November 6, 2023 | 1:30-4pm

Virtual Workshop

Getting Started with Computational Text Analyses

u.mcmaster.ca/scds-events

Devon Mordell

Jay Brodeur

DMDS SCDS Library McMaster University

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

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.

***Closed Captioning** is available for this session—please click the "CC" button at the bottom of the Zoom screen to turn these on. Please feel free to reach out to scds@mcmaster.ca with any other access requests.*

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: https://scds.ca/events/code-of-conduct/

Certificate Program

The Sherman Centre offers a Certificate of Completion that rewards synchronous participation in 7 workshops. We also offer concentrations in Data Analysis and Visualization, Digital Scholarship, and Research Data Management.
Learn more about the Certificate Program: https://scds.ca/certificate-program

Attendance Confirmation

If you would like to be considered for a certificate, verify your participation in today's workshop by completing the form at: https://u.mcmaster.ca/verification

An organizer will enter the code into the session chat window.

Outline & Schedule

Segment	Time Allotted	Key Topics / Activities
Introductory remarks	20 minutes	Introduction to text preparation and analysis Overview of concepts and methods Key considerations for different source materials and analyses
Named Entity Recognition	35 minutes	Introduction to Google Colab & Jupyter Notebooks Get the data Introduction to NER and hands-on exercise
--Break--	**10 minutes**	**--Break--**
Sentiment Analysis	30 minutes	Introduction and hands-on exercise Constellate demonstration
Topic Modeling	30 minutes	Introduction and hands-on exercise
Q & A; Final thoughts (lecture + discussion)	20 minutes	Questions & final thoughts Where to learn more

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

Learning Objectives

By the end of this module, you will be able to:

- List the common methodological approaches used in text preparation and analysis and identify when and how to use them based on source materials and analysis objectives.
- Apply prepared computational techniques to perform common text preparation steps and introductory analyses.
- Identify resources and tutorials for further learning and analyses.

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

An Introduction

Natural Language Processing
(is a big family)

