

Computational Methods for Sociology Research: NVivo, Web Scraping, and Text Analysis

DITI Consultants: Hunter Moskowitz,
Dipa Desai
For SOCL4600
Professor Ineke Marshall
Fall 2023



Northeastern University
NULab for Texts, Maps, and Networks

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

Workshop Agenda/Objectives

- Introduction to Qualitative Coding
 - Introduction and example of NVivo
- Introduction to Web Scraping
 - Reddit Example and Ethics
- Introduction to Text Analysis
 - Introduction and Sociology Examples
- Discussion

Slides, handouts, and data available at

<https://bit.ly/fa23-Marshall-SOCL4600>



Northeastern University
NULab for Texts, Maps, and Networks

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

Introduction to Qualitative Coding and NVivo



Northeastern University
NULab for Texts, Maps, and Networks

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

What can NVivo do?

NVivo is designed for qualitative coding research materials, such as survey results, interviews, audio recording, text documents, articles, and other data formats. It also has other functions:

- create projects that store, organize, and code documents/files
- provide a method for you to code your documents with a user-created coding schema (codes)
- query, summarize, organize, and visualize information about your coding
- conduct forms of computational text analysis, like word counts, on the documents, themselves
- conduct social network analysis with social media datasets



Important Reminders

NVivo can import all types of files, including .docx, .pdf, .doc, .csv, .png, .jpeg, .txt, video/audio files, and more.

You should always **save** your original documents on your local computer or in cloud storage, even if these documents are imported into NVivo. NVivo can store documents, but it is more of an organization and analysis tool, rather than a storage tool.



NVivo Vocabulary

Full definitions available on the handout

- **Data:** your research documents & files
- **Codes:** the ways to annotate the themes/concepts in your research.
- **Nodes:** the themes/concepts that are user-created (NVivo 12 and older only)
- **Relationships:** coding connections between two data items
- **Cases:** units of analysis for your research.
- **Maps:** visualization tool to see connections between the cases and codes
- **Query:** a flexible way to explore and analyze your files, cases, and codes



Coding in practice

The screenshot displays a web-based coding or text editing interface. On the left is a dark blue sidebar with navigation options: 'IMPORT' (Data, Files, Area and Township, Interviews, Literature, News Articles, Social Media, Survey, File Classifications, Externals), 'ORGANIZE' (Coding, Cases, Notes, Sets), and 'EXPLORE' (Queries, Visualizations). The main area is a text editor with a document titled 'Barbara'. The text in the editor includes a list of names, a section titled 'Barbara' with a paragraph about clear-cutting, a section titled 'Q.5. Vision for the future of Down East' with a question, and another section titled 'Barbara' with a paragraph about environmental protection and a final paragraph about renewable resources. On the right side of the editor is a 'CODE STRIPES' sidebar showing a vertical bar with colored segments (purple, blue, orange, green) and labels: 'Coding Density', 'Natural environment', 'Economy', and 'Infrastructure'.



Querying

Querying, or asking something from your data, in NVivo provides multiple ways to explore both your codes and your texts.

- **Word Frequency:** Counts the number of times words appear in one or more files
- **Coding:** Shows the number of codes, the text that was coded, and the files
- **Crosstab:** cross-references codes and case classifications. For example, you might want to know how often a particular code appears in both scholarly articles and your primary texts.



Word Frequency Example (Windows)

“Query” can be found in the
“Explore” Tab

Alternatively, you can right
click on a file and select
Query

To query multiple items,
select the items you would
like to query in the
“Selected Items” tab and
then click “Run Query”

