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

In-Person Workshop

Computational Text Analyses Bootcamp

u.mcmaster.ca/scds-events

Devon Mordell
Jay Brodeur
Subhanya Sivajothy

DMDS SCDS McMaster University Library

Workshop landing page: u.mcmaster.ca/cta-bootcamp-home

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

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: 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: u.mcmaster.ca/verification

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

Outline & Schedule – Day 1

Segment	Time Allotted	Key Topics / Activities
Introductory remarks	20 minutes	Introduction to text preparation and analysis Overview of concepts and methods
Text preparation	120 minutes	Text prep with OpenRefine Building workflows with Python
Lunch (1200 - 1300)	60 minutes	**Lunch**
Text Analysis	180 minutes	Named Entity Recognition [45 mins] Sentiment Analysis [45 mins] Topic Modeling [45 mins] Stylometry [45 mins]

Workshop landing page: u.mcmaster.ca/cta-bootcamp-home

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 A brief Constellate intro
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: u.mcmaster.ca/cta-bootcamp-home

Learning Objectives

By the end of this bootcamp, 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.
- Explain the benefits and challenges of applying a scripted or semi-scripted approach to text preparation and analysis; identify situations where scripting your work will be beneficial.
- Explain the fundamental considerations when visualizing outputs of text analysis for exploration and dissemination purposes.
- Use a variety of computational tools and approaches to prepare, analyze, and disseminate textual data.
- Identify resources and tutorials for further learning and analyses.

Workshop landing page: u.mcmaster.ca/cta-bootcamp-home

Getting to know...you and your objectives

Please take a few minutes to introduce yourself and share:

- What do you hope to take away from this bootcamp?
- Any text analysis projects you have in planning or progress? What are your materials and objectives?
- If you had an unlimited budget for a vacation, where would you go and what would you do?

Workshop landing page: u.mcmaster.ca/cta-bootcamp-home

Creating a successful bootcamp experience

As facilitators, we will strive to:

- Create an environment of open learning and exploration.
- Explain concepts thoroughly and appropriately.
- Respond to questions (though we may not always have the answer).
- Provide you time to work through examples in a hands-on manner.
- Support your learning during exercises and your independent work

As participants, we ask that you:

- Ask questions (to the facilitators and/or your peers, as appropriate).
- Provide feedback to the instructors (e.g., ask us to slow down, revisit an example, etc.).
- Support your peers, when possible.
- Try some new things (even if outside of your comfort zone) and give yourself time and space to learn. Work at your own pace.

Workshop landing page: u.mcmaster.ca/cta-bootcamp-home

An Introduction

Natural Language Processing (is a big family)

Text and speech recognition / processing
OCR, speech recognition, text-to-speech

Lexical semantics
Named entity recognition, Sentiment analysis, word sense disambiguation

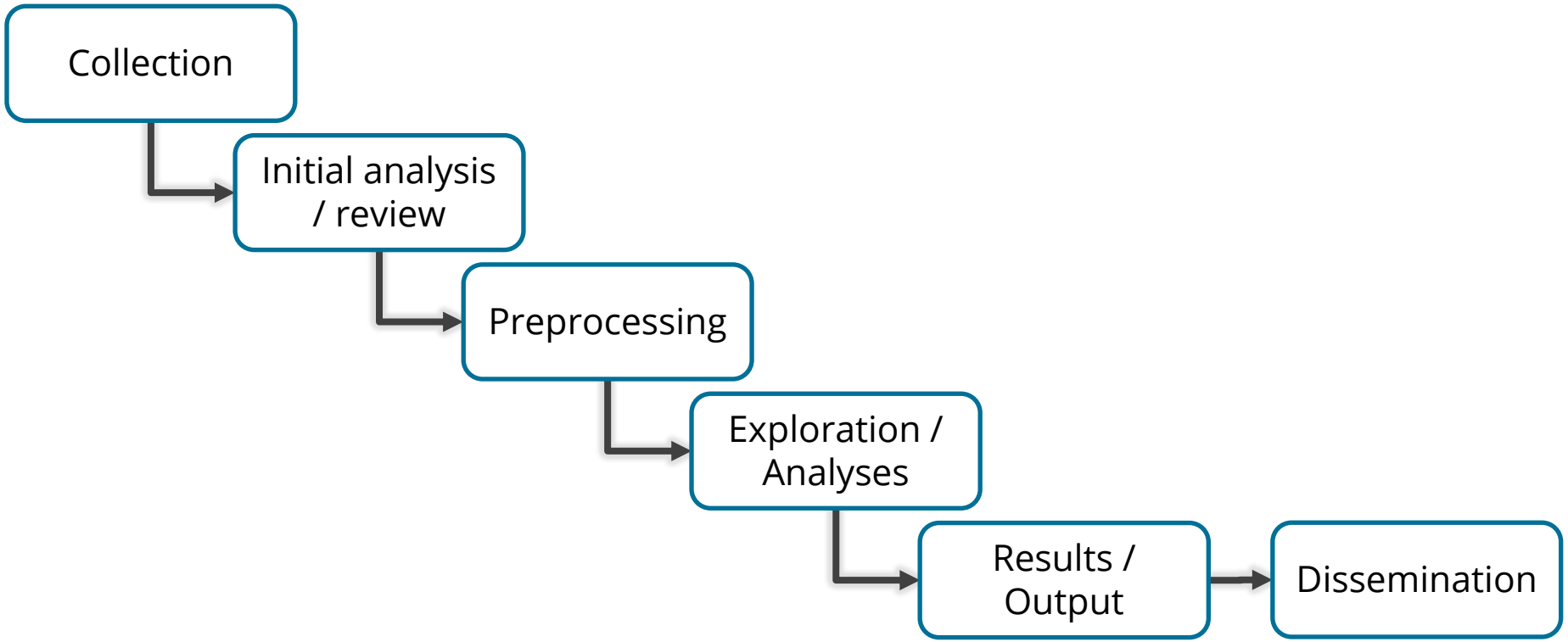
Morphological analysis
Stemming, Lemmatization, Part of Speech Tagging

Relational semantics
Relationship extraction, Semantic parsing

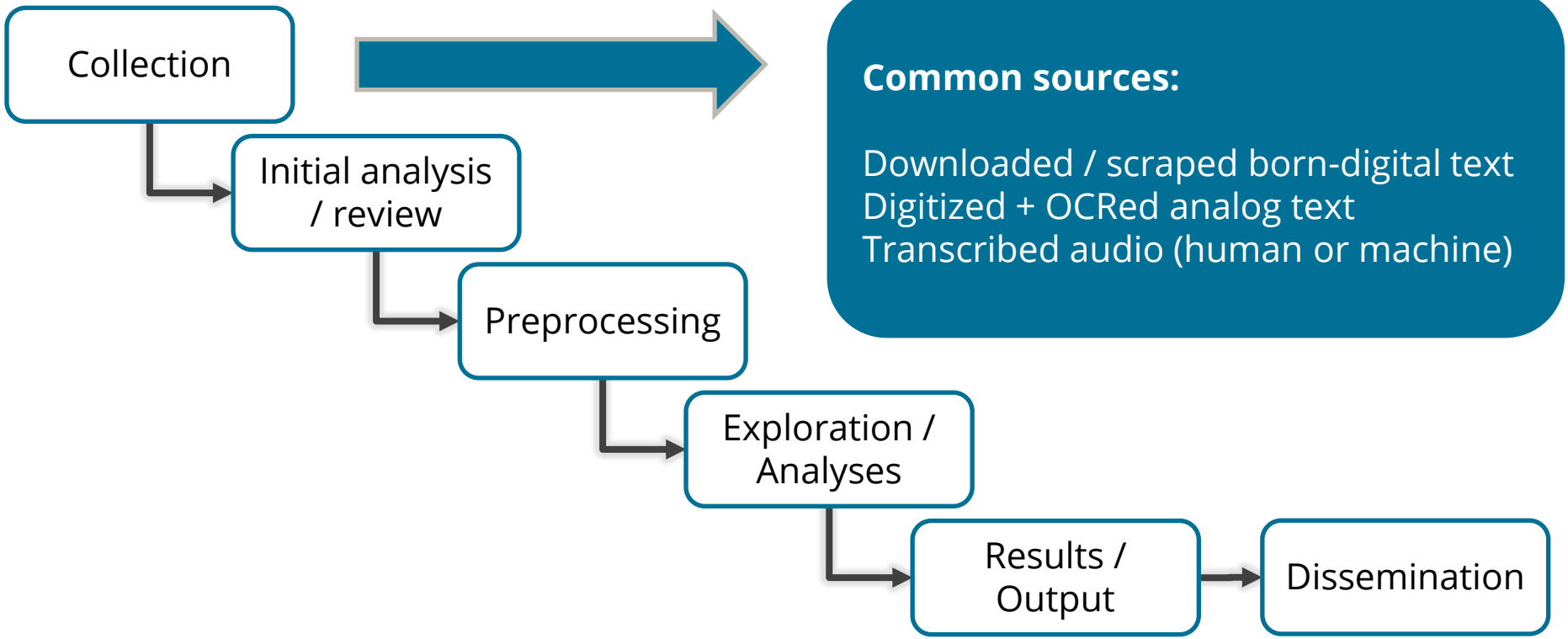
Syntactic analysis
Parsing, Sentence breaking

