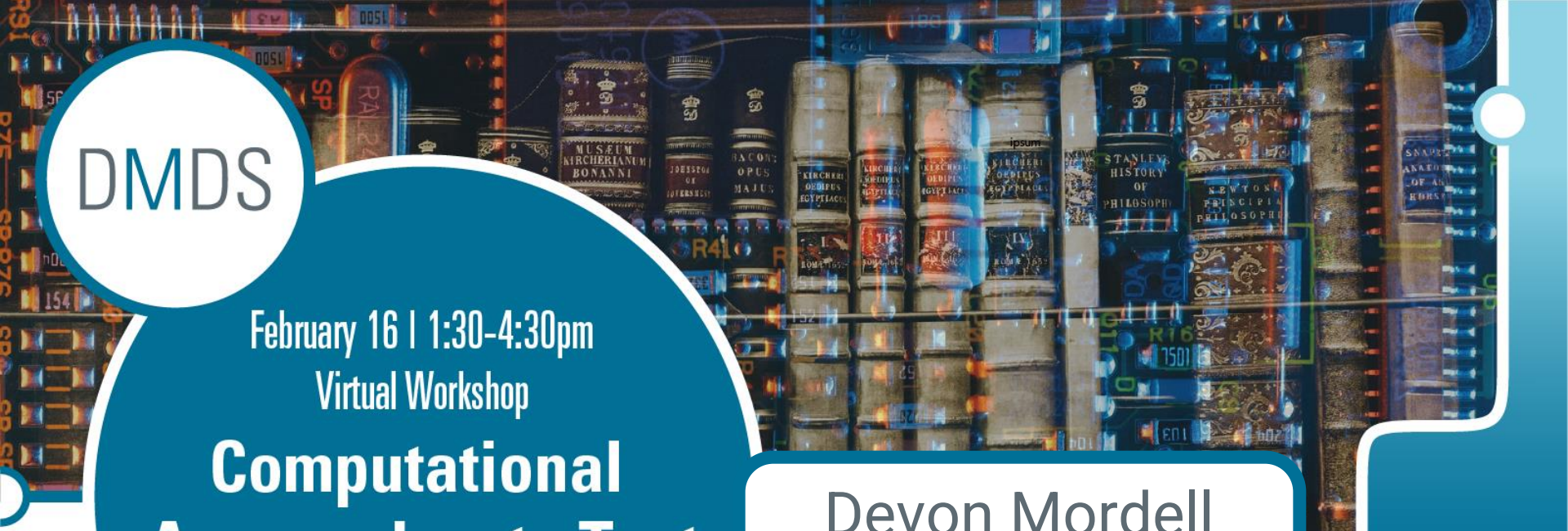


DMDS

February 16 | 1:30-4:30pm

Virtual Workshop

Computational Approaches to Text Prep & Analysis

u.mcmaster.ca/scds-events

Devon Mordell

Jay Brodeur

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
Text prep with OpenRefine	40 minutes	Introduction to OpenRefine Manual cleanup (e.g. find and replace) Faceting
Getting Programmatic with Python	20 minutes	Overview of programmatic approaches The 'what' and 'when' to program Using Python for text preparation
--Break--	**10 minutes**	**--Break--**
Hands-on sampling of text analysis methods	75 minutes	Named entity recognition Topic modeling Sentiment analysis
Q & A; Final thoughts (lecture + discussion)	10 minutes	Questions & final thoughts Where to learn more

Workshop landing page: scds.github.io/dmds-22-23/ComputationalText.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.
- 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.
- Apply prepared computational techniques to perform common text preparation steps and introductory analyses.