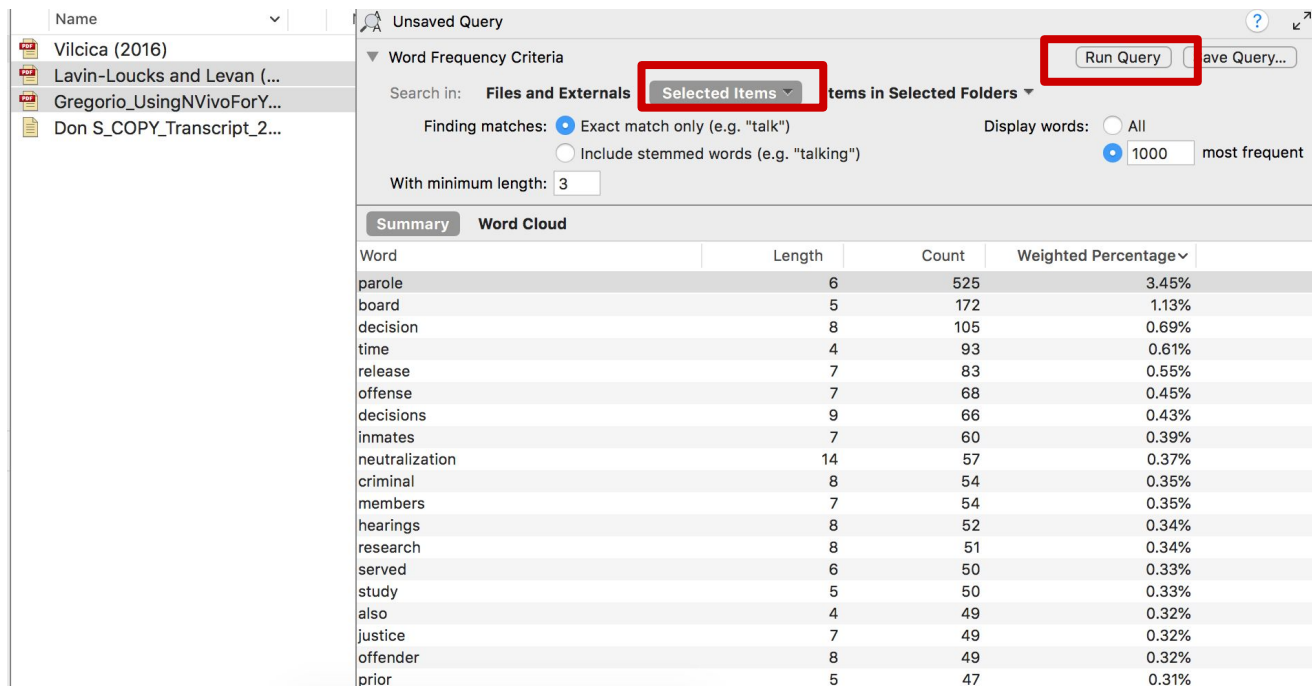
The screenshot displays the NVIVO software interface. The 'Explore' tab is active, showing a 'Word Frequency' query. The 'Files' pane on the left shows a project named 'The wellbeing project - an intr'. The 'Word Frequency Criteria' pane on the right shows the 'Run Query' button. The 'Word Frequency Query Results' pane at the bottom shows a table with columns: Word, Length, Count, and Weighted Percentage (%). The table is currently empty.

Word	Length	Count	Weighted Percentage (%)
------	--------	-------	-------------------------



Word Frequency Example (Mac)

Select the items you would like to query in the “Selected Items” tab and then click “Run Query”



Word	Length	Count	Weighted Percentage
parole	6	525	3.45%
board	5	172	1.13%
decision	8	105	0.69%
time	4	93	0.61%
release	7	83	0.55%
offense	7	68	0.45%
decisions	9	66	0.43%
inmates	7	60	0.39%
neutralization	14	57	0.37%
criminal	8	54	0.35%
members	7	54	0.35%
hearings	8	52	0.34%
research	8	51	0.34%
served	6	50	0.33%
study	5	50	0.33%
also	4	49	0.32%
justice	7	49	0.32%
offender	8	49	0.32%
prior	5	47	0.31%



A Brief Introduction to Web Scraping

slide content courtesy of Alyssa Smith
(smith.alyss@northeastern.edu)



Northeastern University
NULab for Texts, Maps, and Networks

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

Why Internet Data?

- Internet data can give us a way to (very imperfectly) quantify people's social lives online.
 - What are people talking about?
 - Who do people interact with?
 - How do communities form?
- It is especially useful at large scales
 - Getting this kind of information on how people associate without social media data would be very difficult, if not impossible!
- Internet data is very rich in terms of context, content, and usability



We Access This Data Through APIs

- An API is a way for computer programs to talk to each other.
- APIs are code wrappers, allowing a clean way to code communication with websites
- If you are trying to get a lot of information repeatedly from somebody else's computer program, an API is the way to do it!
- This might look like:
 - An analysis of all reddit posts mentioning “potato farming”
 - A program that emails you every time your advisor tweets something with negative sentiment



API Example-Reddit

- Reddit has a Python package - this means you can look up existing code to modify for your own project! It will allow you to see those “potato” posts

```
29 APP_NAME = creds['app_name']
30
31 MY_SUBREDDIT = 'wallstreetbets'
32 SEARCH_TERM = 'gamestop&(potato|facebook)'
33 reddit = praw.Reddit(client_id=REDDIT_ID, client_secret=REDDIT_SECRET, user_agent=APP_NAME)
34 subreddit = reddit.subreddit(MY_SUBREDDIT)
35
```

- Not all APIs have nice Python or R packages, though.