Text and speech recognition / processing

OCR, speech recognition, text-to-speech

Morphological analysis

Stemming, Lemmatization, Part of Speech Tagging

Syntactic analysis

Parsing, Sentence breaking

Lexical semantics

Named entity recognition, Sentiment analysis, word sense disambiguation

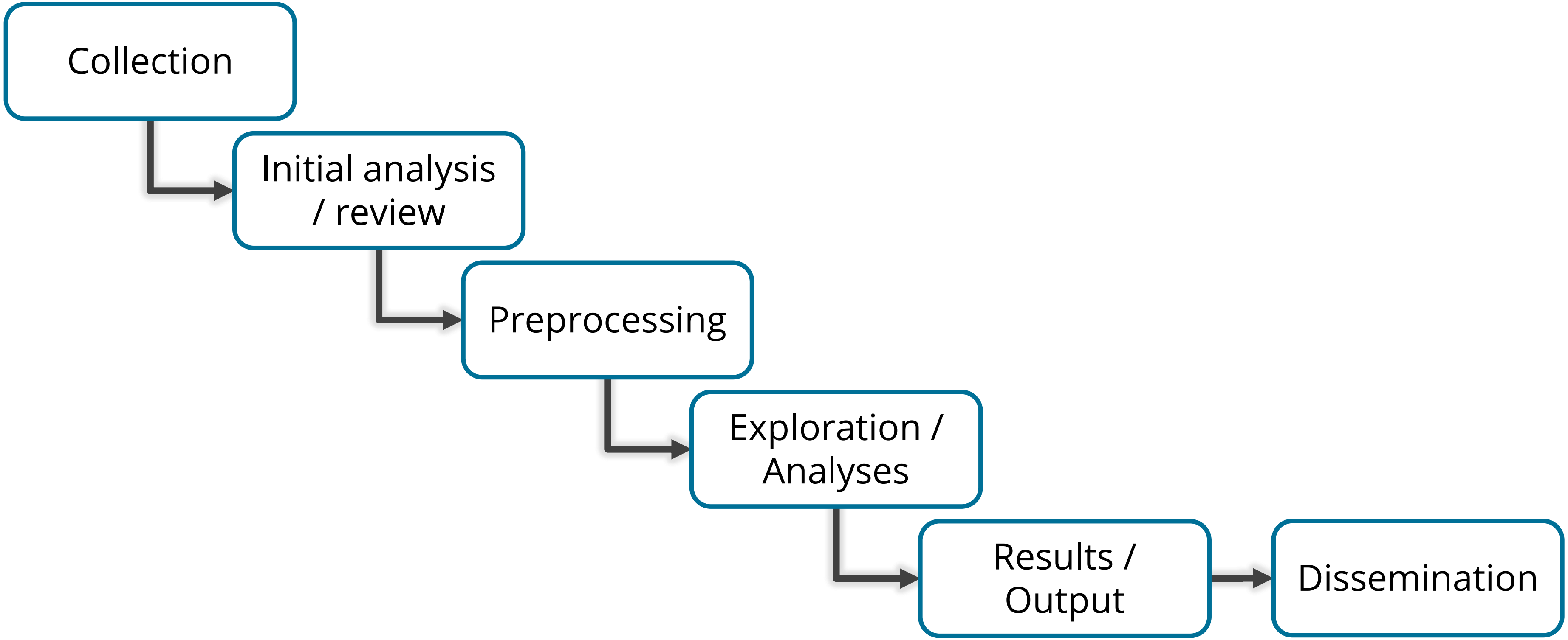
Relational semantics

Relationship extraction, Semantic parsing

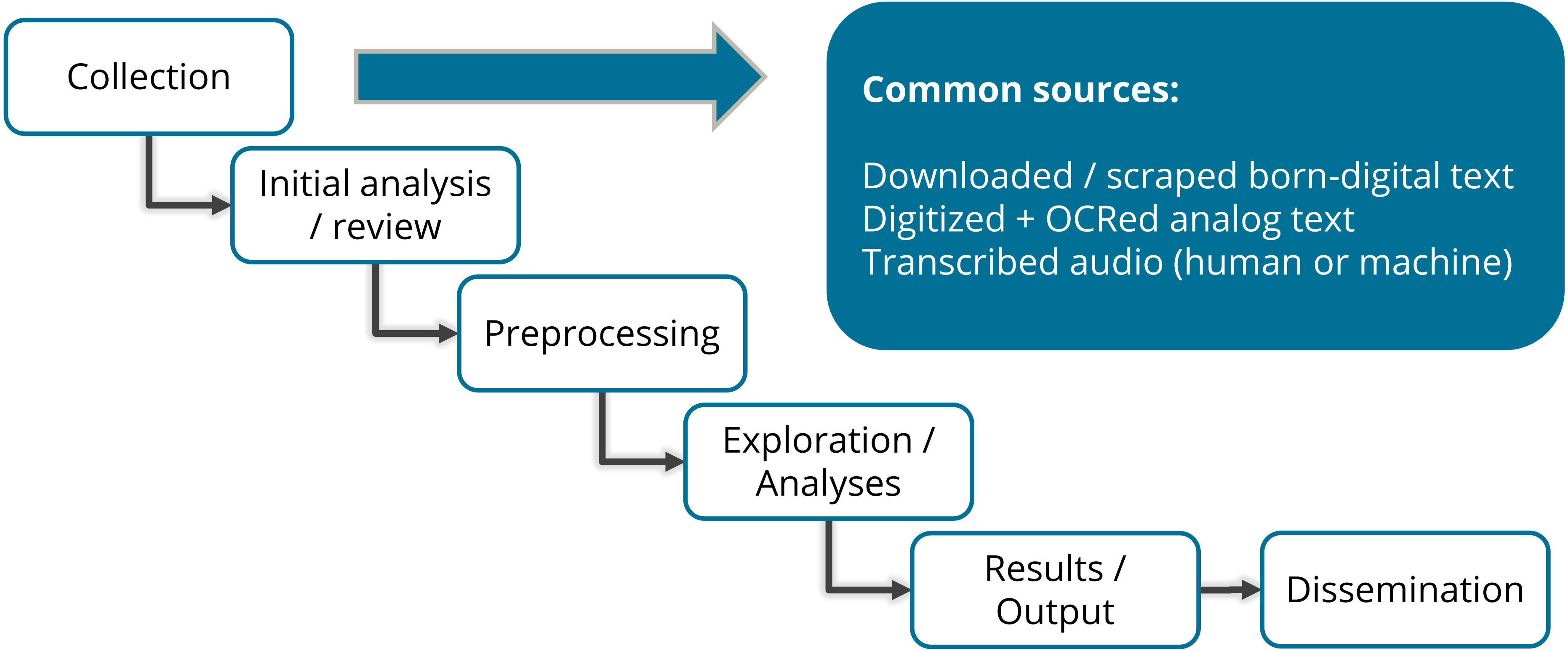
Discourse semantics

Discourse analysis, Topic segmentation, Argument mining

NLP Workflows



NLP Workflows



Collection

Initial analysis / review

Preprocessing

Exploration / Analyses

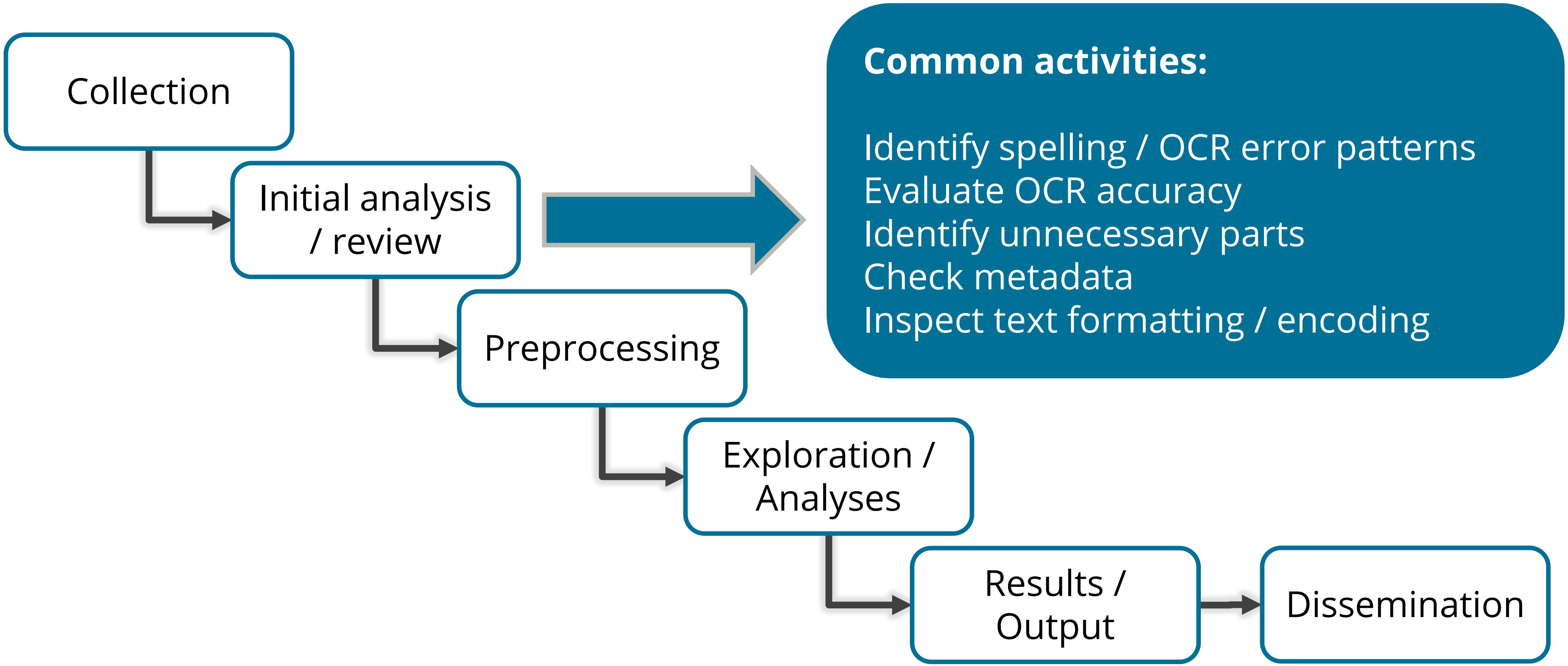
Results / Output

Dissemination

Common sources:

Downloaded / scraped born-digital text
Digitized + OCRed analog text
Transcribed audio (human or machine)