Workshop landing page: scds.github.io/dmds-22-23/ComputationalText.html

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

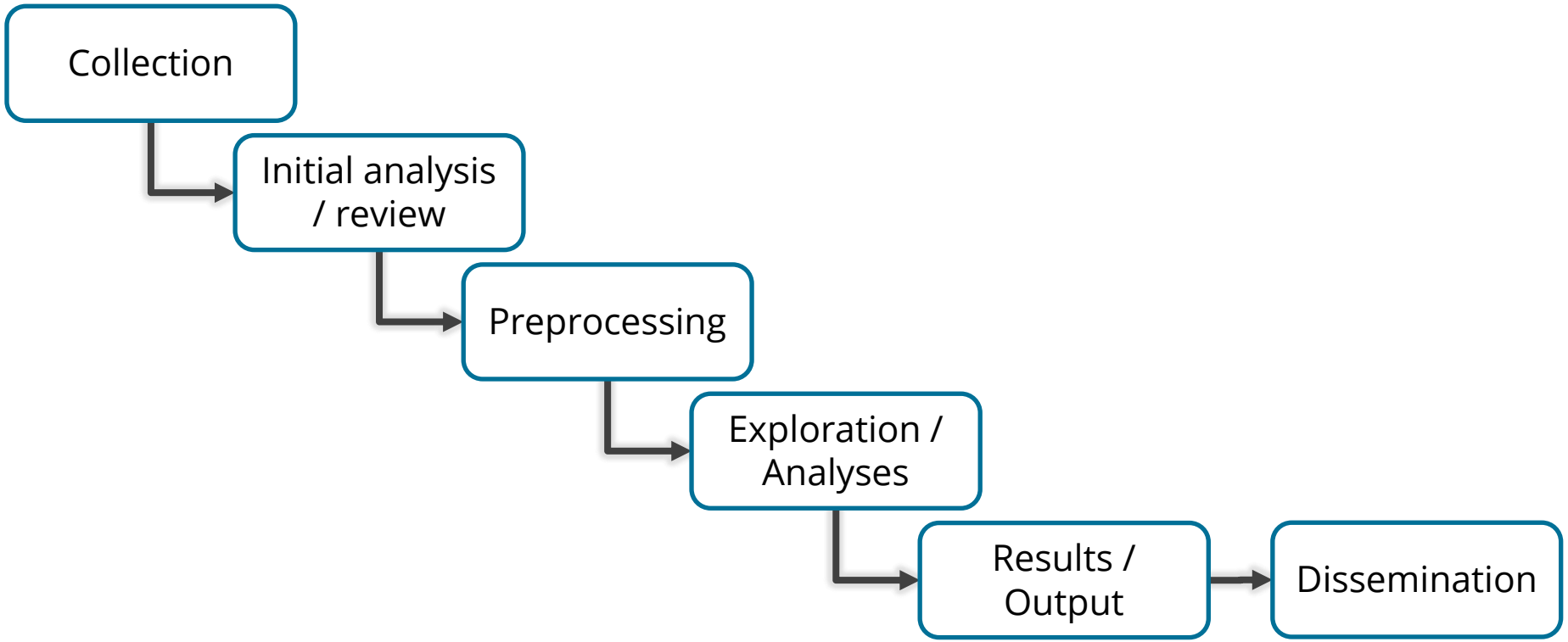
Morphological analysis
Stemming, Lemmatization, Part of Speech Tagging