Web Scraping

- Sometimes websites don't even have an API; in that case, you'll have to scrape the website.
- When you scrape a website, you pull the whole webpage, parse it, and extract the data you want.
- This works better on structured websites that don't block bots (if you are scraping a website, you are a bot).
- Please be mindful of obtaining consent if you are scraping individual info.



Ethical Considerations

- **Contextual Privacy**

- When we think about privacy online we want to think of it as contextual. What someone might be comfortable saying in one context might not be something they're okay saying to a researcher.

- **Keeping People Safe**

- It is risky to publicize the username, profile picture, or exact text of a social media post or profile.
- To show example posts etc, you can make up your own or heavily redact them.



Computational Text Analysis



Northeastern University
NULab for Texts, Maps, and Networks

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

Computational Text Analysis

Text analysis is a **process to make inferences based on textual data**. 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, nGrams, 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 a corpus of text(s). Using text analysis, researchers may **find surprising results** that they would not have discovered from close reading or traditional methods alone.

From collections of texts, researchers can **discover formal continuities or discontinuities in literary genres, or textual similarities across genres**. For example, computational tools reveal textual similarities between detective fiction and science fiction over long periods of time.



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.



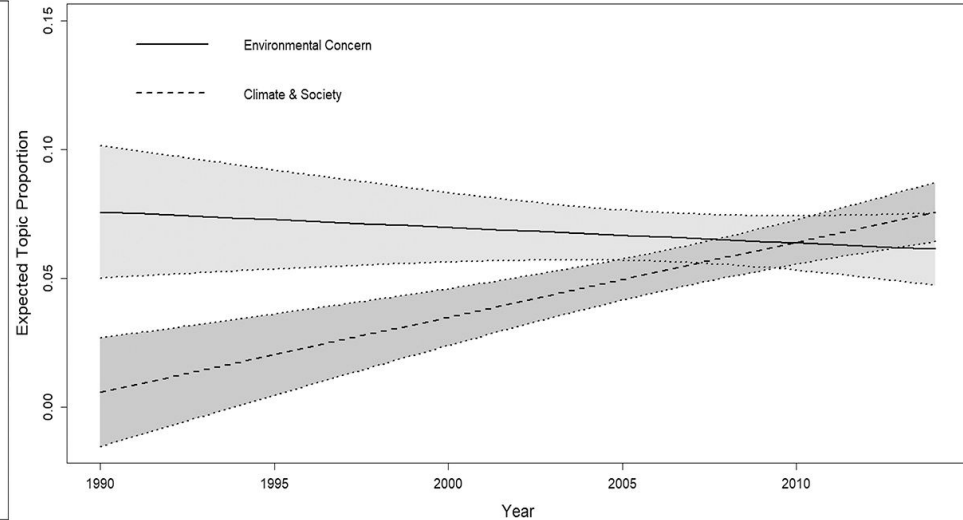
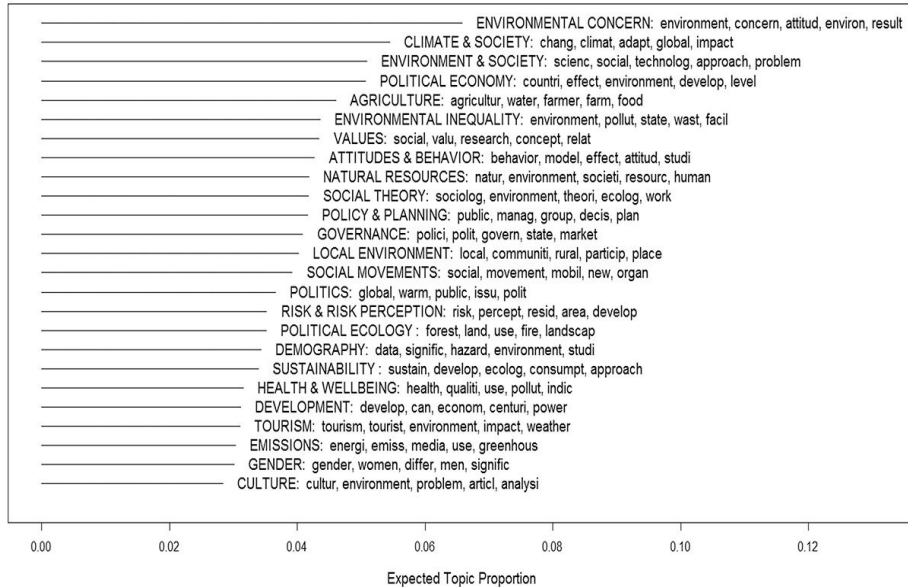
Examples from Practice



Northeastern University
NULab for Texts, Maps, and Networks

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

Key Topics in Environmental Sociology



Topical prevalence of 'Environmental Concern' and 'Climate & Society' over time, 1990 – 2014 (with 95% confidence intervals).