Discourse semantics
Discourse analysis, Topic segmentation, Argument mining

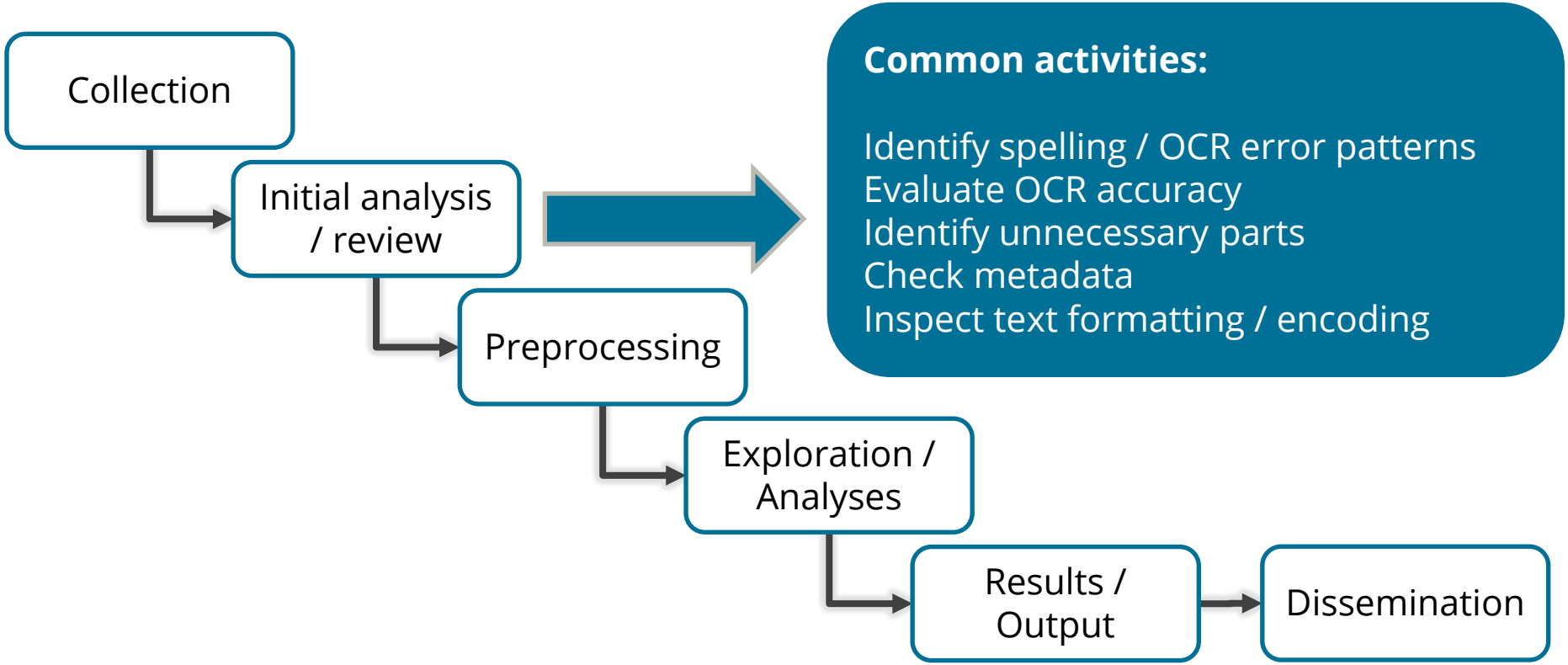
NLP Workflows



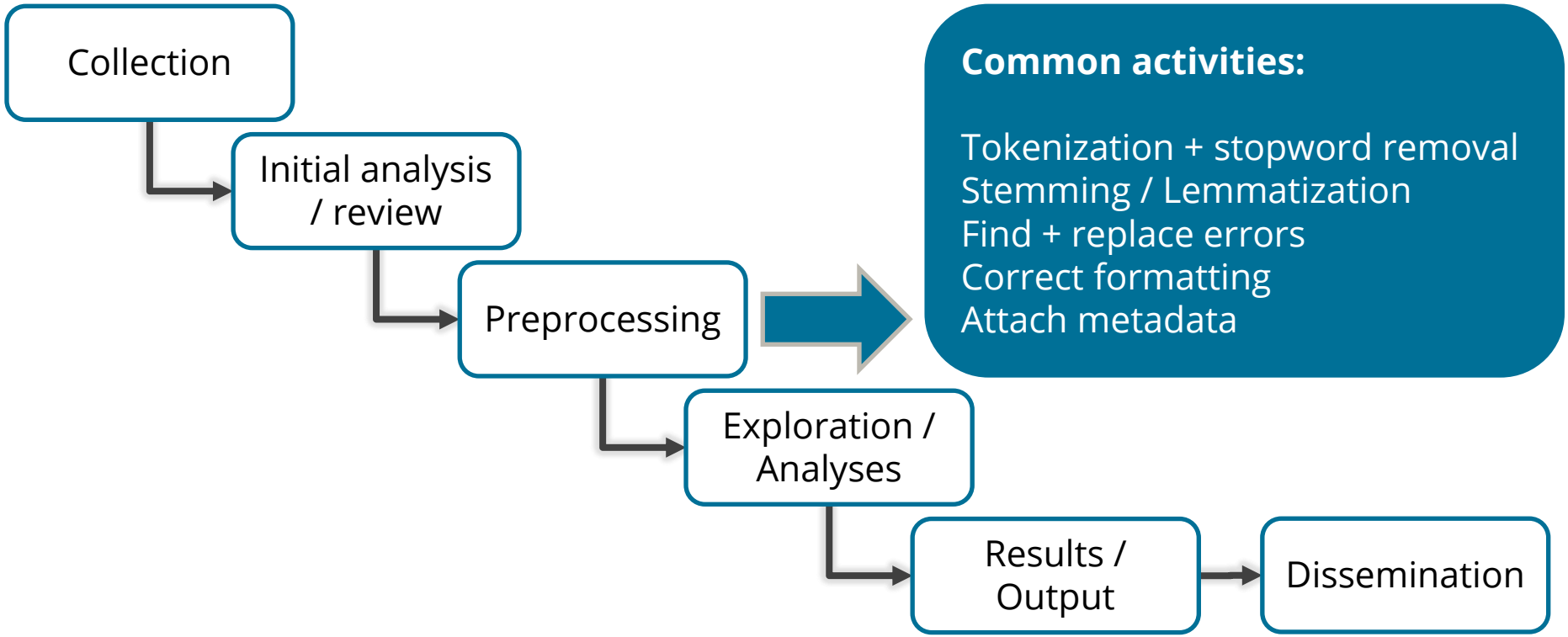
NLP Workflows



NLP Workflows

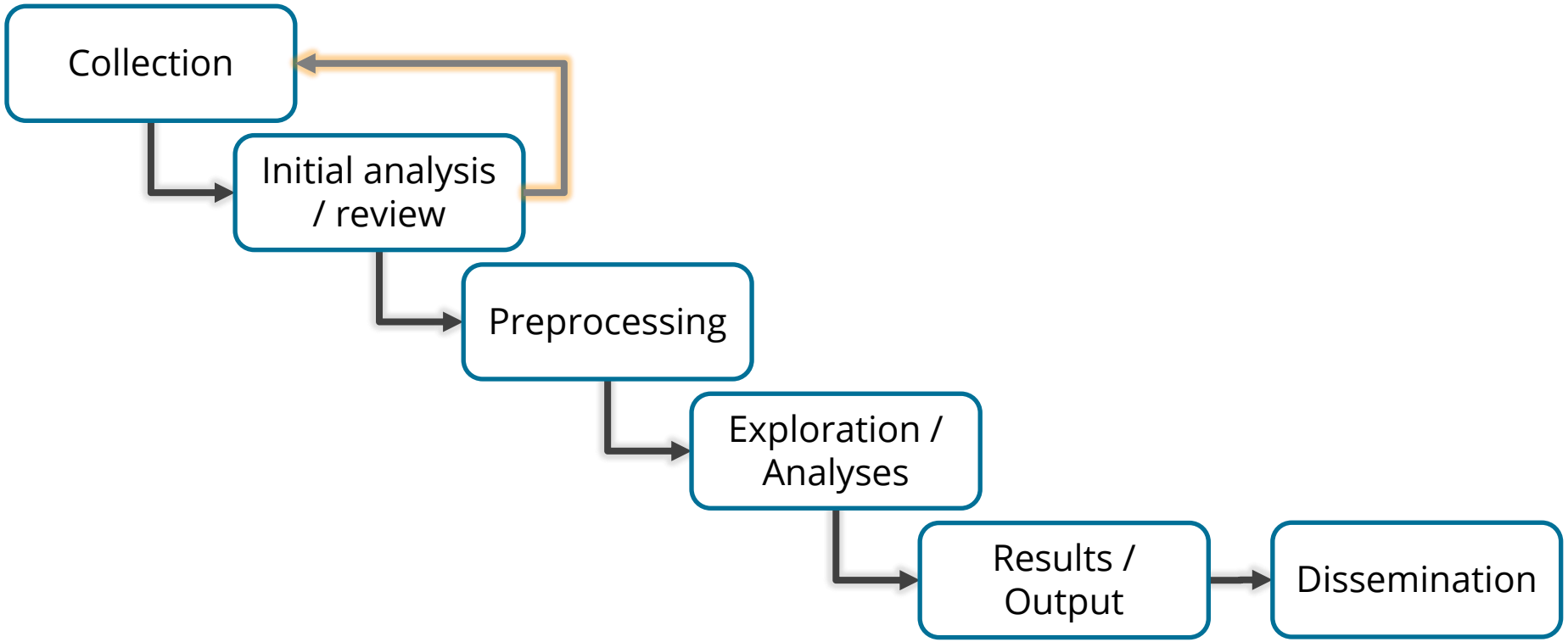


NLP Workflows



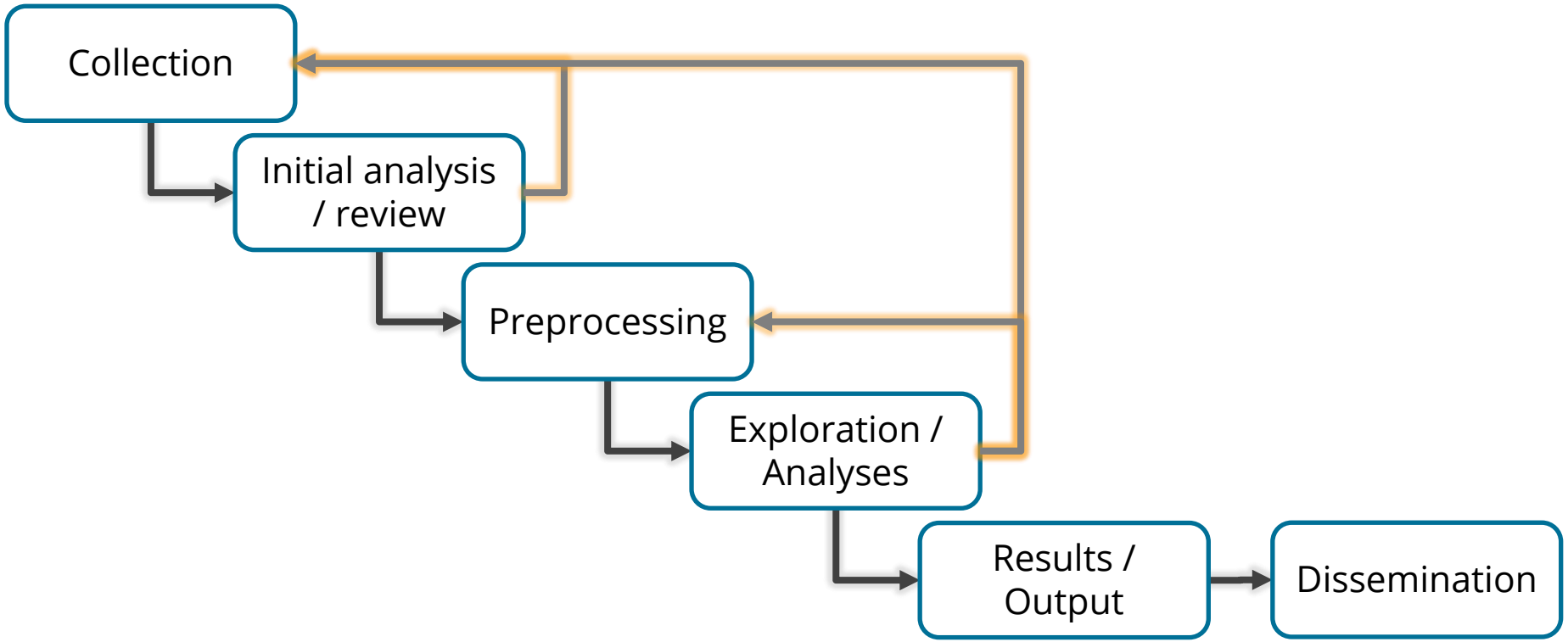
NLP Workflows

… are iterative



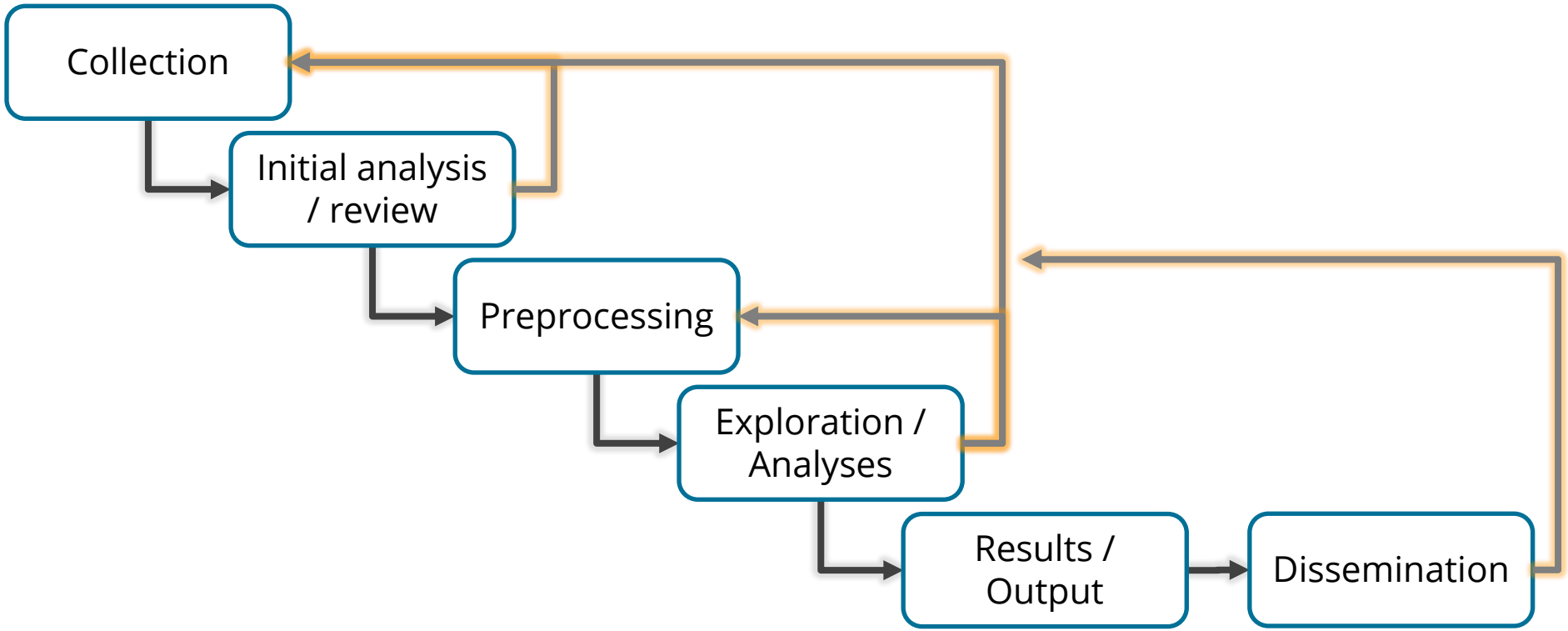
NLP Workflows

… are iterative



NLP Workflows

… are iterative



Getting started:

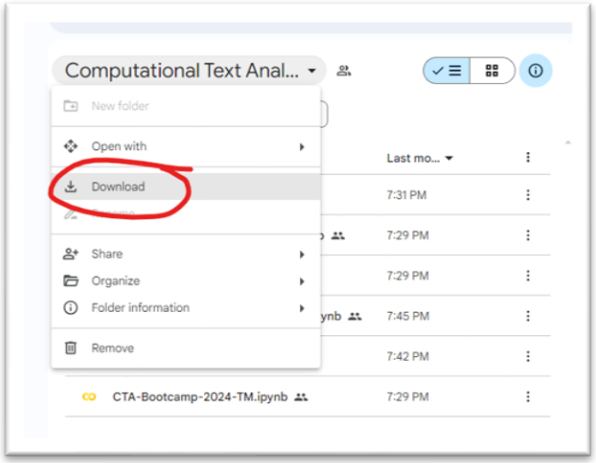
Google Colab & Jupyter notebooks

Get your data

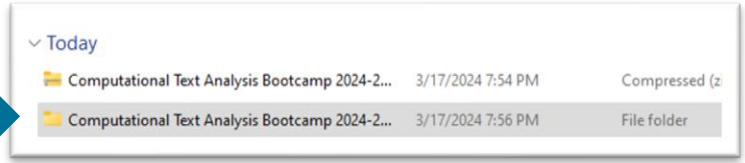
Go to u.mcmaster.ca/cta-bootcamp to download your data and view the workshops for today's workshop.
Follow along with Devon's instructions

Copying materials to your Google Drive

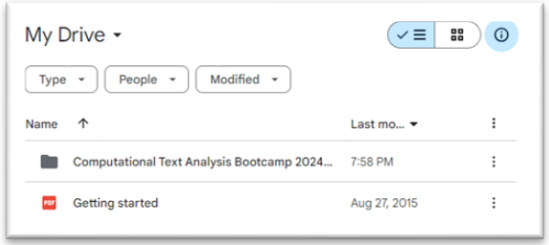
1. Go to u.mcmaster.ca/cta-bootcamp. Download the entire contents to your local computer (as a zipfile).



2. Unzip the zipfile to its own folder on your local computer.



3. Reupload the unzipped folder to your Google Drive



So, why ‘prep’ your text?

Common OCR Issues

SOME COMMENTS ON CORREGGIO IN CONNECTION WITH HIS PICTURES IN DRESDEN.

A few years ago, it would have been hard to tell whether Correggio's *Night* or Raphael's *Madonna Di San Sisto* was the favourite picture of the Dresden Gallery. The little sanctuary where the Virgin with Saint Sixtus floats above the pseudo-altar was then crowded with worshippers as it is now, and Correggio's picture had quite as large and devout a following. But some change in popular taste has evidently taken place, for few people now linger before the *Night*.

What inference is to be drawn? Was the enthusiasm for Correggio merely a fashion which has had its season? He is certainly no longer admired as he was in the first few decades of this century, in the day when no gentleman could afford to be without his theory of the "Correggiosity of Correggio." The explanation is not far to seek.

The enthusiasm for Correggio dates from the time when, all the possible variations having been played upon the themes introduced by Raphael and Michelangelo, the Caracci betook themselves to a comparatively unlaboured field, and founded upon Correggio their school of painting, and thus succeeded in lending a new life to Italian art. Most people, however, appreciate only what is of their own day, and Correggio's interpreters proved far more interesting to their contemporaries than the master himself. The Caracci, Domenichino, Guercino, Guido Reni, and Lanfranco used up all the æsthetic capacity of their admirers, who believed in Correggio as the Catholic peasant doubtless believes in God, although he makes his offerings to the Saints. Furthermore, it was by no means easy to know the master himself. Correggio lived to be scarcely forty. Of his pictures then known the earliest dates from his twenty-first year, and in a career of barely twenty years no painter could have painted enough to fill the various collections of Europe. But in the third decade of this century, the few whose word was law in matters of taste suddenly turned away from Guido, Lanfranco, and their like, and gave themselves up to an unbridled enthusiasm for the Caracci and for their master, Correggio. Later, even the Caracci dropped out of sight, and Correggio stood alone.

The Madonna with St. Francis, No. 150. The Madonna with St. Sebastian, No. 151. The Nativity, called the "Night," No. 152. The Madonna with St. George, No. 153.

73

```
i _ IPS^S^ffclS OME comments on correg ncRftJH GIO IN CONNECTION with
1^58^^^ HIS PICTURES IN DRESDEN. Spl^^T^ES A few years ago, it would
have been hard u|^J^^fev^S to tell whether Correggio's Night or
E^^g^^M Raphael's Madonna Di San Sisto was the ^L^^^^M favourite
picture of the Dresden Gallery. mmSSSmS^mSSm The little sanctuary
where the Virgin with Saint Sixtus floats above the pseudo-altar was
then crowded with worshippers as it is now, and Correggio's picture
had quite as large and devout a fol lowing. But some change in
popular taste has evidently taken place, for few people now linger
before the Night. What inference is to be drawn ? Was the enthusiasm
for Correggio merely a fashion which has had its season ? He is
certainly no longer admired as he was in the first few decades of
this century, in the day when no gentleman could afford to be without
his theory of the " Correggiosity of Correggio." The explanation is
not far to seek. The enthusiasm for Correggio dates from the time
when, all the possible variations having been played upon the themes
introduced by Raphael and Michelangelo, the Caracci betook themselves
to a comparatively unlaboured field, and founded upon Correggio their
school of painting, and thus succeeded in lending a new life to
Italian art. Most people, however, appreciate only what is of their
own day, ana Correggio's in terpreters proved far more interesting to
their contemporaries than the master himself. The Caracci,
Domenichino, Guer cino, Guido Reni, and Lanfranco used up all the
aesthetic capacity of their admirers, who believed in Correggio as
the Catholic peasant doubtless believes in God, although he makes his
offerings to the Saints. Furthermore, it was by no means easy to know
the master himself. Correggio lived to be scarcely forty. Of his
```

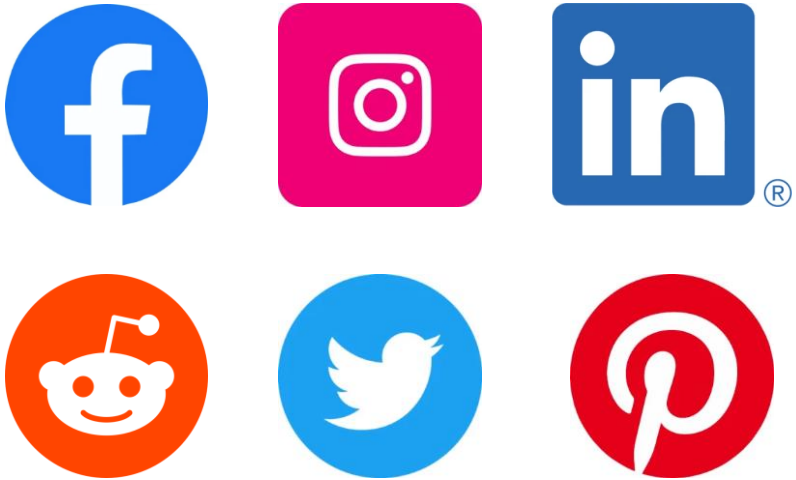
Berenson, B. (1892). Some Comments on Correggio in Connection with His Pictures in Dresden. *The Knight Errant, 1*(3), 73-85. doi:10.2307/25515893

Common Transcription Issues

```
10
00:06:56.910 --> 00:07:07.200
aaa: We work and study on the traditional territory shared between the
holden has shown a confederacy and the addition of a nations, which is
acknowledged in the dish with one spoon off of.

11
00:07:08.220 --> 00:07:19.440
aaa: The wampanoag uses the symbolism of a dish to represent the
territory and one spoon to represent that the people are to share the
resources of this land and take only what they need.
```

We work and study on the traditional territory shared between the Haudenosaunee confederacy and the Anishinabe nations, which was acknowledged in the Dish with One Spoon Wampum belt. The wampum uses the symbolism of a dish to represent the territory, and one spoon to represent that the people are to share the resources of the land and only take what they need

Srsly, this stuff can #lackconsistency



Born-digital text (especially from SM) may be well-structured, but can also:

- contain a lot of spelling errors (sometimes intentionally) and non-words
- use non-traditional representations and abbreviations
- include non-textual data like markup and embedded scripts
- have different encodings

Hands-on text prep with OpenRefine

OpenRefine – for text preparation???

- Graphical interface (GUI)
- Non-destructive editing
- Self-documenting
- Reproducibility of steps
- New: portable!

Using OpenRefine for text prep is best suited to....

- Scanned print documents:
  - good contrast
  - clearly defined boundaries
  - no or few tables, images or equations
- e.g. typed correspondence, minutes, manuscripts, reports, etc.

conversations and correspondence with municipal employees and regional and local MOE staff, a subsurface soils drilling program, a ground water monitoring program, a surface water monitoring program, and a landfill gas monitoring program. The work carried out under each of these parts of the work program is outlined in Sections 2.1 through 2.5.

2.1 DESK TOP INVENTORY

A series of air photos from 1953 to 1987 were used to identify the extent and process of filling over the period the landfill was operational. Historical water quality data for the Bay of Quinte and the Moira River provided by the MOE were reviewed. Recent study results on the Bay of Quinte were provided by the Bay of Quinte Remedial Action Plan. Geological and hydrogeological information was provided by the Ministry of Transportation through their work on the construction of Highway 62 on Zwick's Island. All of the background data reviewed was used in the design and interpretation of the drilling and monitoring programs.

2.2 SITE VISIT

A site visit was conducted on April 2, 1990 in order to relate the data collected during the desk top study to actual field conditions, and to collect additional data on the physical setting of the site. In addition to meeting with the MOE, GLL staff also met with a Municipal employee who worked at the landfill in the 1960's. This meeting provided first-hand information with respect to the nature of the refuse and the filling locations. The information obtained through the site visit assisted GLL in finalizing the drilling program as well as providing input to the health and safety protocols which would be followed during the course of field work.

