspeciesact1969.txt	The Endangered Species Conservation Act of 1969 (Public Law 91-135) was an expansion of the Endangered Species Preservation Act... ...eral agencies, the United States Fish and Wildlife Service(FWS) and the National Oceanic and Atmospheric Administration (NOAA).
	3	endangeredspeciesact1973		endangeredspeciesact1973.txt	The Endangered Species Act of 1973 (ESA or "The Act"; 16 U.S.C. § 1531 et seq.) is the primary law in the United States for pro... ...e property; establish a refuge, reserve, preserve, or other conservation area; or allow government access to private land.[117]
	4	internationalunionforconservationofnature		internationalunionforconservationofnature.txt	The International Union for Conservation of Nature (IUCN; officially International Union for Conservation of Nature and Natural... ...th by 2020 since an agreement between the world's nations at the Convention on Biological Diversity, held in Japan in 2010.[34]
	5	iucn_red_list		iucn_red_list.txt	The International Union for Conservation of Nature (IUCN) Red List of

Download ?

Humanities Lexos v4.0 © 2019 Wheaton Lexomics

Active Documents : 5

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

Lexos: Prepare (scrub)

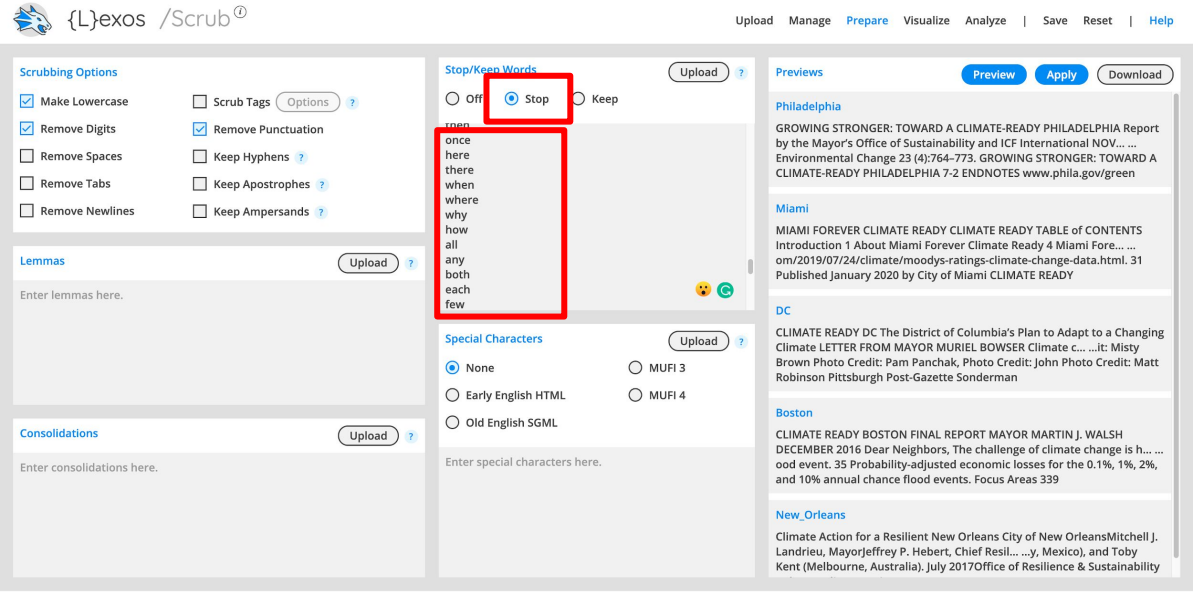
Lexos demonstrates some more advanced options you have for preparing your corpus. By "scrubbing," you are transforming the texts in your corpus and making choices that will impact your results. Here are some possibilities:

- **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.
- **Stop/Keep Words**: remove a list of words. Usually these would be **stopwords**, or the most common words in a language (English: the, a she, her, it, him, they, etc).
- **Lemmas**: standardize to the *stem* of word. For example, you can stem all forms of talk: talking, talked, talks, etc. to "talk"

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

Lexos: Removing Stopwords

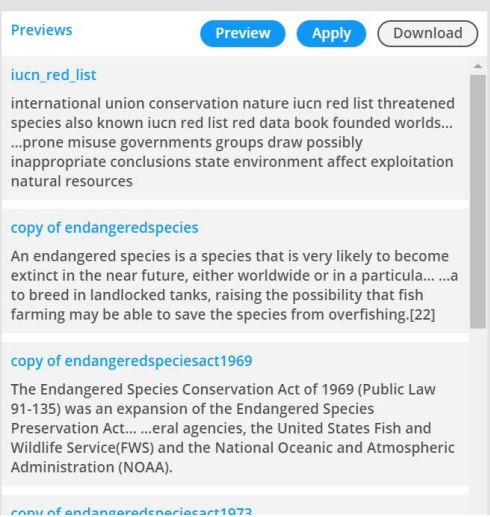
Get a list of English stopwords here: https://gist.github.com/sebleier/554280 (there is also a copy on the GitHub page). Copy and paste the stopwords (or upload the .txt file) into the “Stop/Keep Words” box then select “Stop”



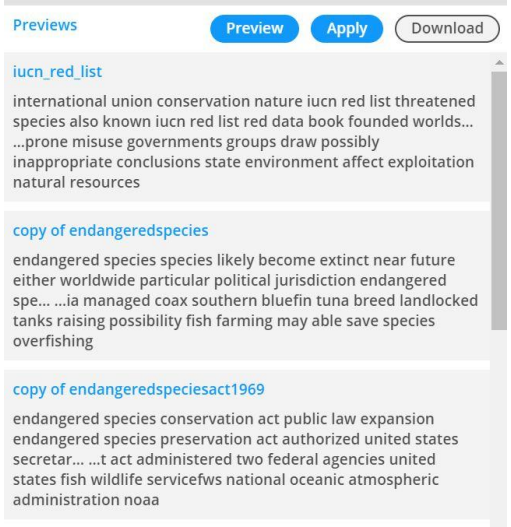
Feel free to ask questions at any point during the presentation!

