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

DITI Consultants: Ana Abraham,
Juniper Johnson
For SOCL4600
Professor Ineke Marshall
Spring 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/sp23-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

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 (nodes)
- Query, summarize, organize, and visualize information about your coding
- Conduct forms of computational text analysis, like word counts, on the documents themselves

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

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.

Northeastern University
NULab for Texts, Maps, and Networks

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

NVivo Vocabulary

Full definitions available on the handout

- **Data**: your research documents & files
- **Codes**: the method to annotate the themes/concepts
- **Nodes**: the actual themes/concepts that you create
- **Relationships**: coding connections between two data items

- **Cases:** units of analysis for your research.
- **Maps**: visualization tool to see connections between the cases and nodes
- **Query**: a flexible way to explore and analyze your files, cases, and nodes

Northeastern University
NULab for Texts, Maps, and Networks

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

Coding in practice

key, at that point.

M: Sure sure, so I did not understand, what was the comparison with other five, with other locals, did they have more issues or less, or did they also have that similar experiences like you, across all those different locals [26]

R1: They had, you know, similar experiences that I did. Very strong, present, you know, walking. I didnt see hear, early on, anyways for the initial, when we pulled them that day, the Thursday, that anyone was saying "i'm not going". Didnt hear that from anybody. Didnt hear it from my local, didnt hear it, you know because all the locals have facebook pages too. You know, they have, social pages, so you can also keep in touch that way, and, you know, between reading and [unclear], you know taliking to [unclear] and stuff like that, we were hearing nothing but positive feedback, on on, people walking. People were ready. They were angry, with the contract that was being presented, and they were ready, they were ready to take it, to go to the sidewalk [27] so to speak. And then, and to state their point, yep.

M: Can I maybe ask you yes, related to that anger, like, how did people, hm how to say, how did they precisely rationalize their anger, or, how did they justify their anger. What did they say, who, or how did they understand, why that unjustice was being done to them. How did they understand why the managers at all decided to offer them such a bad contract, how did they think about that anger?

R1: Sure. So, with that, when we, when we, with the organizer, we met earlier, weeks earlier to, so, we met up in [unclear], the organizer brought as many - you know, every, its open invitation, to any local member of the store, to come to this, rather large gathering of people [28] where he broke down the contract. So, not only if you couldn't make, these contract breakdowns came back to the store, and were handed out to, individuals to read, and were posted on the union board. So, the people had time to see, what the contract, you know, here

Northeastern University
NULab for Texts, Maps, and Networks

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

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 nodes and case classifications. For example, you might want to know how often a particular node appears in both scholarly articles and your primary texts.

Northeastern University
NULab for Texts, Maps, and Networks

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

Word Frequency Example (Windows)

"Query" can be found in the "Explore" Tab
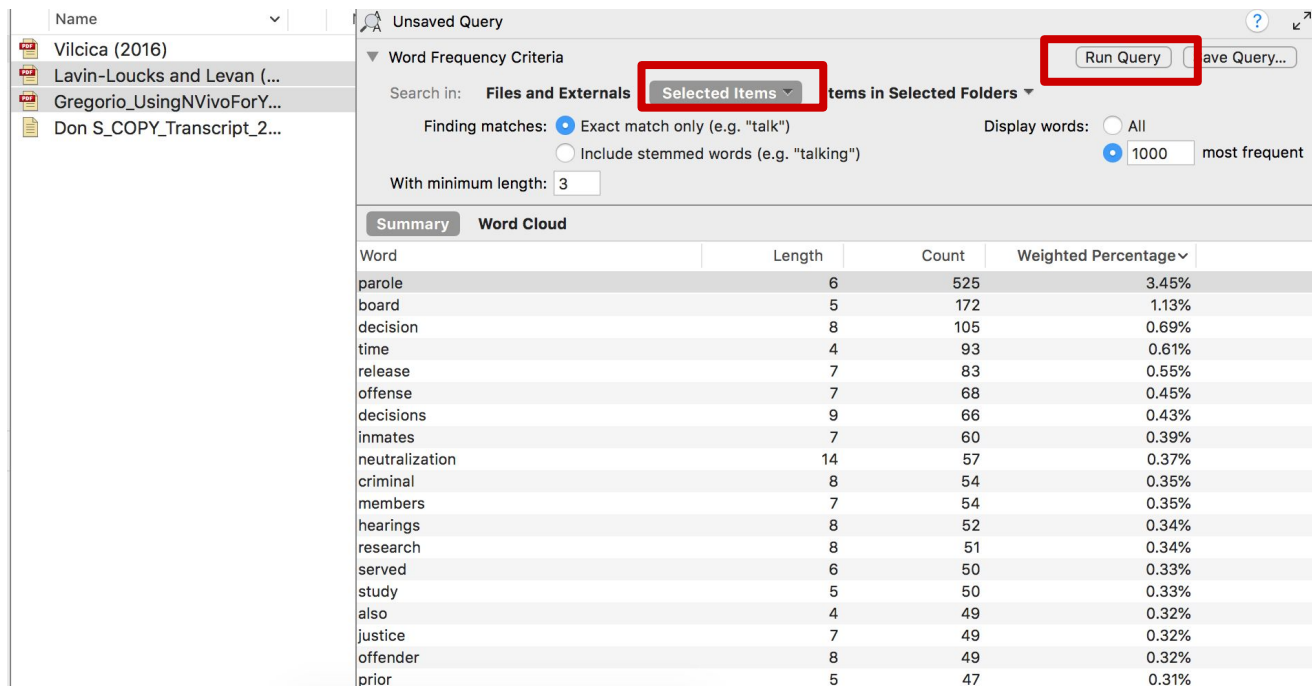
Alternatively, you can Command/Control+click on a file and select "Query"