NLP Workflows



Collection

Initial analysis / review

Preprocessing

Exploration / Analyses

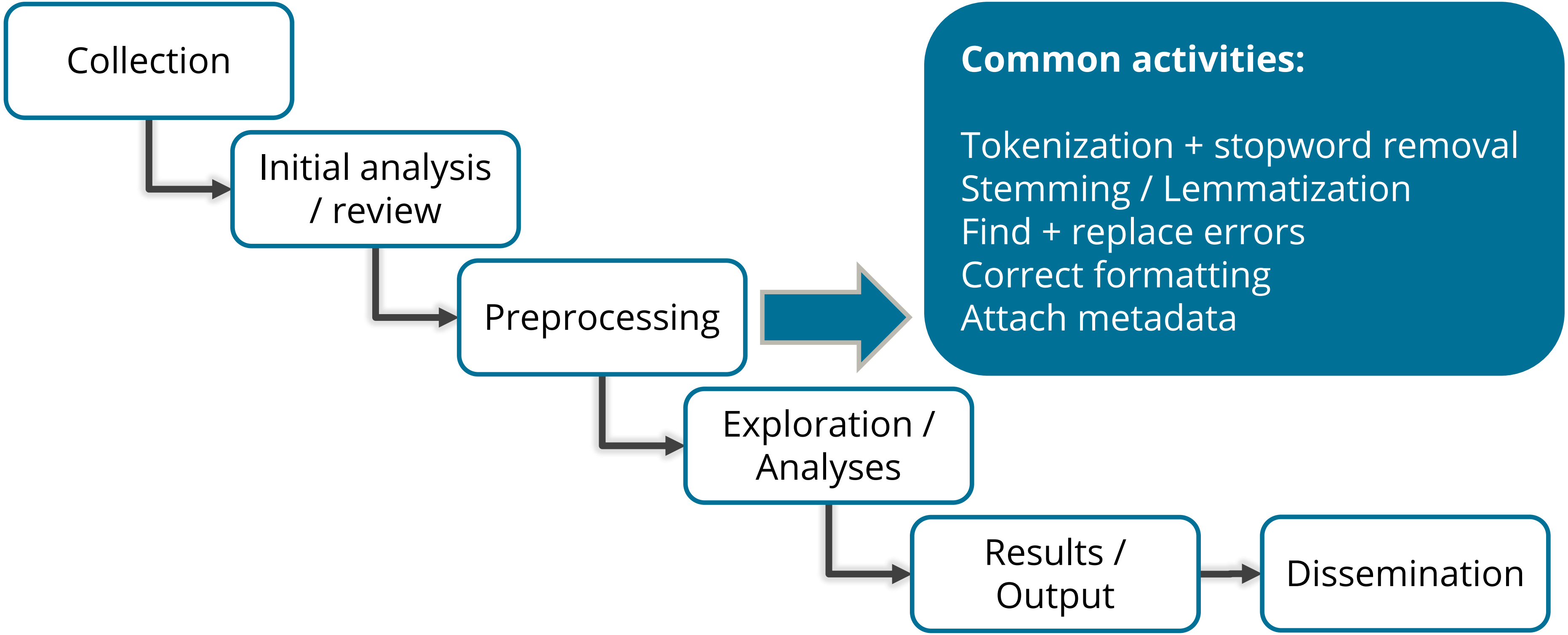
Results / Output

Dissemination

Common activities:

Identify spelling / OCR error patterns
Evaluate OCR accuracy
Identify unnecessary parts
Check metadata
Inspect text formatting / encoding

NLP Workflows

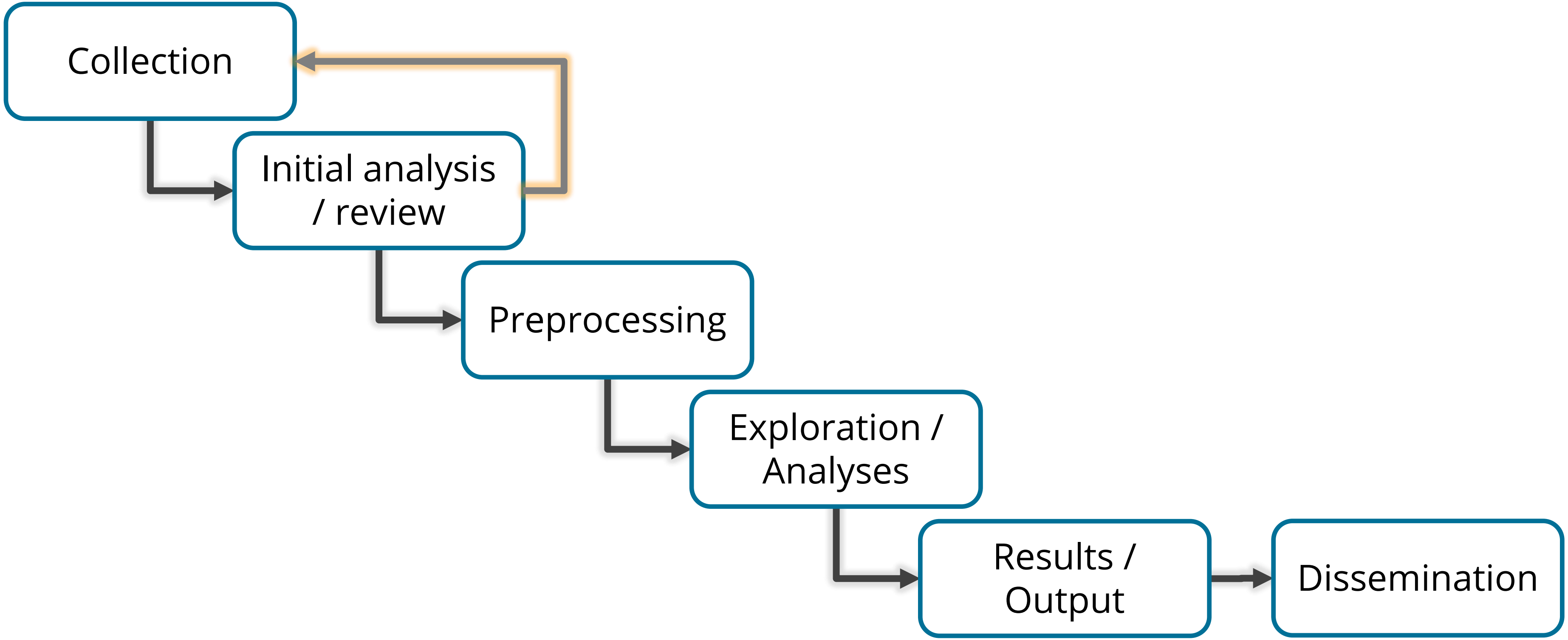


Common activities:

Tokenization + stopword removal
Stemming / Lemmatization
Find + replace errors
Correct formatting
Attach metadata

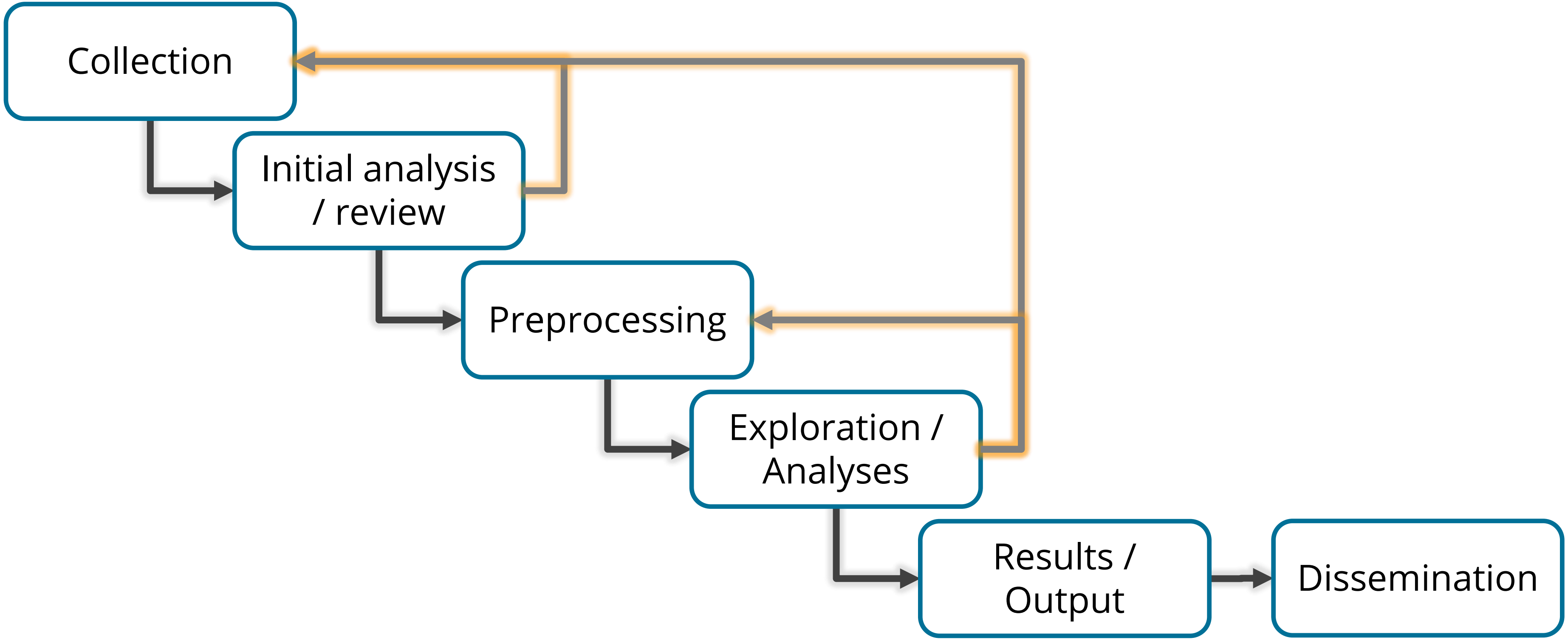
NLP Workflows

… are iterative



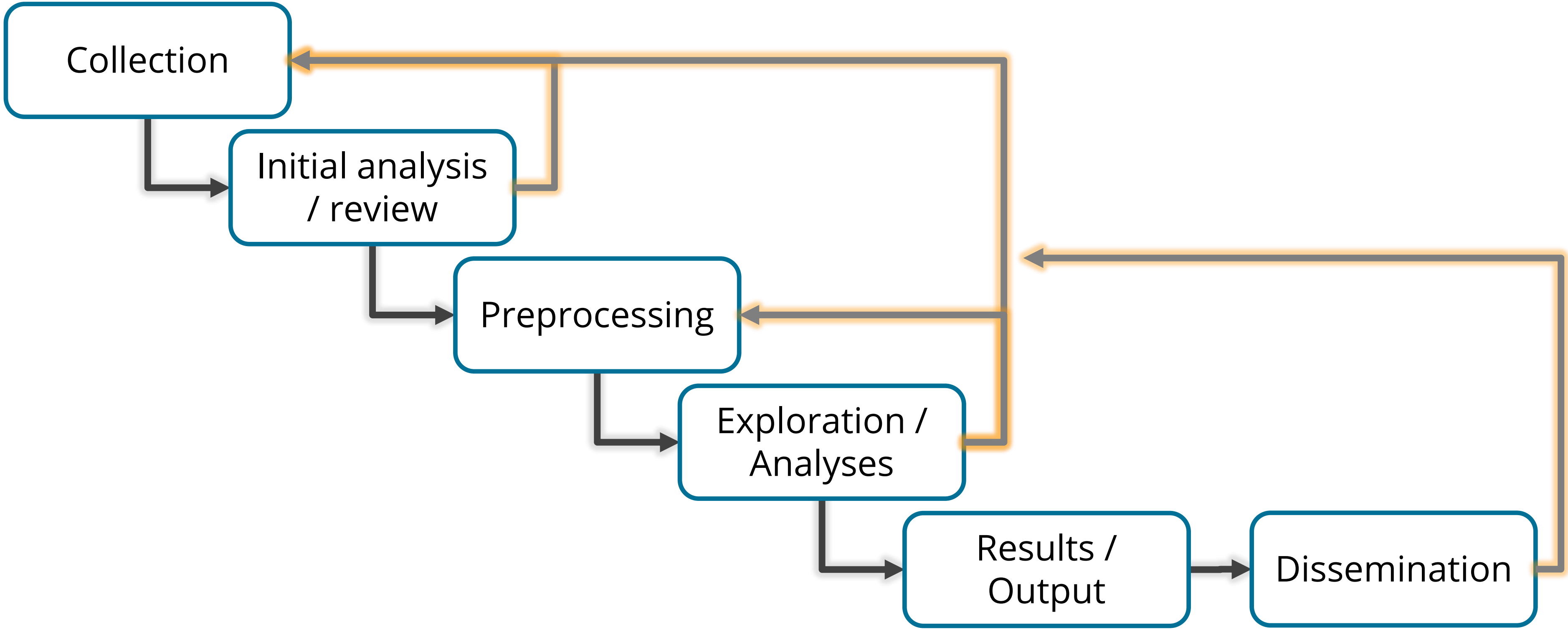
NLP Workflows

… are iterative



NLP Workflows

… are iterative



Collection

Initial analysis / review

Preprocessing

Exploration / Analyses

Results / Output

Dissemination

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

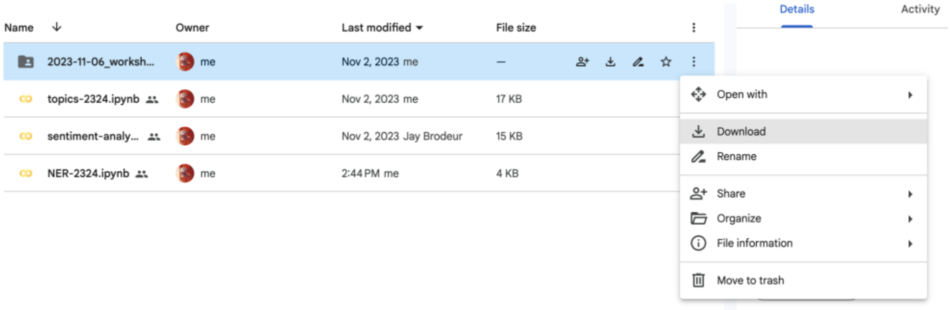
- With which tools are you familiar? Do feasible solutions exist within those?
- How much time and interest do you have to learn new approaches and tools?
- Do you have time to explore, test, and iterate?
- Can you apply your acquired knowledge & workflows to future projects?

Getting started:

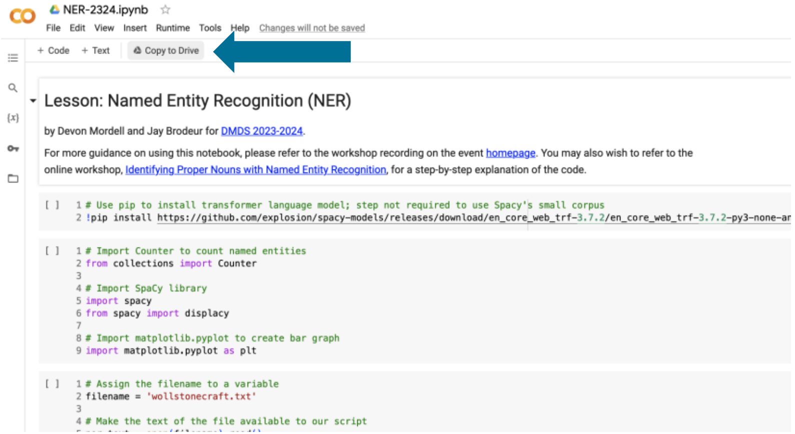
Google Colab & Jupyter notebooks
Get your data

Go to u.mcmaster.ca/dmds-text-2324 to download your data and view the workshops for today’s workshop.
Follow along with Devon’s instructions

Orientation to Google Colab / Jupyter Notebooks



Name	Owner	Last modified	File size
2023-11-06_worksh...	me	Nov 2, 2023 me	—
topics-2324.ipynb	me	Nov 2, 2023 me	17 KB