Lexos: Applying your Preparations

BEFORE PREP



AFTER PREP

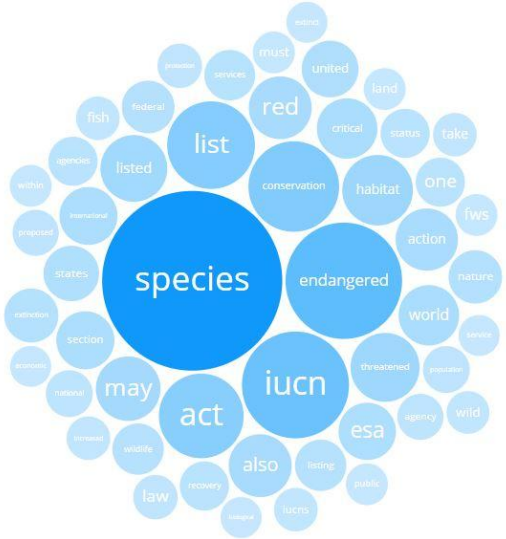


Once you have made decisions about your preparations, click "**Apply**" and wait a few minutes. Because the program is going through each document and completing all the processes you selected, it needs some time. Then, you will see the final results of your preparation! You can also **download** your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

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

Lexos: Visualize

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



Word Cloud: visualize a wordcloud across the entire text/corpus.

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

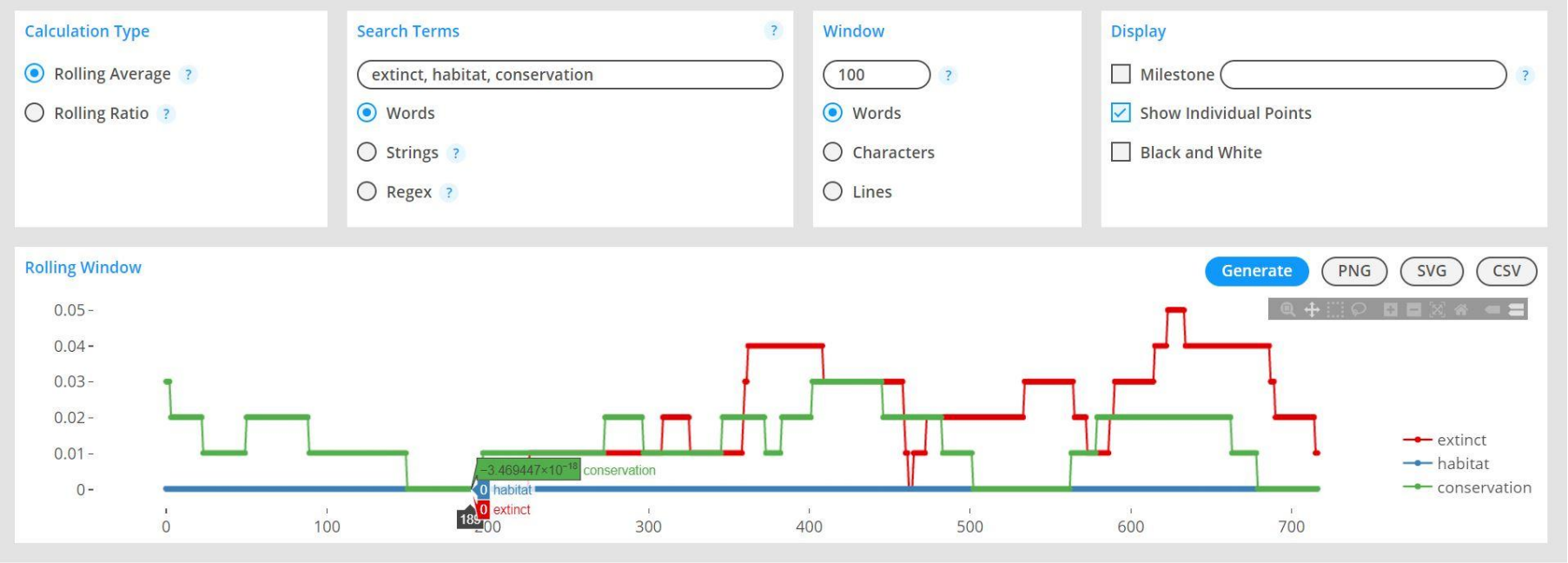
Lexos: Rolling Window

Rolling windows allow you to look at word trends across **one** document. To use a rolling window:

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”

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

Lexos: Rolling Window Results



Northeastern University
NULab for Texts, Maps, and Networks

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

Lexos: Analyze > Dendrogram

The dendrogram demonstrates similarity between the different documents. Dendrograms require at least two documents to compare. Dendrograms are able to show the hierarchy between objects. Dendrograms 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

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

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.

Conclusion

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

Your Turn!

Use the sample text or texts of your choice and begin practicing web-browser text analysis. **Explore different Lexos and Voyant features!**

Discussion Prompts

- What do you find challenging or exciting about these tools?
- What interesting or surprising results came up?
- How might you interpret those results based on what you know about your field?

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 Adam Tomasi and Vaishali Kushwaha
Delivered by Adam Tomasi and Milan Skobic
DITI Research Fellows
Digital Integration Teaching Initiative

Slides, handouts, and data available at

http://bit.ly/diti-spring2021-musselman2

You also have access to DITI Canvas Module on Computational Text Analysis.

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

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