Relational semantics
Relationship extraction, Semantic parsing

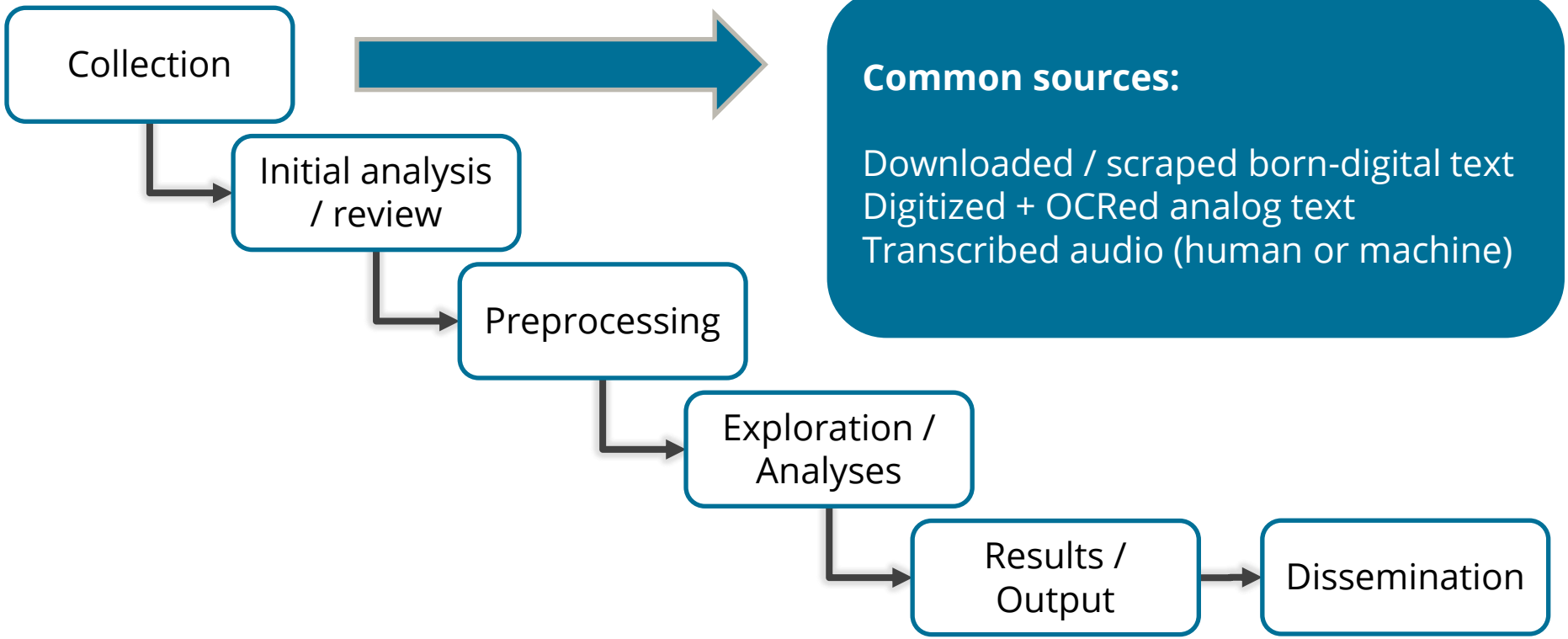
Syntactic analysis
Parsing, Sentence breaking