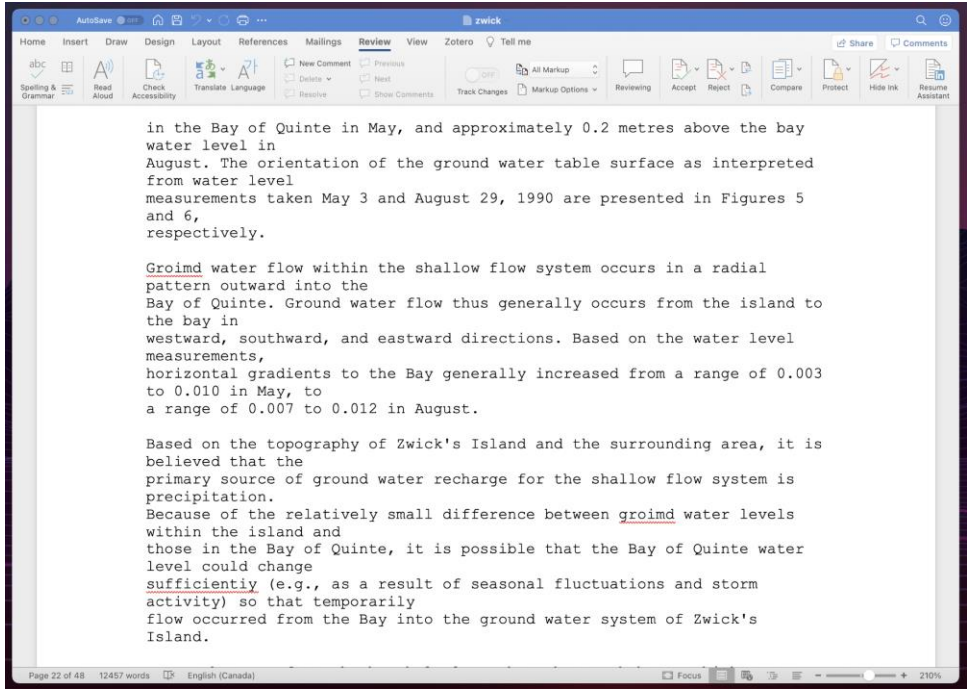
During the site visit surface water drainage and pathways were observed and noted. Conductivity measurements of surface water were taken at several locations around the site along with observations of iron staining and vegetation loss in roadside and drainage ditches in the northeast corner of the Island. Observations were also made as to the occurrence of ground settlement, locations of exposed refuse, and evidence of leachate seeps.

During the course of the field visit, GLL staff visited the City of Belleville Town Offices and obtained historical maps of Zwick's Island. This information assisted in establishing the original shoreline of the island and the landfilling locations.

The Dataset

- “Zwick's Island landfill environmental investigations” (1991)
  - Copied and pasted from full text on Internet Archive
- Transformations:
  - removed preamble
  - removed tabular data

Initial Data Analysis: in MS Word



in the Bay of Quinte in May, and approximately 0.2 metres above the bay water level in
August. The orientation of the ground water table surface as interpreted from water level
measurements taken May 3 and August 29, 1990 are presented in Figures 5 and 6,
respectively.

Groimd water flow within the shallow flow system occurs in a radial pattern outward into the
Bay of Quinte. Ground water flow thus generally occurs from the island to the bay in
westward, southward, and eastward directions. Based on the water level measurements,
horizontal gradients to the Bay generally increased from a range of 0.003 to 0.010 in May, to
a range of 0.007 to 0.012 in August.

Based on the topography of Zwick's Island and the surrounding area, it is believed that the
primary source of ground water recharge for the shallow flow system is precipitation.
Because of the relatively small difference between groimd water levels within the island and
those in the Bay of Quinte, it is possible that the Bay of Quinte water level could change
sufficientiy (e.g., as a result of seasonal fluctuations and storm activity) so that temporarily
flow occurred from the Bay into the ground water system of Zwick's Island.

A closer look at our errors...

- Pay attention to surrounding letters

  (i.e. note **context**, not just the errors)
- Try to observe and record patterns
  - We will use the patterns to correct multiple errors at a time