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

Northeastern University
NULab for Texts, Maps, and Networks

Word Frequency Example (Mac)

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

Name		
Vilcica (2016)		
Lavin-Loucks and Levan (...		
Gregorio_UsingNVivoForY...		
Don S_COPY_Transcript_2...		

Unsaved Query

▼ Word Frequency Criteria

Run Query Save Query...

Search in: **Files and Externals** **Selected Items** ▼ Items in Selected Folders ▼

Finding matches: ● Exact match only (e.g. "talk")
○ Include stemmed words (e.g. "talking")

Display words: ○ All
● 1000 most frequent

With minimum length: 3

Summary **Word Cloud**

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%

Northeastern University
NULab for Texts, Maps, and Networks

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

A Brief Introduction to Web Scraping

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

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

Northeastern University
NULab for Texts, Maps, and Networks

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

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

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

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.

Northeastern University
NULab for Texts, Maps, and Networks

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.

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

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.

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

Computational Text Analysis

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

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, 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.

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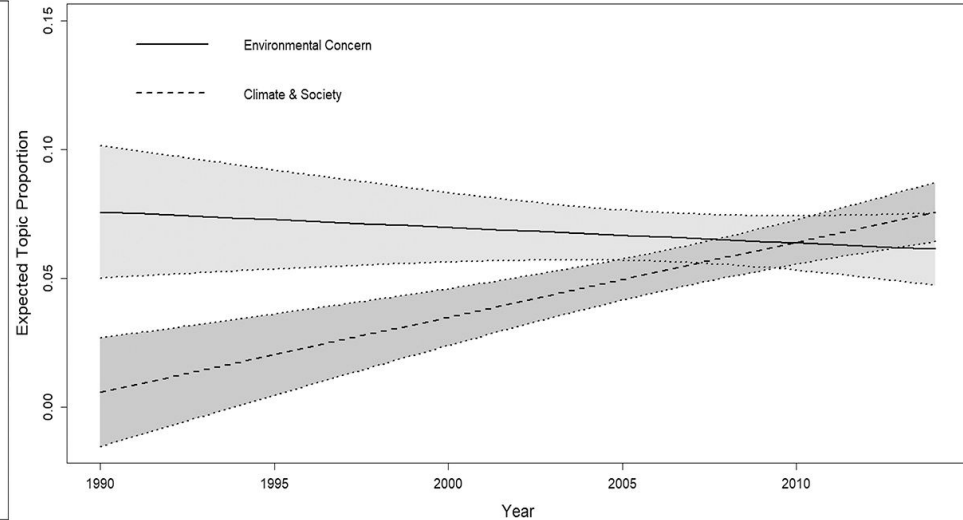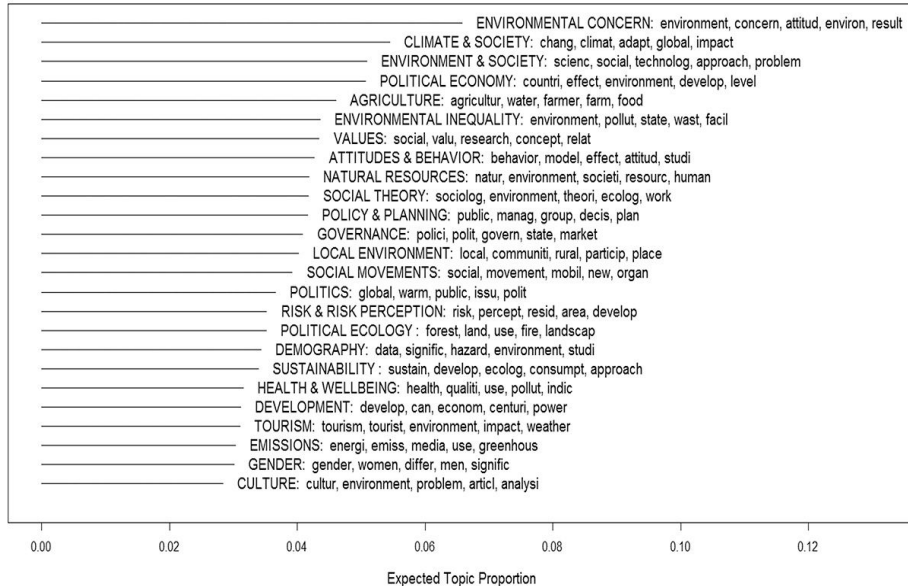
Feel free to ask questions at any point during the presentation!

