May 6-7, 2024 | 9:30 am - 4:30 pm

In-Person Workshop

Computational Text Analyses Bootcamp

u.mcmaster.ca/scds-events

Subhanya Sivajothy
Devon Mordell
Jay Brodeur

DMDS SCDS McMaster University Library

McMaster University is located on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the “Dish With One Spoon” wampum agreement.

Laslovarga, “Webster Falls in Winter, Waterdown, Hamilton, Ontario, Canada - Spencer Gorge / Webster's Falls Conservation Area,” 23 January 2011, Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Waterdawn_Webster_Falls_in_Winter8.jpg

Outline & Schedule – Day 2

Segment	Time Allotted	Key Topics / Activities
Corpora Selection	30 minutes	Sources and types Key considerations for different source materials and analyses case studies
Visualization for Dissemination	75 minutes	Core concepts Visualization types hands-on exercises
Working Period	75 minutes	Work on your own data or a pre-selected project
Lunch (1230 - 1330)	60 minutes	**Lunch**
Working Period	120 minutes	Continue project work
Share Back, Closing Comments	30 minutes	Share your work Questions and wrap-up

Workshop landing page: scds.github.io/dmds23-24/comptext-bootcamp.html

Building your corpus

What is the "stuff" of your analysis?

- Do you have a specific research question in mind?
- Or do you have a corpus you want to explore for insights?

- Where will you get the data?
- What pre-processing steps may be required?

- What documents must be included in your corpus?
- Are there any documents that should be omitted from your corpus?

...

Working with open corpora

- Project Gutenberg
  - 70,000 eBooks in the public domain (no restrictions)
  - Largely free of errors, limited pre-processing required
- Internet Archive
  - Not necessarily in the public domain (Fair Dealing exception for research / private study)
  - Many text files are created through OCR of scanned documents (error prone)
- HathiTrust
  - Similar to Internet Archive: not necessarily PD and text files may be OCR generated

Some text prep and analysis considerations

Text preparation and analysis are task specific

Your approaches should be informed by:

1. **Your analysis objectives**
2. **Your source materials and their common traits, inconsistencies, errors**
3. **Your abilities, time, interests, and familiarity with tools**

Considerations

1. Your analysis objectives

- Do you have a defined research question or are you experimenting?
- What analyses are required to meet your objectives and create desired outputs?
- Are your methods sensitive to particular types of errors and imprecision?
- For which applications were the methods developed? How were they trained/validated? Are they appropriate for your purposes?

Considerations

2. Your source materials and their common traits, inconsistencies, errors

- Born-digital vs. digitized
- The quality of the source materials
- The methods used to digitize materials and create text
- The structure of the materials and the text within
- The nature of communication within the materials
- Which (if any) processing operations can be automated?

Considerations

3. Your abilities, time, interests, and familiarity with tools

- With which tools are you familiar? Do feasible solutions exist within those?
- How much time and interest do you have to learn new approaches and tools?
- Do you have time to explore, test, and iterate?
- Can you apply your acquired knowledge & workflows to future projects?

Case study: Working with Social Media Data

1192318107173285888	RT @MikeSchreiner: Bill 132 is another attack on environmental protection. Mo I'll keep saying it: Protecting drinking water and holding polluters accountable i https://t.co/garmgfMhA8 #onpoli	Thu Nov 07 05:49:23 +0000 2019	07/11/2019 05:49:23			
1192317960058232832	RT @StephenPunwasi: Sure, but it's the Conservatives that are destroying the #onpoli #cdnpoli https://t.co/lkobmzv9C5	Thu Nov 07 05:48:48 +0000 2019	07/11/2019 05:48:48			
1192317634970312704	RT @drpmonaghan: My thoughts on the OAP Advisory Panel Report. The recommendations are good. Really good. @ToddSmithPC - It's time to get to work @sletersky @JillDunlop1 @AmyFeePC @JR_Ottawa #NeedsBasedTherapy #onpoli https://t.co/hdhLCUTJCX	Thu Nov 07 05:47:31 +0000 2019	07/11/2019 05:47:31			
1192317567978917888	RT @coteau: Doug Ford's Fall Econ Statement is all smoke & mirrors to at	Thu Nov 07 05:47:15 +0000 2019	07/11/2019 05:47:15			
1192317383429496834	RT @StephenPunwasi: "The Ford government has proposed eliminating an exi Fines for dumping toxins in the drinking water is "red tape" to Ontario? 🤔	Thu Nov 07 05:46:31 +0000 2019	07/11/2019 05:46:31			

Case study: Working with Historical Newspapers

Discussion

Let’s talk about your projects and / or experiences with analyzing texts...