sentiment-analy...	me	Nov 2, 2023 Jay Brodeur	15 KB
NER-2324.ipynb	me	2:44 PM me	4 KB



NER-2324.ipynb
File Edit View Insert Runtime Tools Help Changes will not be saved
+ Code + Text
Copy to Drive

Lesson: Named Entity Recognition (NER)

by Devon Mordell and Jay Brodeur for DMDS 2023-2024.

For more guidance on using this notebook, please refer to the workshop recording on the event homepage. You may also wish to refer to the online workshop, Identifying Proper Nouns with Named Entity Recognition, for a step-by-step explanation of the code.

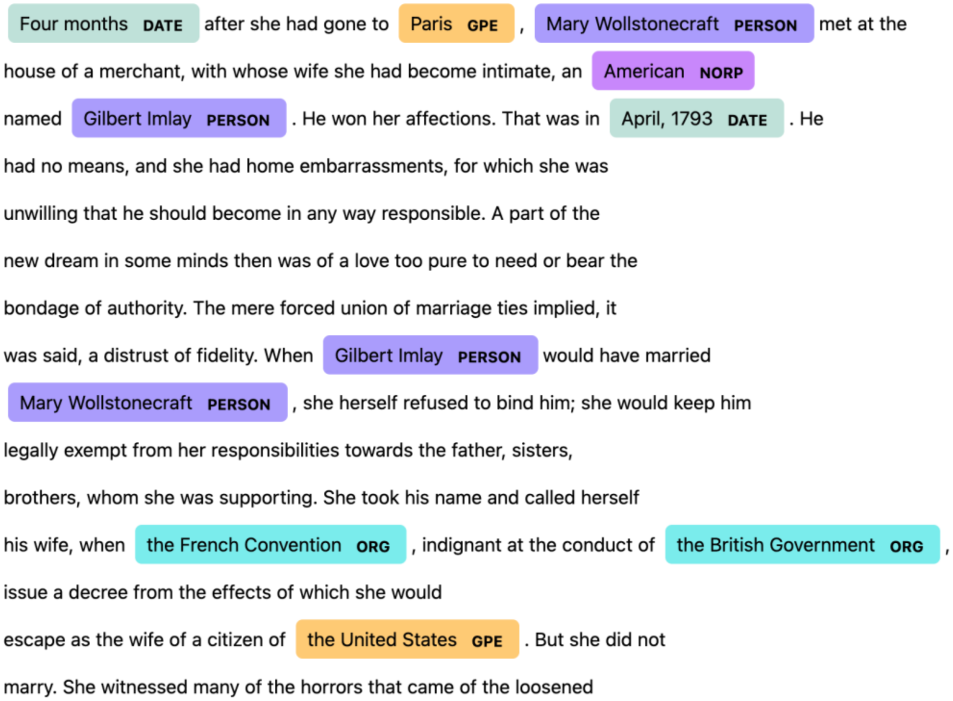
```
[ ]  # Use pip to install transformer language model; step not required to use Spacy's small corpus
     !pip install https://github.com/explosion/spacy-models/releases/download/en_core_web_trf-3.7.2/en_core_web_trf-3.7.2-py3-none-ar
```

```
[ ]  # Import Counter to count named entities
     from collections import Counter

     # Import SpaCy library
     import spacy
     from spacy import displacy

     # Import matplotlib.pyplot to create bar graph
     import matplotlib.pyplot as plt
```

```
[ ]  # Assign the filename to a variable
     filename = 'wollstonecraft.txt'

     # Make the text of the file available to our script
```