Examples from Practice

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

Key Topics in Environmental Sociology

25 topics ranked from most to least prevalent in the corpus of 815 environmental sociology articles, including the top five associated word stems. The *x*-axis represents the proportion of each topic within the overall corpus.

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
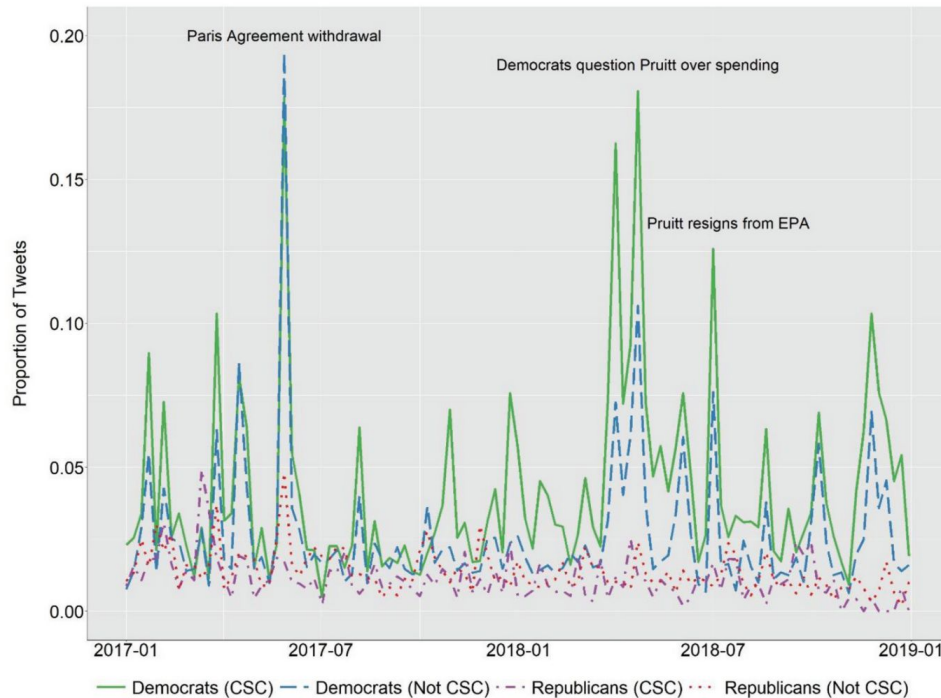
Northeastern University
NULab for Texts, Maps, and Networks

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

Northeastern University
NULab for Texts, Maps, and Networks

Next Steps

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/

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!**

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?

Northeastern University
NULab for Texts, Maps, and Networks

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

Further Resources

To learn more:

- https://bit.ly/NVivoSlides
- https://bit.ly/ScrapingSlides
- https://bit.ly/exampletextanalysisslides

To try out:

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

Northeastern University
NULab for Texts, Maps, and Networks

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

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 Ana Abraham and Juniper Johnson
Digital Integration Teaching Initiative
DITI Research Fellows

Slides, handouts, and data available at

https://bit.ly/sp23-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://calendly.com/diti-nu**

Northeastern University
NULab for Texts, Maps, and Networks

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