undenake

concepmal

smdy

smdy

smdy

backgroimd

acmal

coUect

Mimicipal

fmalizing

Dtiring

aroimd

stammg

landfiUed

moimted

Figiu

Stratigraphie

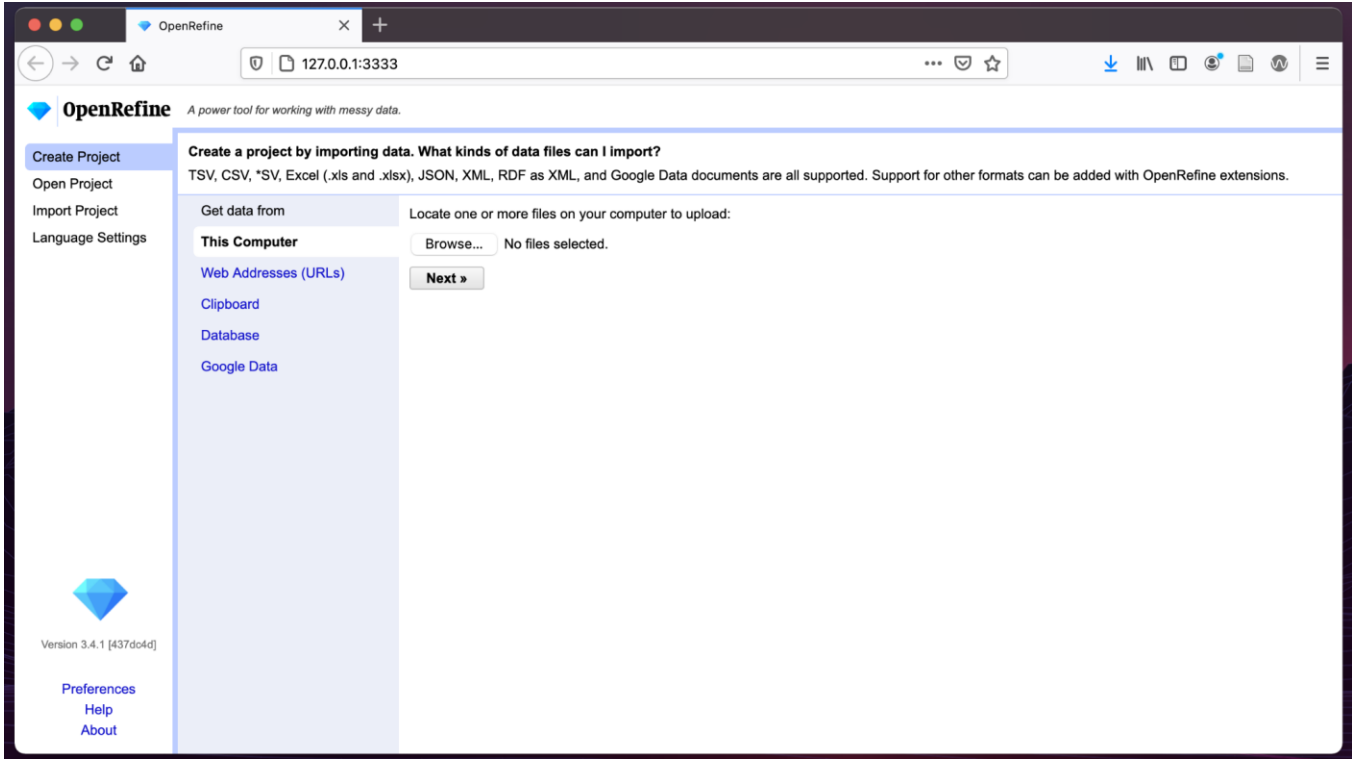
concurrenLly

Ruoride

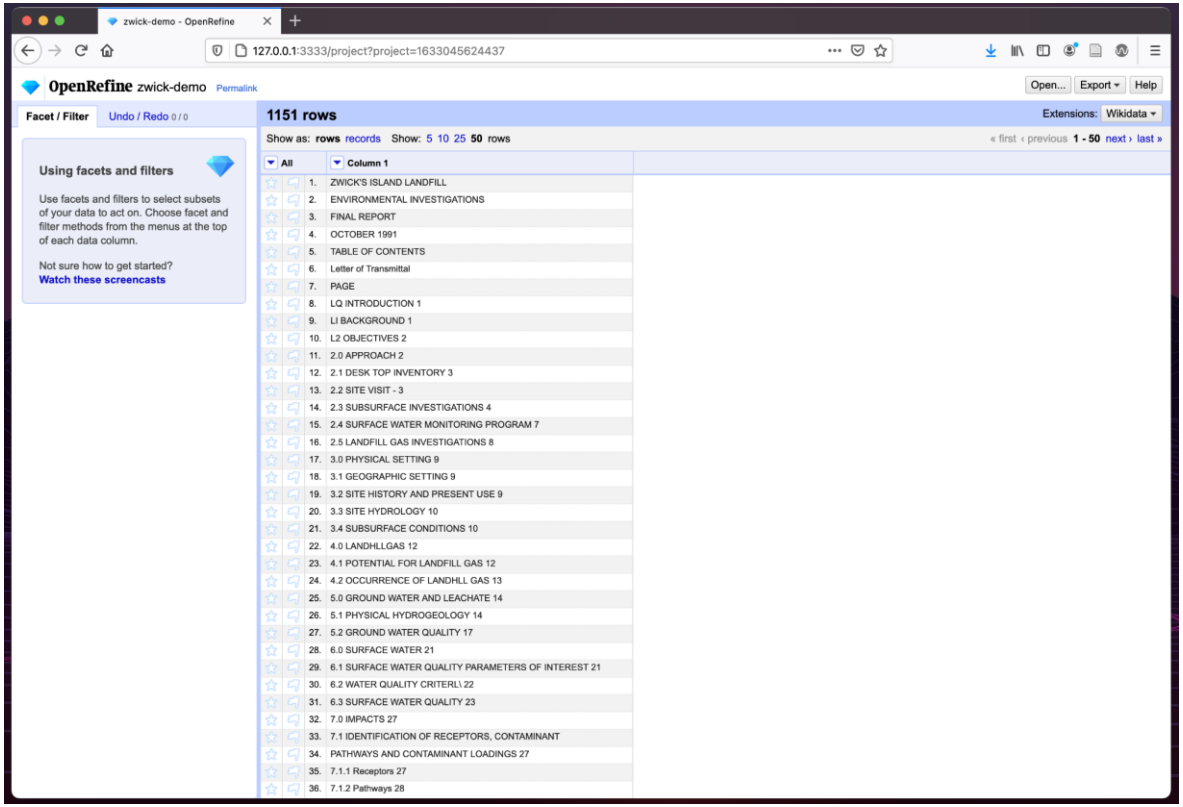
Qiloride

Organo

Open OpenRefine

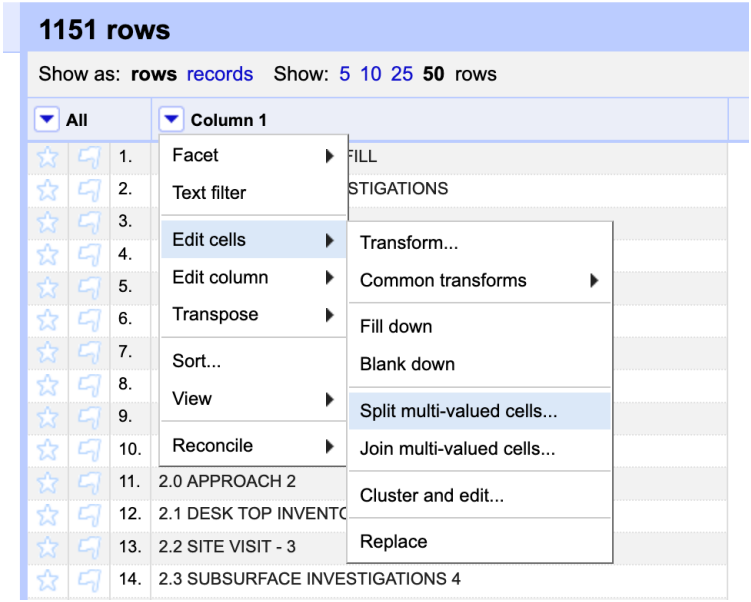


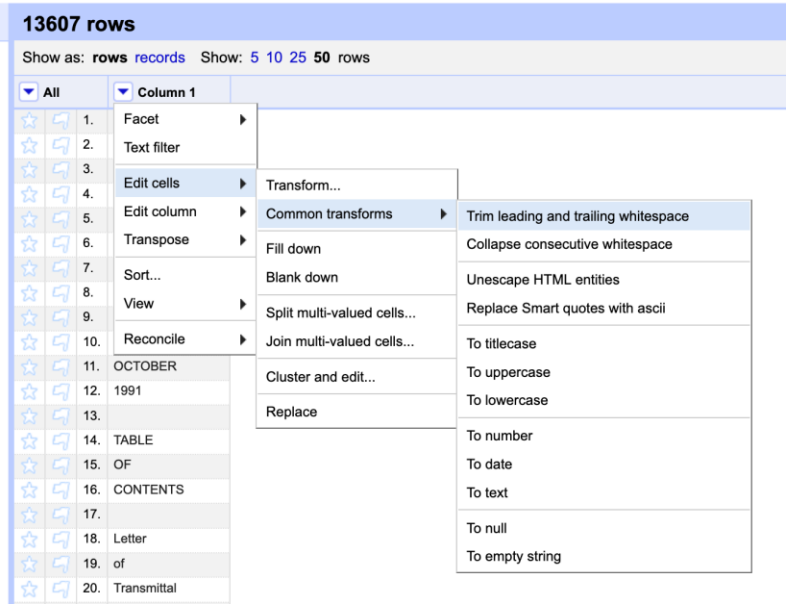
Initial Data Analysis: in OpenRefine



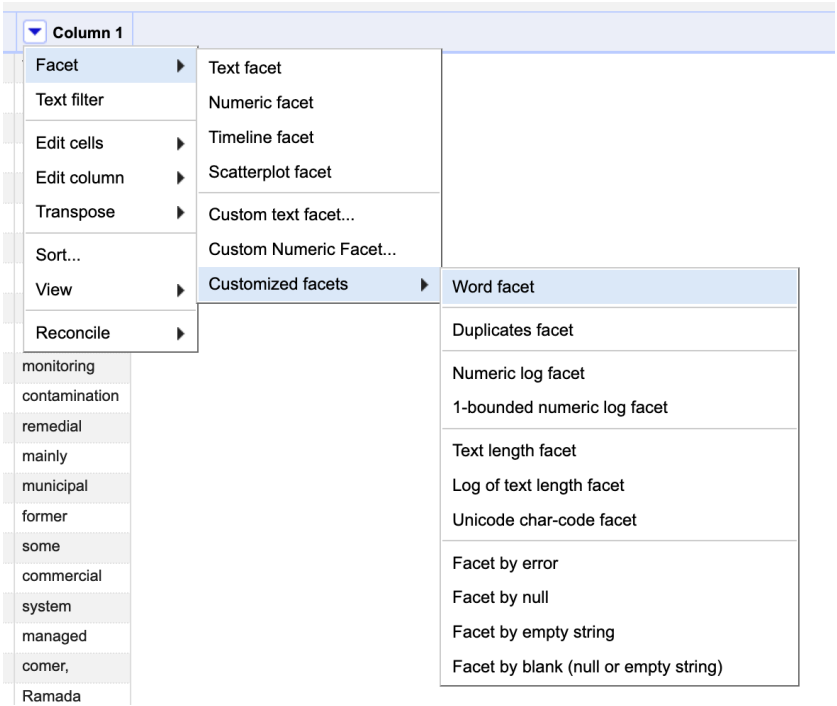
Prepare Dataset for Text Analysis

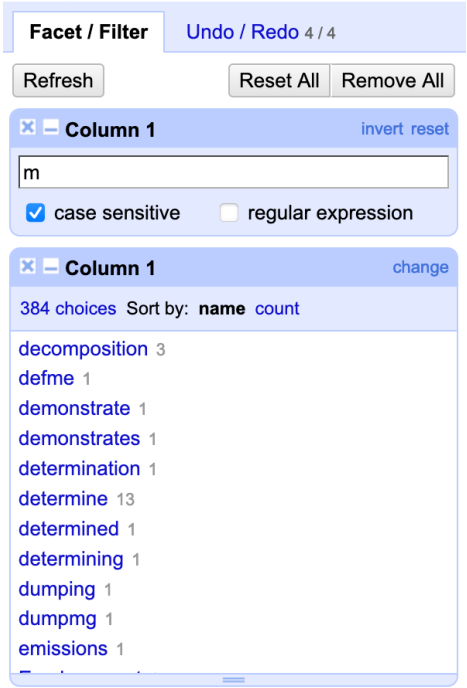
Tokenize, trim and remove blank rows



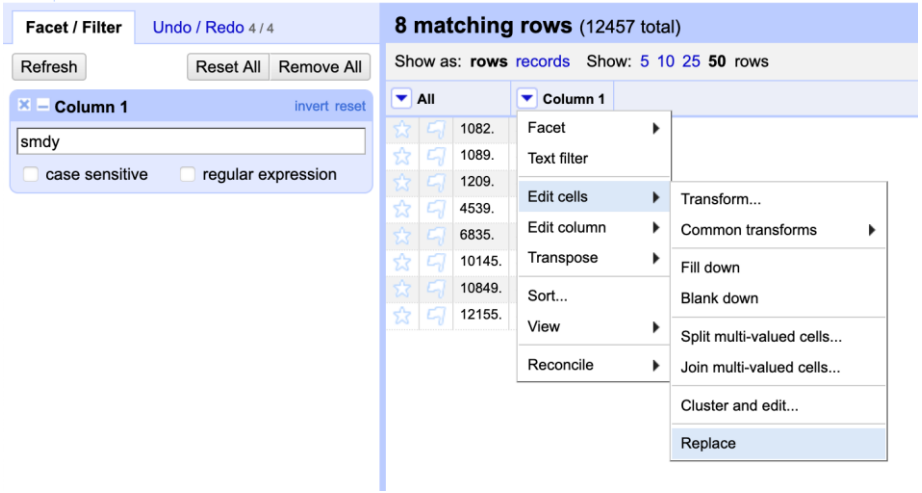


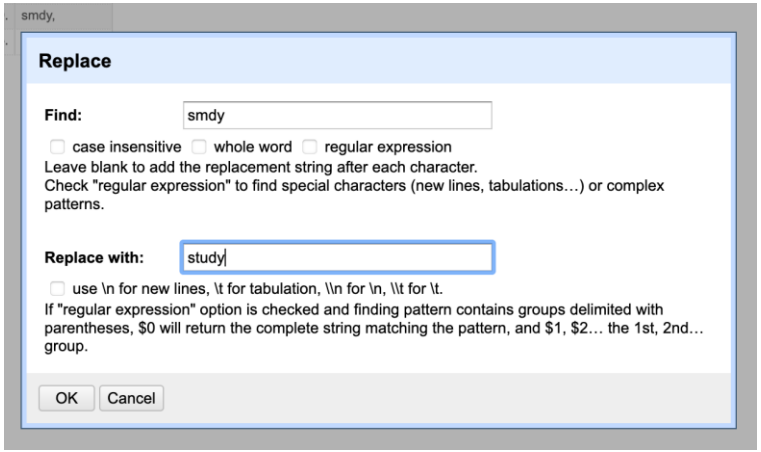
Filter and Facet to Find Errors





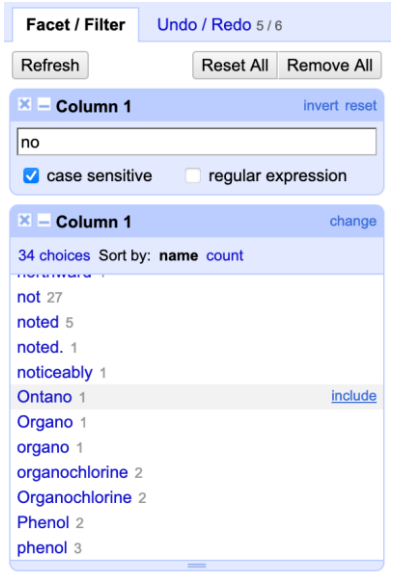
Find and Replace with Text Filter

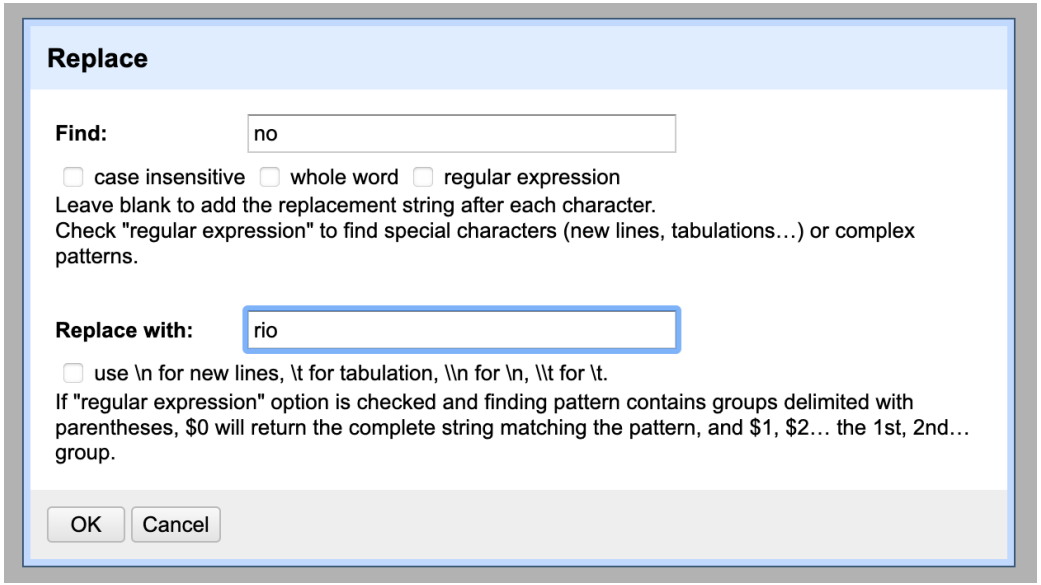




Find and Replace with Text Filter

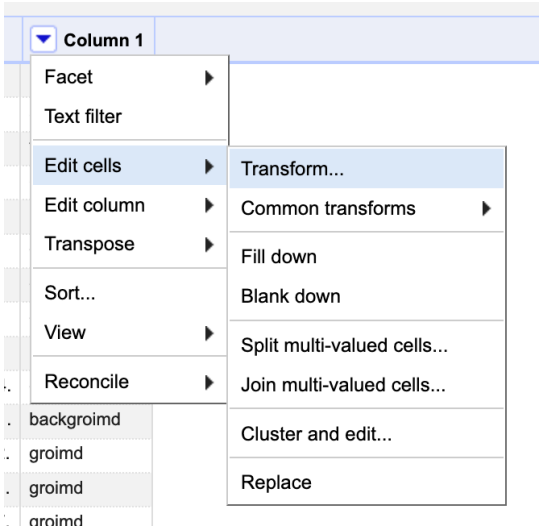
Filter by “no” as case sensitive, then word facet & include