Discourse semantics
Discourse analysis, Topic segmentation, Argument mining

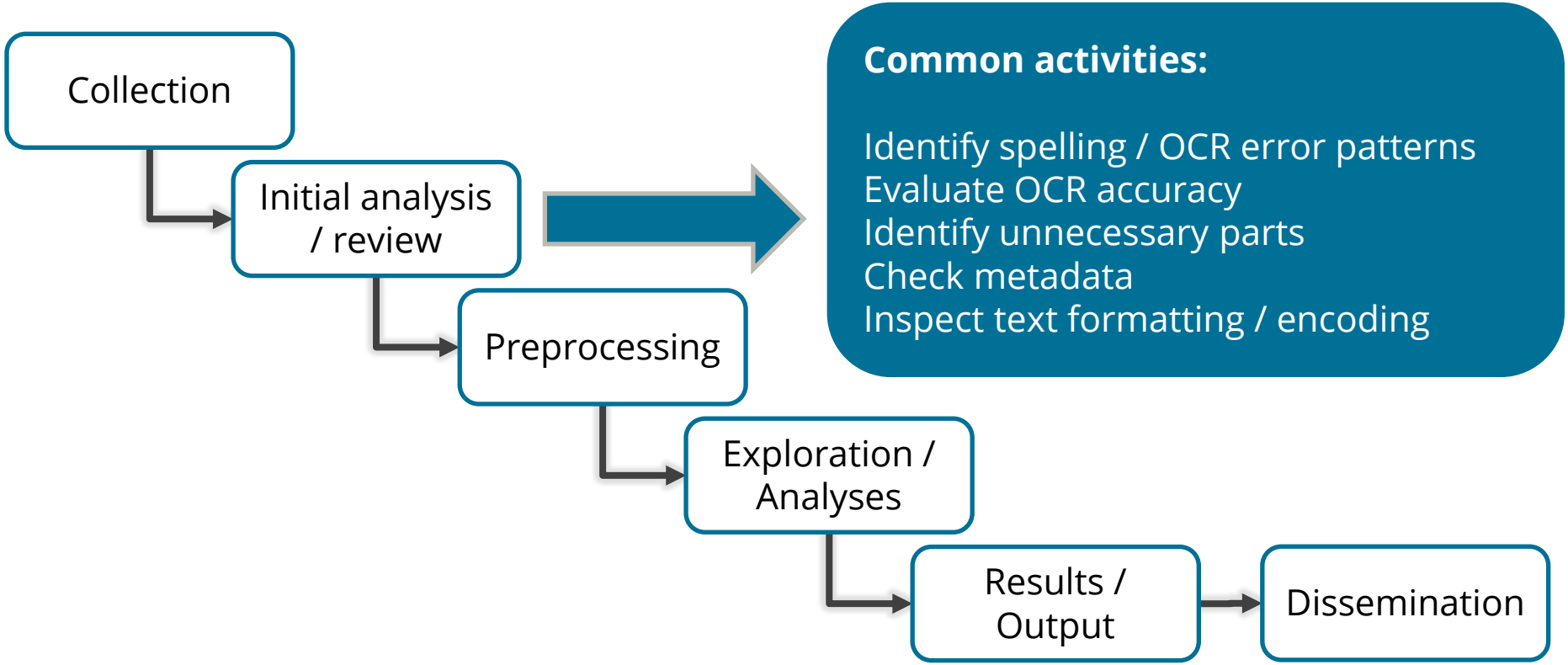
NLP Workflows