- What use cases have you encountered?
- What kinds of documents do you plan on working with?
- What issues do you foresee with your corpus?

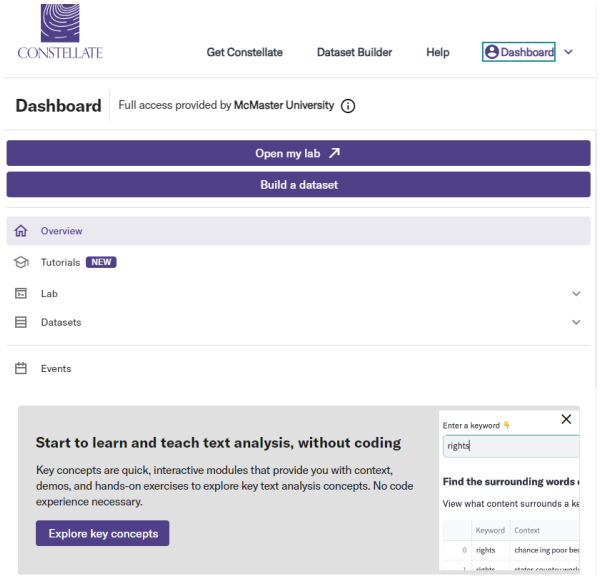
A brief Constellate demo

What is Constellate?

Constellate is a text analysis learning and analysis platform supported by JSTOR Labs and ITHAKA.

It combines:

- A comprehensive set of interactive Jupyter Notebook-based tutorials for text analysis.
- Analytical access to content from 35+ million articles, books, and newspapers from JSTOR, Portico, Chronicling America, etc.
- A computational platform to develop notebooks and collect, create, analyze, and store data.*
- Access to advanced support*



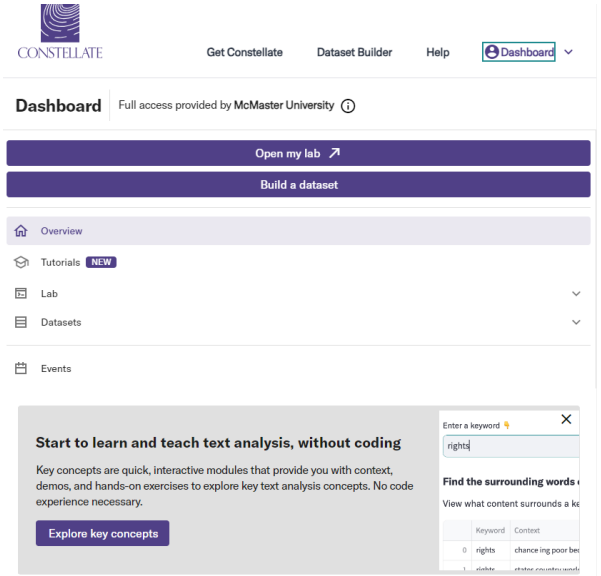
https://constellate.org/

* Available to members of pedagogy package-subscribing institutions like McMaster

Getting started with Constellate

To access the features of the pedagogy package (McMaster members):

1. Sign up for an account using your McMaster email.
    - If on campus, navigate to https://constellate.org/register
    - If off-campus, log in via the Library's off-campus access service: https://u.mcmaster.ca/constellate-signup
2. Follow the instructions to verify your account
3. Log in via https://constellate.org/login
    - If you are off campus and aren't recognized as a McMaster member, log out and back in via off-campus access: https://u.mcmaster.ca/constellate-login



https://constellate.org/

Visualization for Dissemination

Working Period 1
75 minutes

Choose Your Own Adventure...