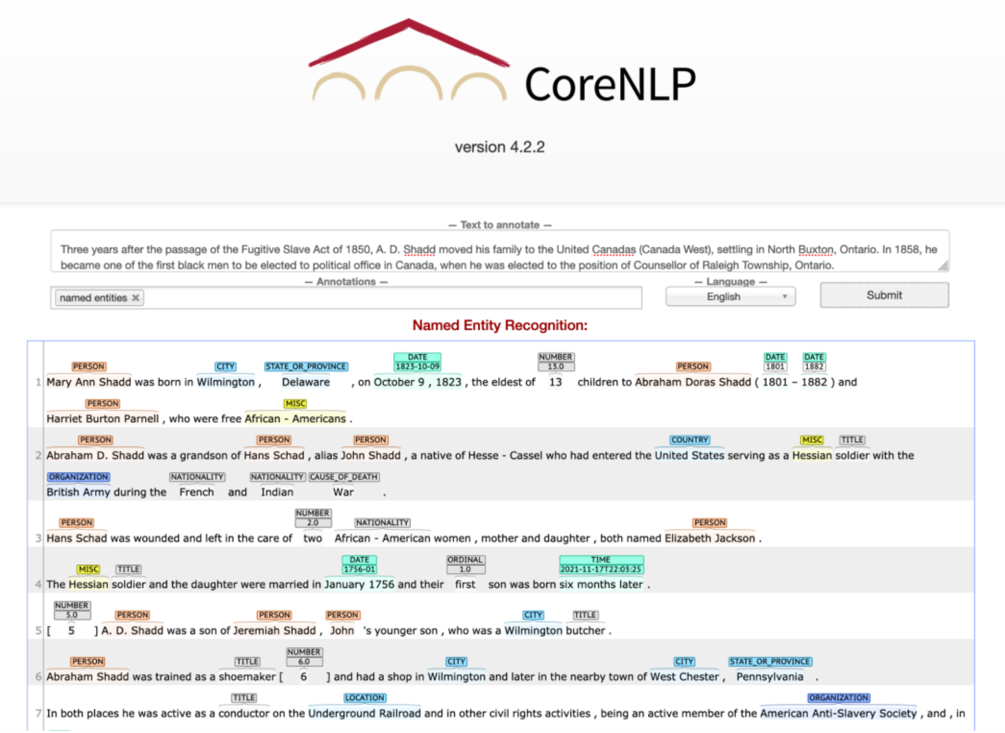
Named Entity Recognition

Analyzing Texts with Named Entity Recognition



Four months DATE after she had gone to Paris GPE , Mary Wollstonecraft PERSON met at the house of a merchant, with whose wife she had become intimate, an American NORP named Gilbert Imlay PERSON . He won her affections. That was in April, 1793 DATE . He had no means, and she had home embarrassments, for which she was unwilling that he should become in any way responsible. A part of the new dream in some minds then was of a love too pure to need or bear the bondage of authority. The mere forced union of marriage ties implied, it was said, a distrust of fidelity. When Gilbert Imlay PERSON would have married Mary Wollstonecraft PERSON , she herself refused to bind him; she would keep him legally exempt from her responsibilities towards the father, sisters, brothers, whom she was supporting. She took his name and called herself his wife, when the French Convention ORG , indignant at the conduct of the British Government ORG , issue a decree from the effects of which she would escape as the wife of a citizen of the United States GPE . But she did not marry. She witnessed many of the horrors that came of the loosened

Named Entity Recognition (NER) in Practice



Three years after the passage of the Fugitive Slave Act of 1850, A. D. Shadd moved his family to the United Canadas (Canada West), settling in North Buxton, Ontario. In 1858, he became one of the first black men to be elected to political office in Canada, when he was elected to the position of Counsellor of Raleigh Township, Ontario.

Named Entity Recognition:

1 Mary Ann Shadd was born in Wilmington , Delaware , on October 9 , 1823 , the eldest of 13 children to Abraham Doras Shadd (1801 – 1882) and Harriet Burton Parnell , who were free African - Americans .

2 Abraham D. Shadd was a grandson of Hans Schad , alias John Shadd , a native of Hesse - Cassel who had entered the United States serving as a Hessian soldier with the British Army during the French and Indian War .

3 Hans Schad was wounded and left in the care of two African - American women , mother and daughter , both named Elizabeth Jackson .

4 The Hessian soldier and the daughter were married in January 1756 and their first son was born six months later .

5 [5] A. D. Shadd was a son of Jeremiah Shadd , John 's younger son , who was a Wilmington butcher .

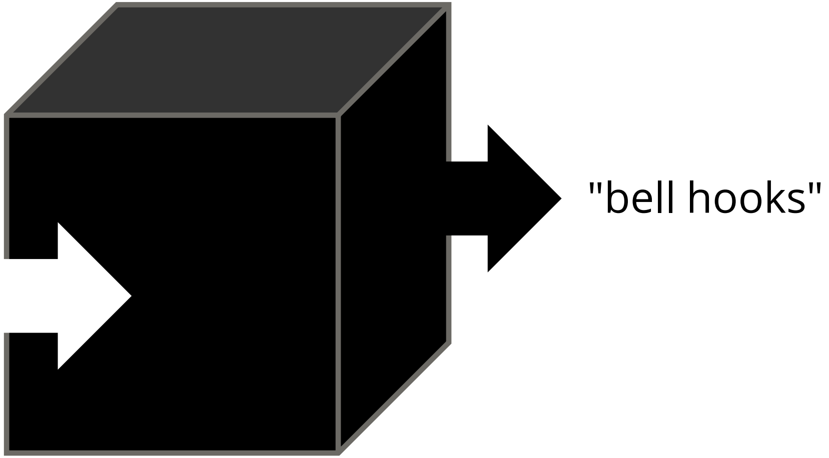
6 Abraham Shadd was trained as a shoemaker [6] and had a shop in Wilmington and later in the nearby town of West Chester , Pennsylvania .