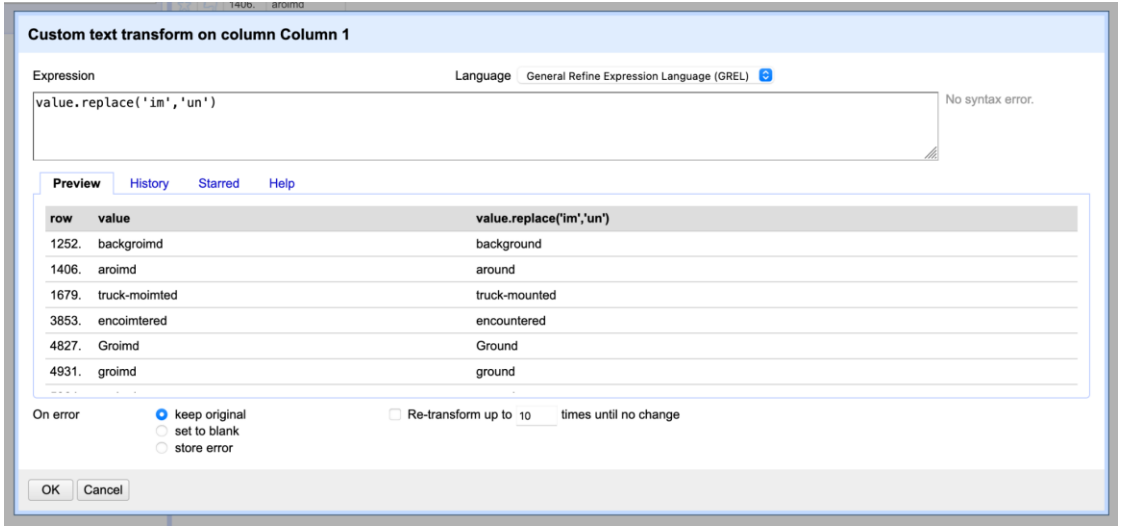




Find and Replace with GREL

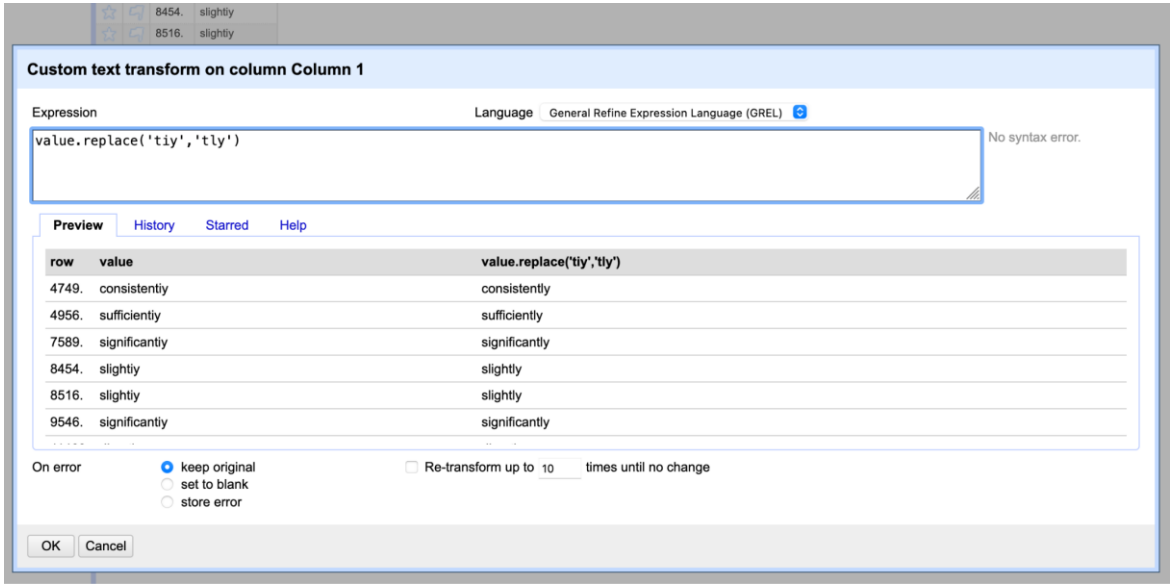
Filter by “oim” (oun) → value.replace(‘im’, ‘un’)





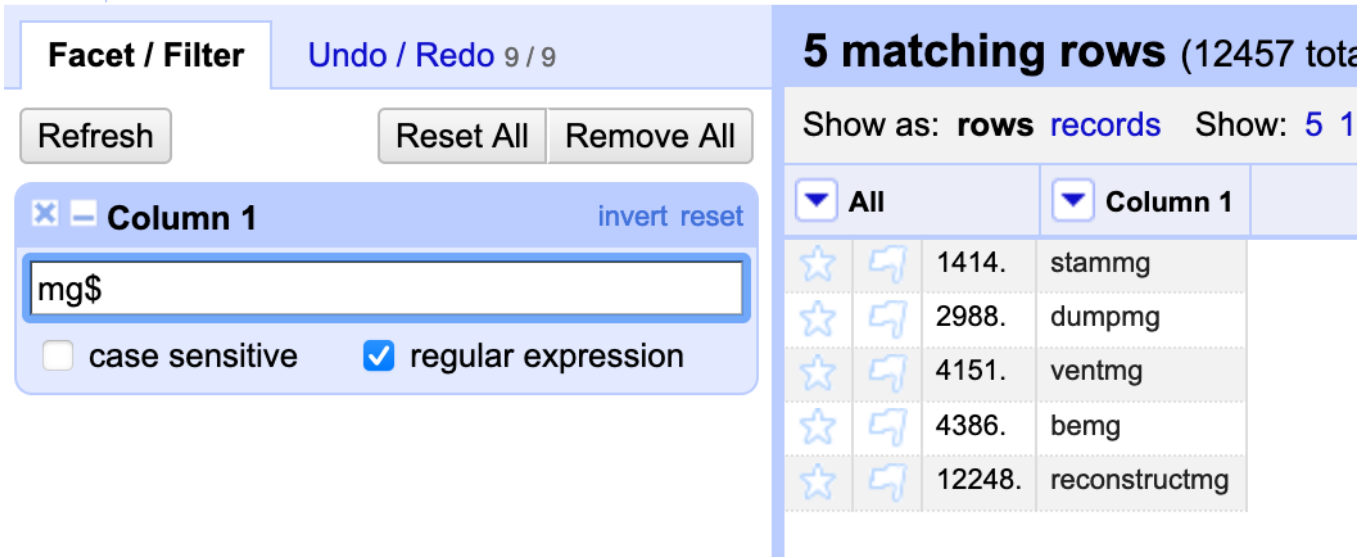
Find and Replace with GREL

Try it out: filter by “tiy” (tly) → `value.replace(‘tiy’, ‘tly’)`



Find and Replace with Regular Expression (Regex)

Filter by "mg" → try `mg$` with "regular expression" checked ("ing" as "mg" error)



Quick Guide to Regex

^ *start of expression*

$ *end of expression*

E.g. **^T$** will only return cells with “T”

^mn will only return cells that **start** with “mn”

ent$ will only return cells that **end** with “ent”

Quick Guide to Regex

[string] - contains any of the letters

[^string] - does not contain the letters

E.g. **[iou]m** will return words that contain “im,” “om” and “um”

ti[^o] will exclude “tion”

Find and Replace with Regex ctd.

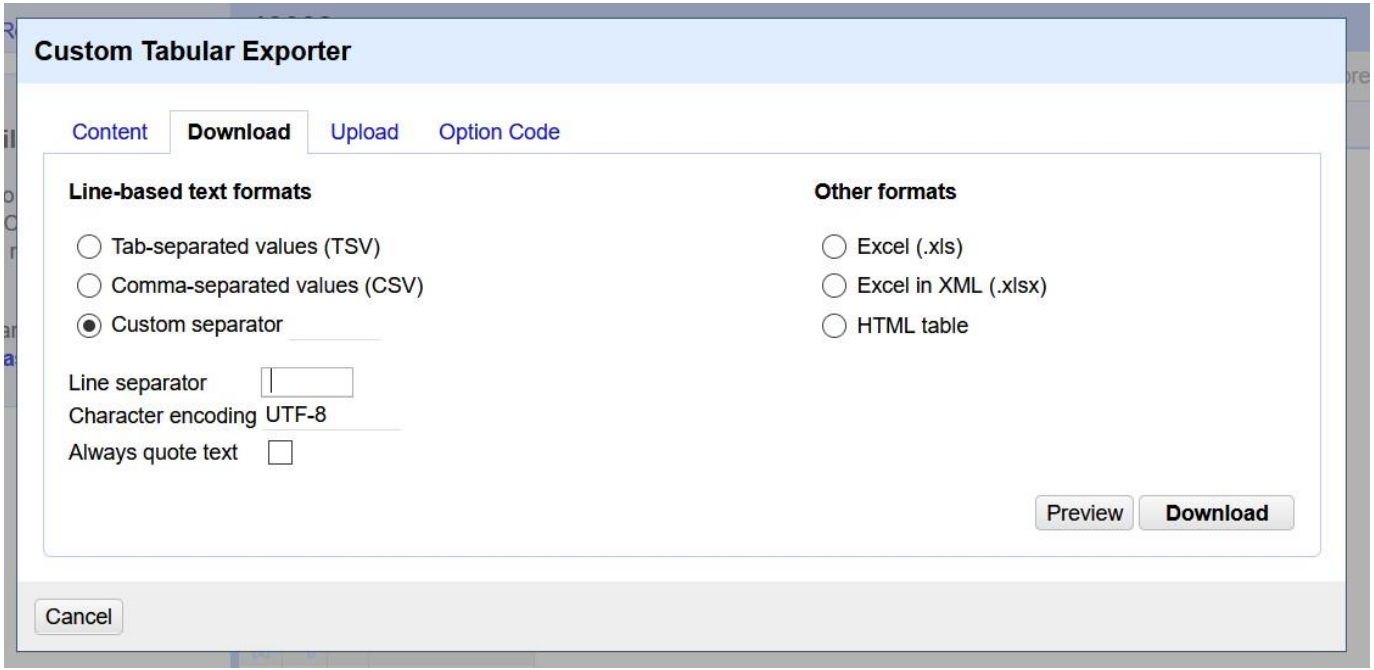
Try it out:

^mt → value.replace(‘m’ , ‘in’)

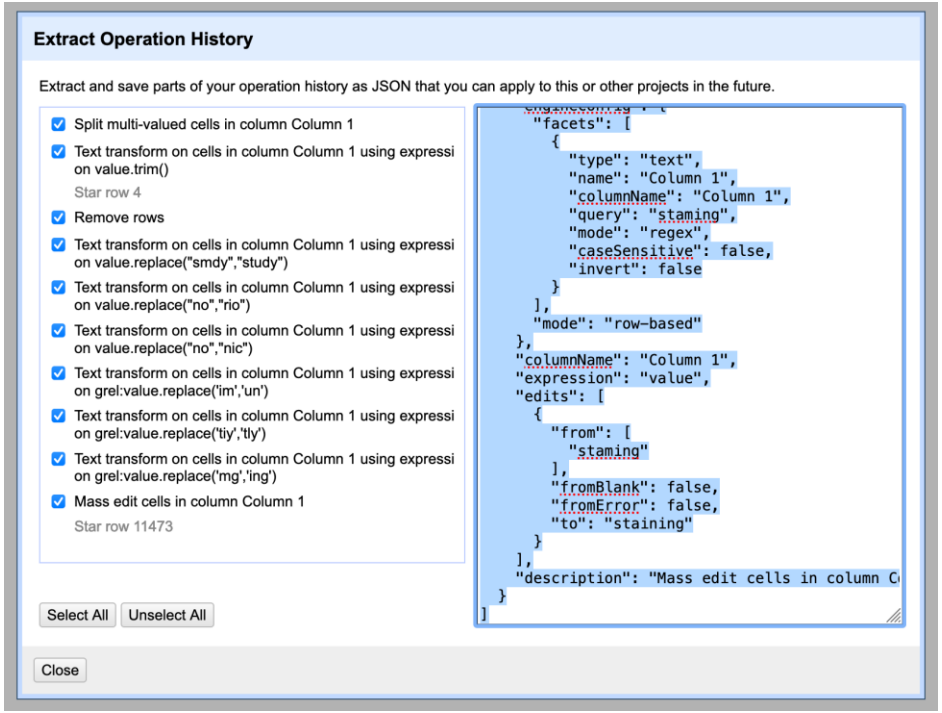
[a-z]U → value.replace(‘U’ , ‘ll’) [with case sensitive checked]

Others…?