Practice is the best teacher! Now that you have a robust toolkit of techniques at your disposal, you can deepen your learning through practice.

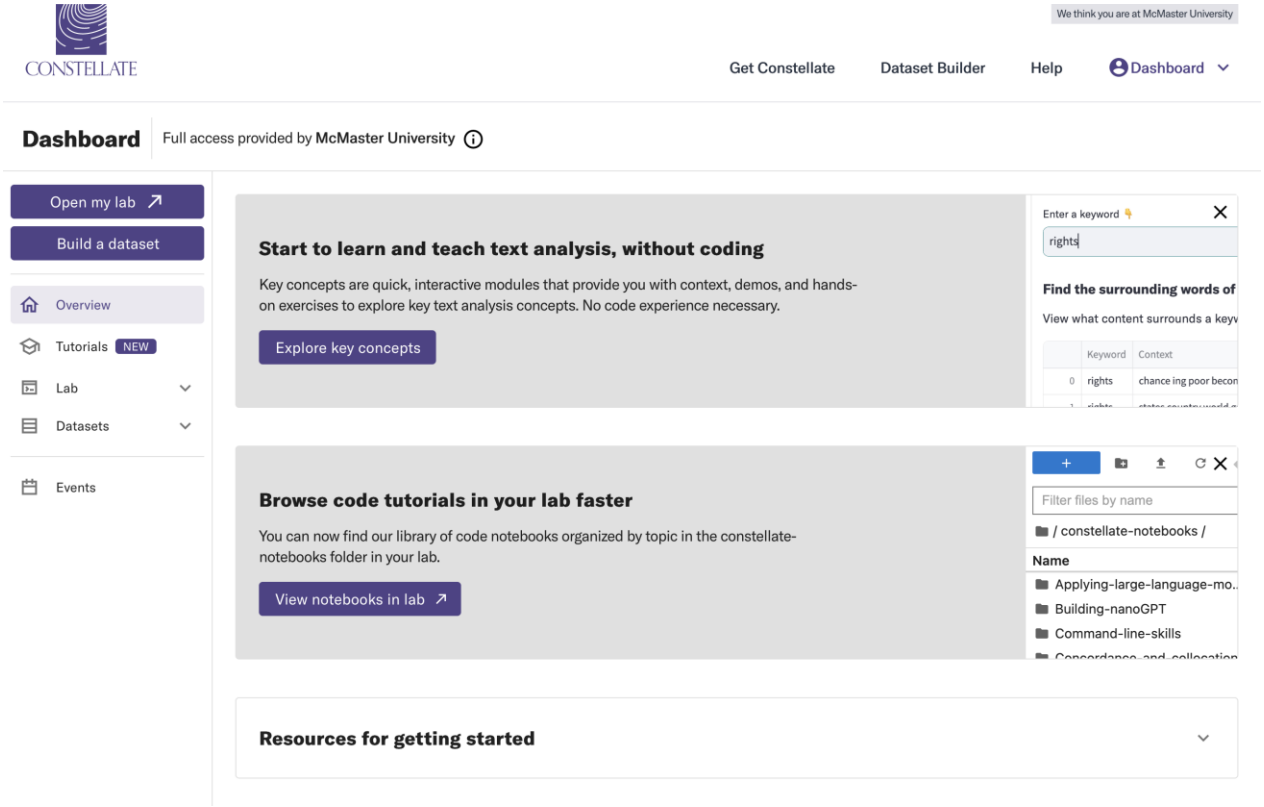
If you have a project in mind...

- Use Voyant to determine what pre-processing tasks need to be done
- Try out the techniques we learned on your own corpus
- Consult with us to figure out next steps

If you do not yet have a project in mind...

- Explore the repositories discussed to identify a possible corpus
- Work through a Constellate tutorial to learn another skill

Introduction to Constellate



Working Period 2
120 minutes

Share back &
Closing comments

Some final thoughts

- Begin with your goals in mind
- Experiment and iterate
- Understand your methods
- Start small and scale up
- Document your sources, methods, rationale, and outcomes **as you develop them**