7 In both places he was active as a conductor on the Underground Railroad and in other civil rights activities , being an active member of the American Anti-Slavery Society , and , in

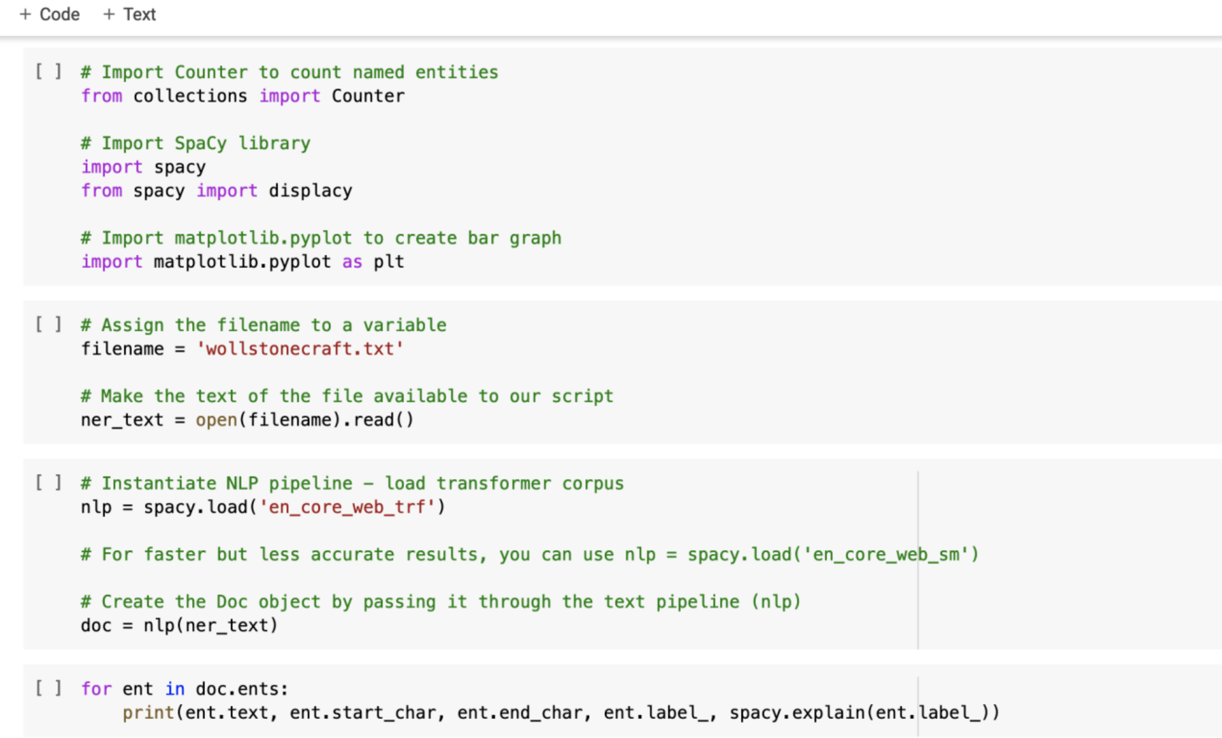
How Named Entity Recognition (NER) Works

Training dataset

"..and soon the white walls and flowery garden of Fort William, the Hudson Bay Company's trading post. The rockery in the centre of the garden would have gladdened the heart of an Ontario gardener. I believe that wealthy people there have had large fragments of Lake Superior rock brought down to adorn their lawns and gardens. We found friends at the fort in the factor and his family, with whom we spent a pleasant half-hour. Mr. McIntyre is well known, and many will owe him gratitude for kindness as long as Fort William or the Canada Pacific Railway remains in their memory."



Try it out in Jupyter Notebooks... Make a copy!



```
[ ] # Import Counter to count named entities
    from collections import Counter

    # Import SpaCy library
    import spacy
    from spacy import displacy

    # Import matplotlib.pyplot to create bar graph
    import matplotlib.pyplot as plt
```

```
[ ] # Assign the filename to a variable
    filename = 'wollstonecraft.txt'

    # Make the text of the file available to our script
    ner_text = open(filename).read()
```

```
[ ] # Instantiate NLP pipeline - load transformer corpus
    nlp = spacy.load('en_core_web_trf')

    # For faster but less accurate results, you can use nlp = spacy.load('en_core_web_sm')

    # Create the Doc object by passing it through the text pipeline (nlp)
    doc = nlp(ner_text)
```

```
[ ] for ent in doc.ents:
        print(ent.text, ent.start_char, ent.end_char, ent.label_, spacy.explain(ent.label_))
```

Interpreting the Results

- **PERSON** - People (including fictional ones)
- **NORP** - Nationalities, or religious or political groups
- **GPE** - Geopolitical Entity, e.g. city, country, states
- **LOC** - Non GPE locations, mountain ranges, bodies of water
- **FAC** - Buildings, airports, highways, bridges, etc.
- **ORG** - Companies, agencies institutions
- **EVENT** - battles, wars, sports events, etc.

Sentiment Analysis

Sentiment Analysis

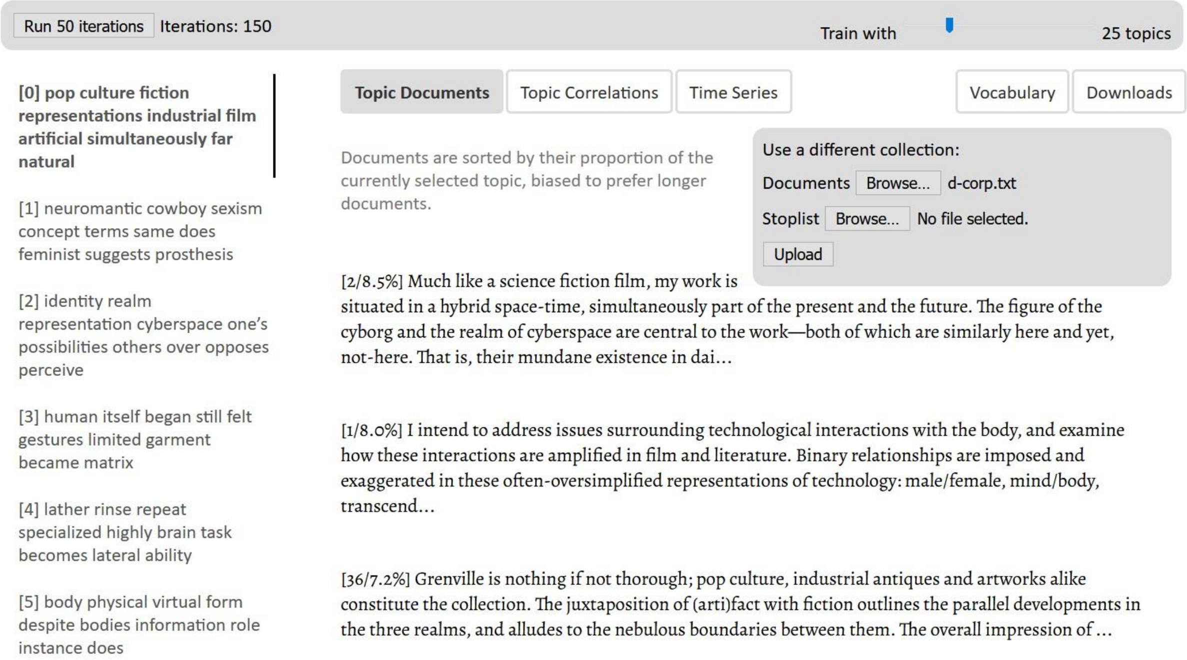
- Go to our shared materials for this workshop: `u.mcmaster.ca/dmds-text-2324`