Jeremiah Bohr & Riley E. Dunlap (2018) Key Topics in environmental sociology, 1990–2014: results from a computational text analysis, *Environmental Sociology*, 4:2, 181-195, DOI: [10.1080/23251042.2017.1393863](https://doi.org/10.1080/23251042.2017.1393863)



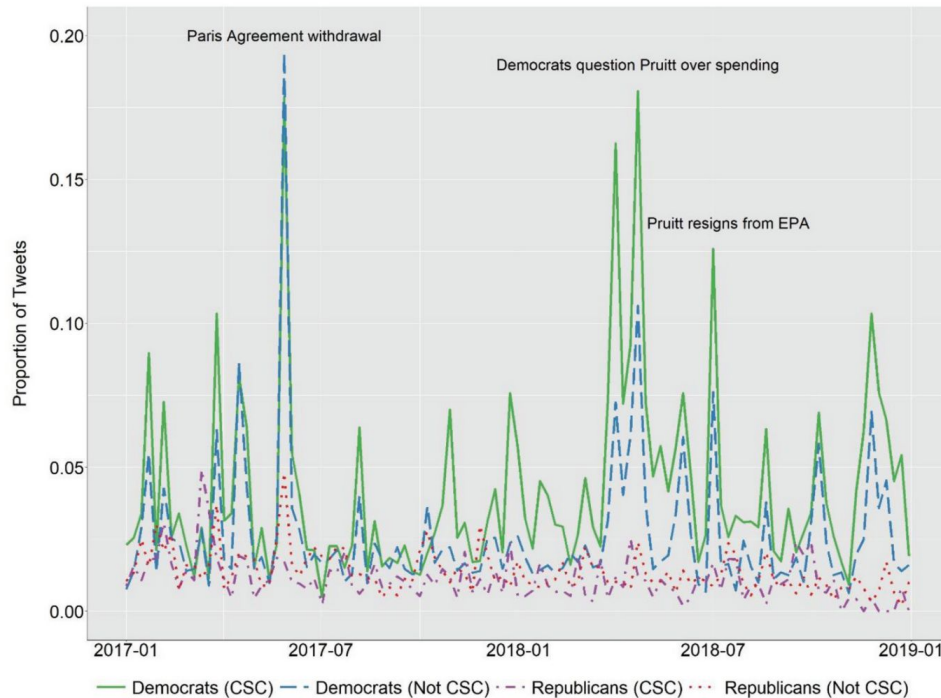
Northeastern University
NULab for Texts, Maps, and Networks

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

U.S. Environmental Politics

To what extent politicians publicly discuss environmental issues in line with public opinion and economic characteristics of their constituents?

- Nominally pro-environment Republicans representing more moderate constituents fail to oppose their partisan colleagues, particularly during the Trump administration's withdrawal from the Paris Agreement. At the same time, very few openly attacked climate science



Weekly proportions of tweets discussing environmental issues sent by the 115th House of Representatives.

[Key events and challenges: a computational text analysis of the 115th house of representatives on Twitter](#) - Jeremiah Bohr in Environmental Politics (2021), 30 (3): 399-422



Next Steps



Northeastern University
NULab for Texts, Maps, and Networks

*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.**

- This includes word counts, nGrams (n=number of words), word frequency distributions, word trends across documents, and concordances. It also makes nice visualizations!
- Voyant is a great place to start exploring texts, but it just scratches the surface!

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



Northeastern University
NULab for Texts, Maps, and Networks

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 Voyant's features!**

Sample corpus: [State of the Unions](#)

Discussion Prompts

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



Further Resources

To learn more:

- <https://bit.ly/NVivoSlides>
- <https://bit.ly/ScrapingSlides>
- <https://bit.ly/exampletextanalsisslides>

To try out:

- <https://voyant-tools.org/>
- <http://lexos.wheatoncollege.edu/upload>
- <https://www.jasondavies.com/wordtree/>



Thank you!

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

Developed by Ana Abraham
Digital Integration Teaching Initiative
Assistant Director

**Taught by Hunter Moskowitz and
Yutong Si**
Digital Integration Teaching Initiative
DITI Research Fellows

Slides, handouts, and data available at
<https://bit.ly/fa23-Marshall-SOCL4600>

We'd love your feedback! Please fill out a short survey here:
<https://bit.ly/diti-feedback>

Schedule an appointment with us! <https://bit.ly/diti-meeting>



Northeastern University
NULab for Texts, Maps, and Networks

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