Reconstitute your Document



Export your “Recipe” of Tasks



Programmatic approaches with Python

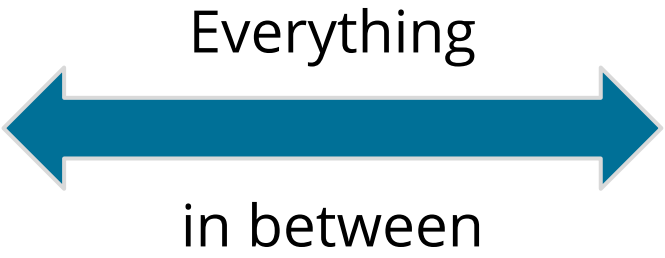
Text prep and analysis as a continuum of mediation

Completely manual

Completely automated

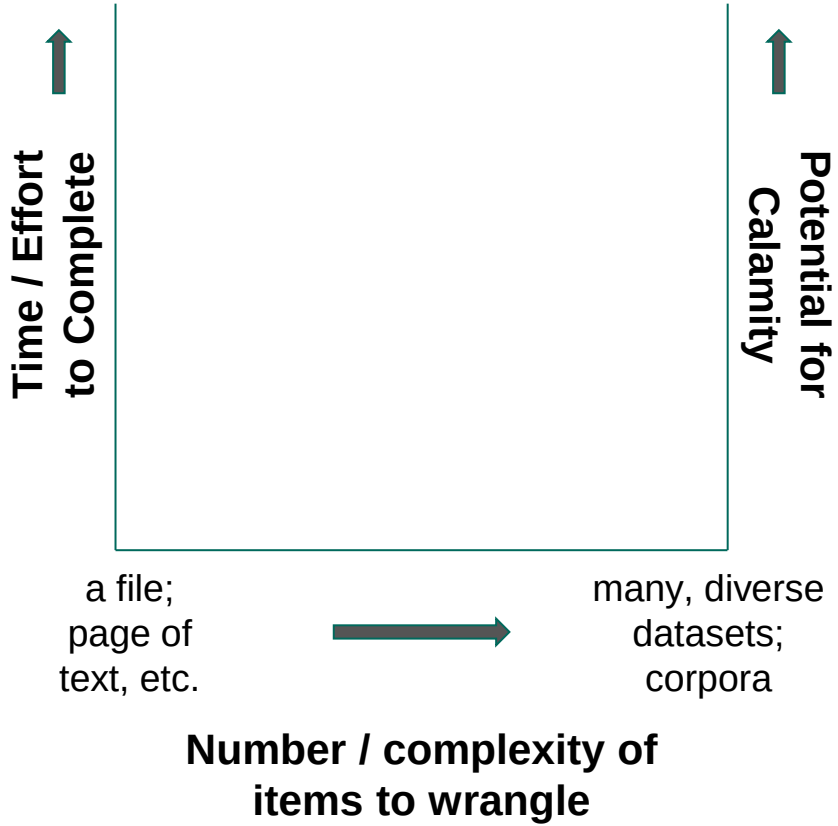
Text prep and analysis as a continuum of mediation

Completely manual

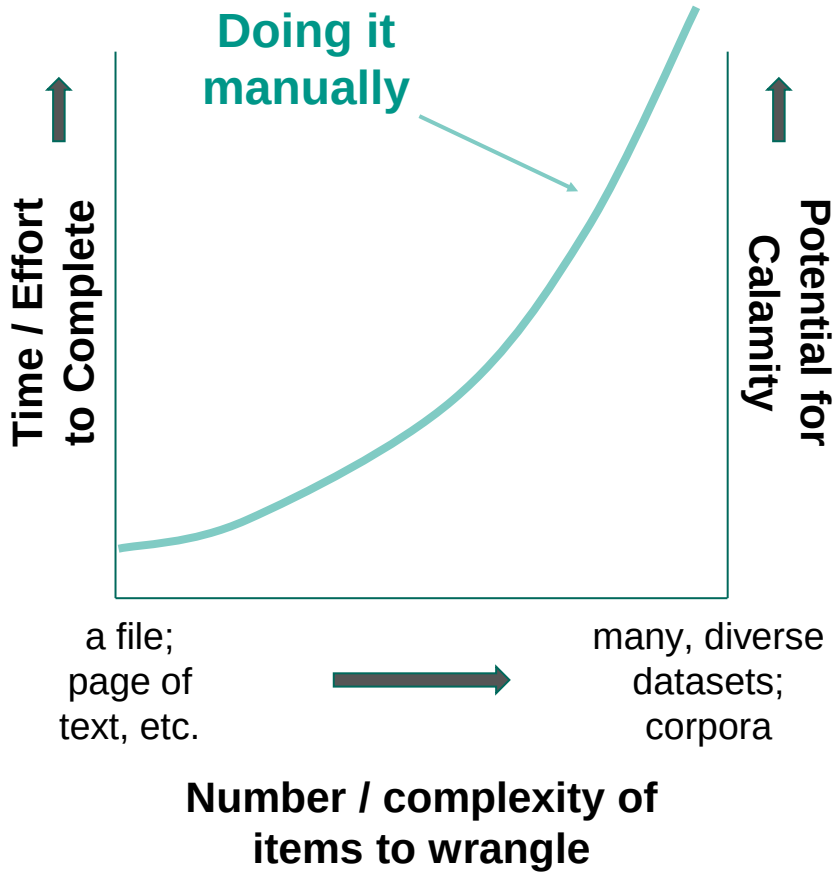


Completely automated

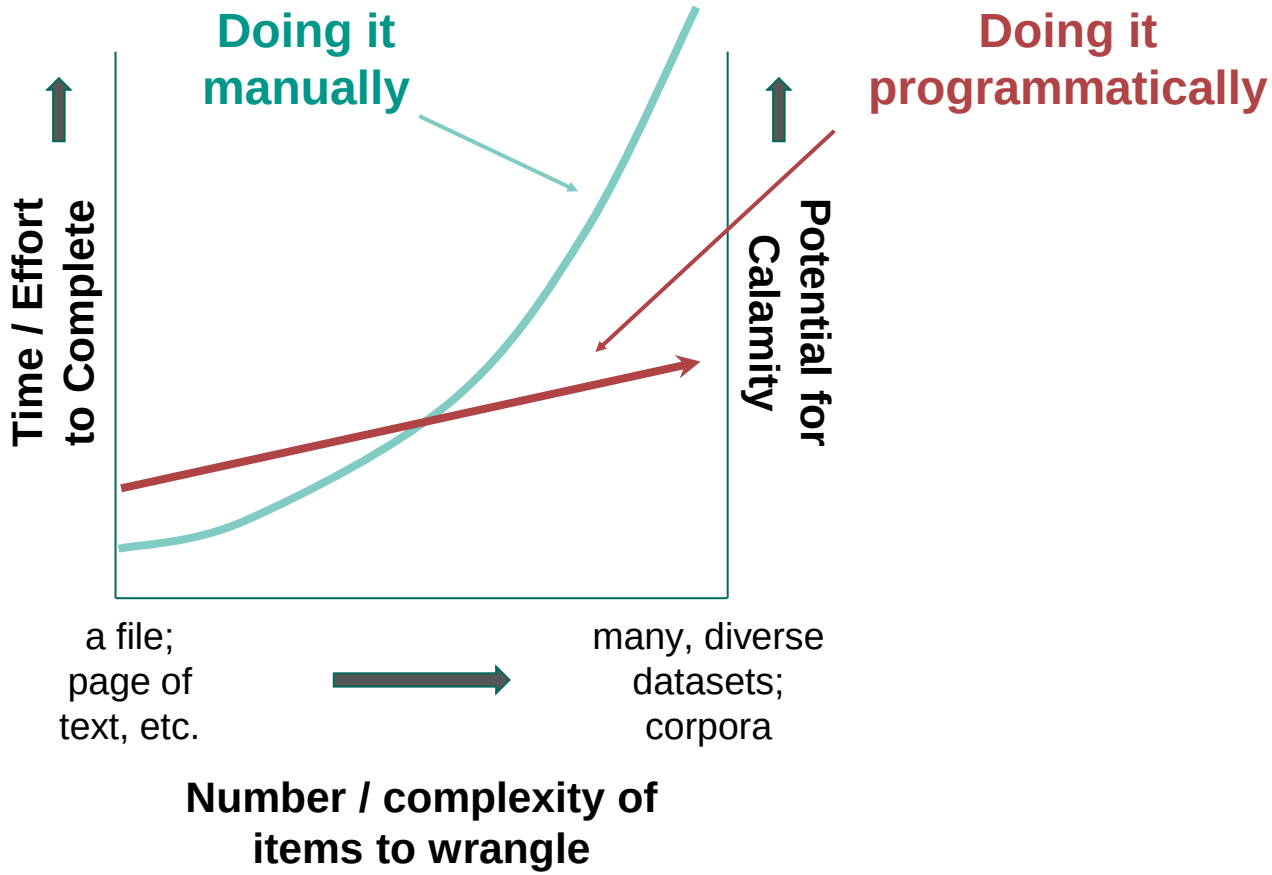
So, when to let the computer take over?



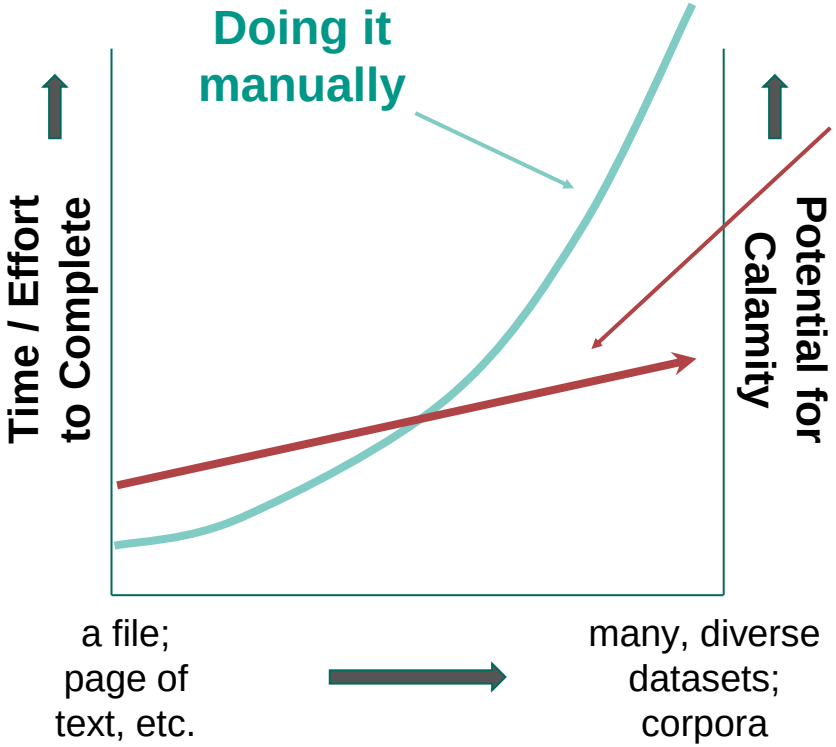
Spreadsheets: The frenemy of research

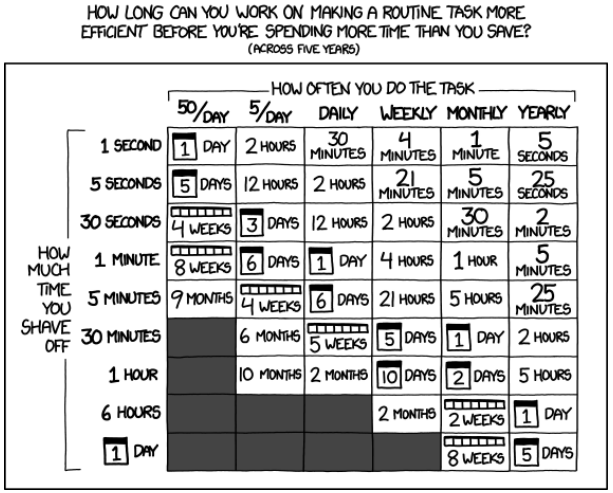


Spreadsheets: The frenemy of research



Spreadsheets: The frenemy of research





Relevant xkcd:

https://xkcd.com/1205/

Reasons to code some/all of your approach

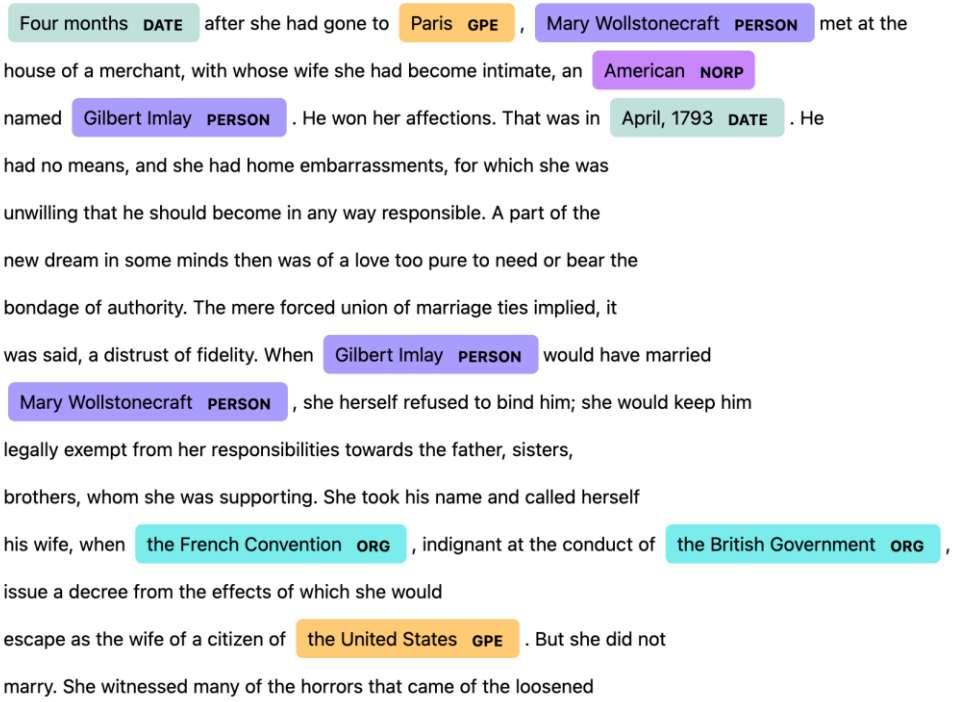
- To save you time
- To scale your approaches
- To reduce analytical toil
- Because (some people think) it is fun
- To build your own ‘toolkit’ of analytical scripts, functions, modules
- To enhance tractability, transparency, reproducibility, and reuse

To our Jupyter Notebook

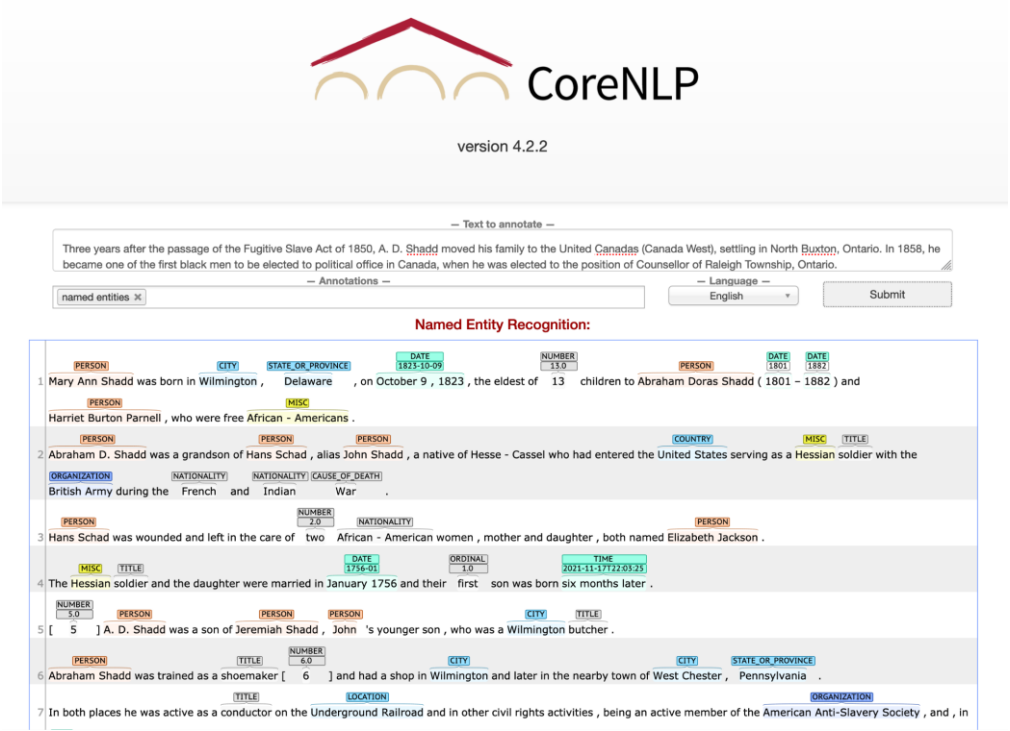
Open the file `CTA-Bootcamp-2024-python-prep.ipynb`.
Make a copy of the file (if you haven't already)
Follow along with Jay’s instructions

Named Entity Recognition

Analyzing Texts with Named Entity Recognition



Named Entity Recognition (NER) in Practice



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

+ Code + Text

```
[ ] # 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()
```

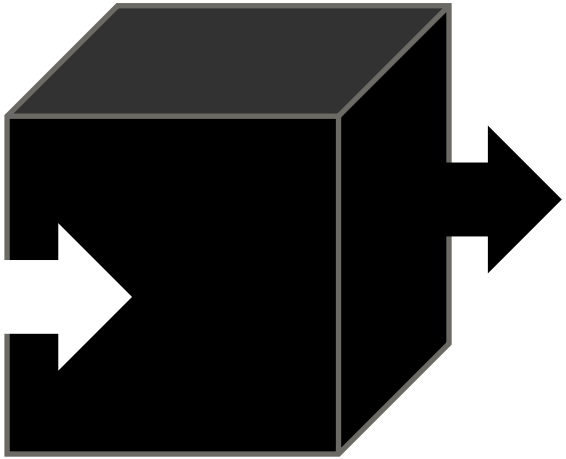
```
[ ] # Instantiate NLP pipeline - load transformer corpus
    nlp = spacy.load('en_core_web_trf')

    # 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_))
```

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."



"bell hooks"

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/cta-bootcamp

- Open the file **CTA-Bootcamp-2024-SA.ipynb** and save a copy to your Google Drive.

- Follow along with Jay

Stylometry

What is Stylometry?

"[T]he quantitative study of literary style through computational distant reading methods... based on the observation that authors tend to write in relatively consistent, recognizable and unique ways."

Francois Dominic Laramée, *Introduction to stylometry with Python* (Programming Historian)

The John Burrows' Delta Method

- One of the most widely used stylometric methods
- Basically:
  - Corpus has ***x*** number of authors
  - Corpus has ***n*** most frequently used words as features
  - Calculate share of each of the author's use of each ***n***
  - Calculate the mean and the standard deviation of each ***n*** across all ***x*** values
  - Use the calculated mean and SD for the feature over the corpus
  - Calculate a z-score for each of the ***n*** features and ***x*** sub-corpora
  - Calculate the same z-scores for each feature in the text of unknown provenance
  - Calculate a delta score comparing the anonymous text with each author's sub-corpus

Strongest candidate is the author for whom the delta score between the author's sub-corpus and the anonymous text is the lowest.

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

```
[ ]  1 # Install the Fast Stylometry library
     2 !pip install faststylometry
```

```
[ ]  1 # Import the required internal Python libraries
     2
     3 # For corpus pre-processing, the Natural Language Toolkit (nltk)
     4 import nltk
     5 nltk.download("punkt")
     6
     7 # For working with data as data frames in Pandas
     8 from sklearn.decomposition import PCA
     9 import re
    10 import pandas as pd
```

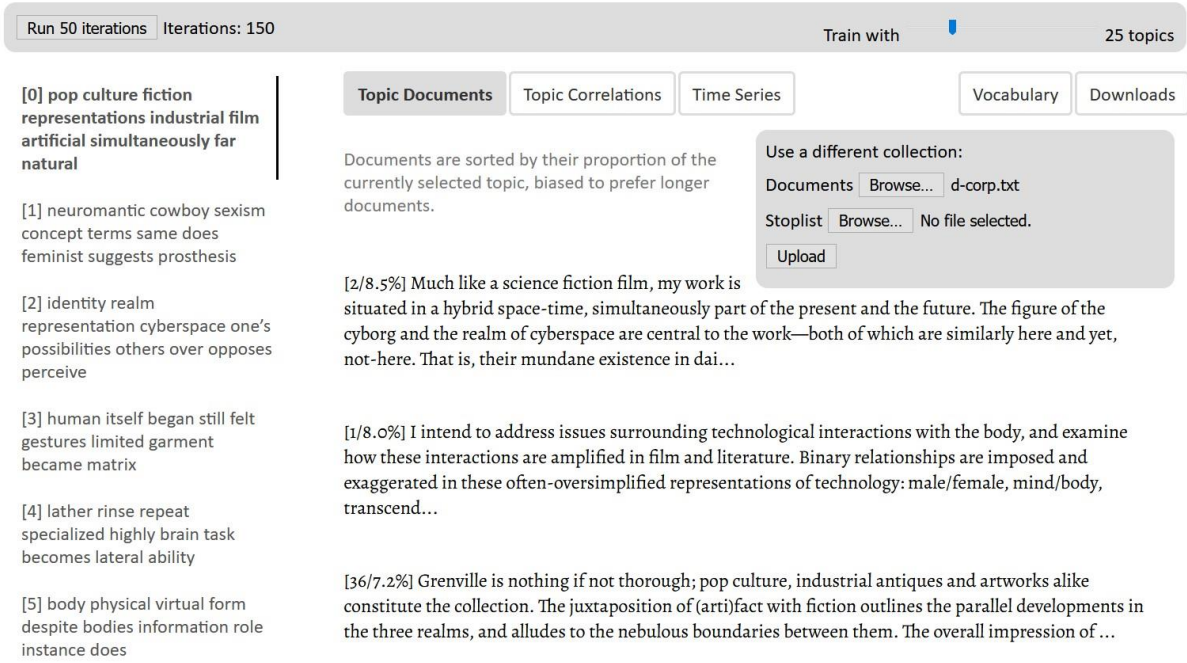
```
[ ]  1 # Import the Fast Stylometry components required for the script
     2 from faststylometry import Corpus
     3 from faststylometry import load_corpus_from_folder
     4 from faststylometry import tokenise_remove_pronouns_en
     5 from faststylometry import calculate_burrows_delta
     6 from faststylometry import predict_proba, calibrate, get_calibration_curve
```

```
[ ]  1 # Load the corpus and do minimal pre-processing: tokenise and remove pronous
     2 train_corpus = load_corpus_from_folder("data/train")
     3 train_corpus.tokenise(tokenise_remove_pronouns_en)
```

```
[ ]  1 # Set pattern to a string value to load a subset of the corpus if there is more than one text with an unknown author
     2 test_corpus = load_corpus_from_folder("data/test", pattern=None)
     3 test_corpus.tokenise(tokenise_remove_pronouns_en)
```

Topic Modeling

Discerning Corpus "Topics" with Topic Modeling



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

+ Code + Text

```
[ ] # 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())
```

Day 1 wrap-up