- Open the file **`sentiment-analysis-2324.ipynb`** and save a copy to your Google Drive.

- Follow along with Jay

Topic Modeling

Discerning Corpus "Topics" with Topic Modeling



Iterations: 150

Train with 25 topics

[0] pop culture fiction representations industrial film artificial simultaneously far natural

[1] neuromantic cowboy sexism concept terms same does feminist suggests prosthesis

[2] identity realm representation cyberspace one's possibilities others over opposes perceive

[3] human itself began still felt gestures limited garment became matrix

[4] lather rinse repeat specialized highly brain task becomes lateral ability

[5] body physical virtual form despite bodies information role instance does

Documents are sorted by their proportion of the currently selected topic, biased to prefer longer documents.

Use a different collection:

Documents d-corp.txt

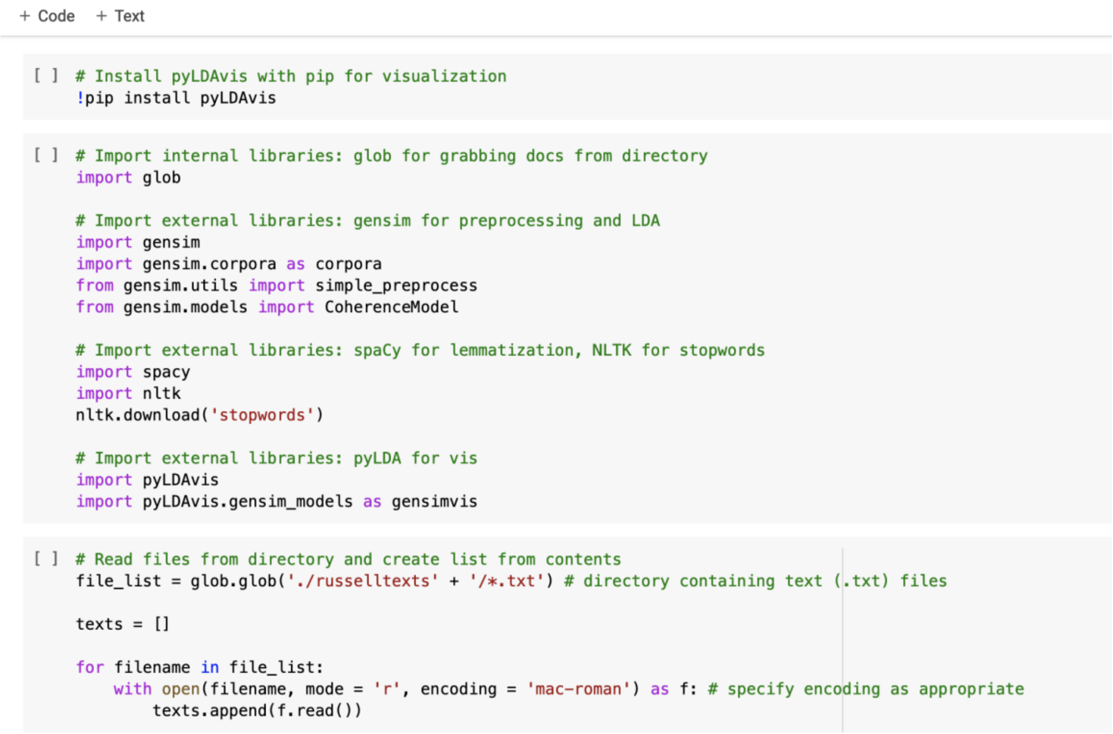
Stoplist No file selected.

[2/8.5%] Much like a science fiction film, my work is situated in a hybrid space-time, simultaneously part of the present and the future. The figure of the cyborg and the realm of cyberspace are central to the work—both of which are similarly here and yet, not-here. That is, their mundane existence in dai...

[1/8.0%] I intend to address issues surrounding technological interactions with the body, and examine how these interactions are amplified in film and literature. Binary relationships are imposed and exaggerated in these often-oversimplified representations of technology: male/female, mind/body, transcend...

[36/7.2%] Grenville is nothing if not thorough; pop culture, industrial antiques and artworks alike constitute the collection. The juxtaposition of (arti)fact with fiction outlines the parallel developments in the three realms, and alludes to the nebulous boundaries between them. The overall impression of ...

Try it out in Jupyter Notebooks... Make a copy!



```
# Install pyLDAvis with pip for visualization
!pip install pyLDAvis
```

```
# Import internal libraries: glob for grabbing docs from directory
import glob

# Import external libraries: gensim for preprocessing and LDA
import gensim
import gensim.corpora as corpora
from gensim.utils import simple_preprocess
from gensim.models import CoherenceModel

# Import external libraries: spaCy for lemmatization, NLTK for stopwords
import spacy
import nltk
nltk.download('stopwords')

# Import external libraries: pyLDA for vis
import pyLDAvis
import pyLDAvis.gensim_models as gensimvis
```

```
# Read files from directory and create list from contents
file_list = glob.glob('./russelltexts' + '/*.txt') # directory containing text (.txt) files

texts = []

for filename in file_list:
    with open(filename, mode = 'r', encoding = 'mac-roman') as f: # specify encoding as appropriate
        texts.append(f.read())
```

Questions &
Final thoughts

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**