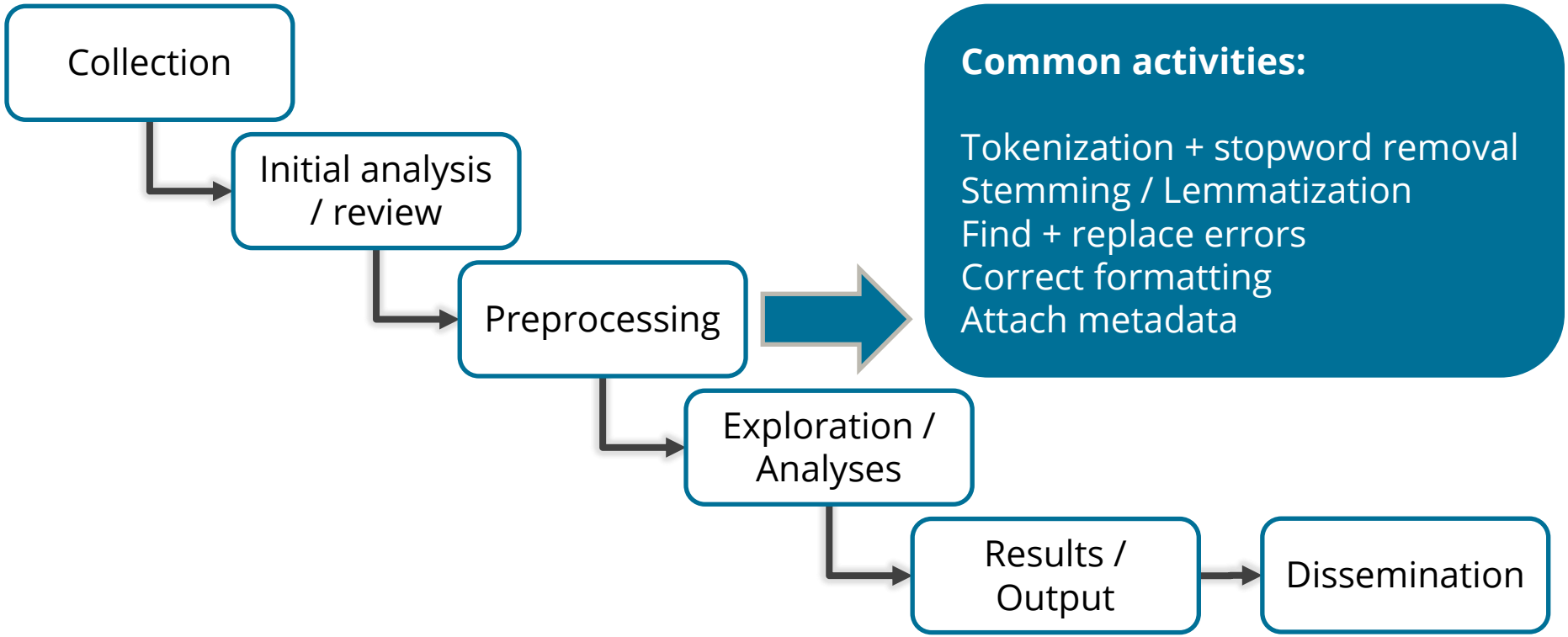
NLP Workflows



NLP Workflows

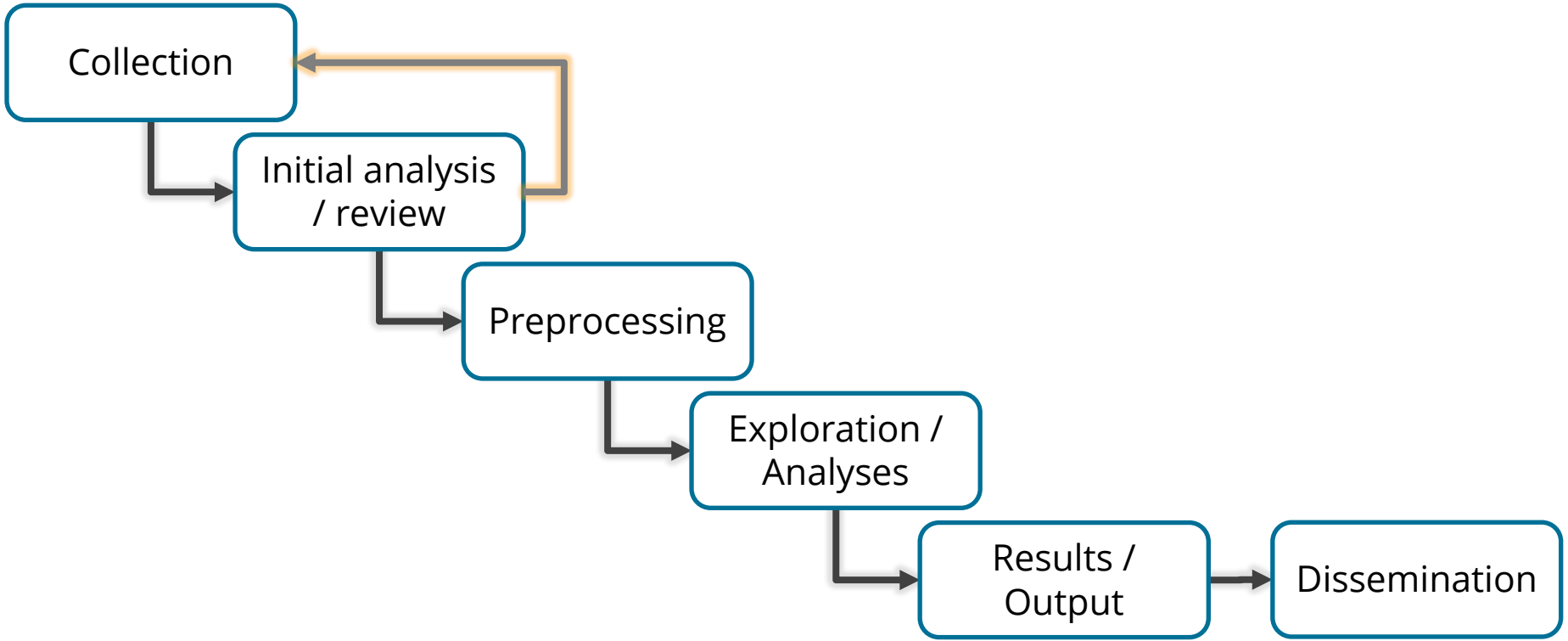


NLP Workflows



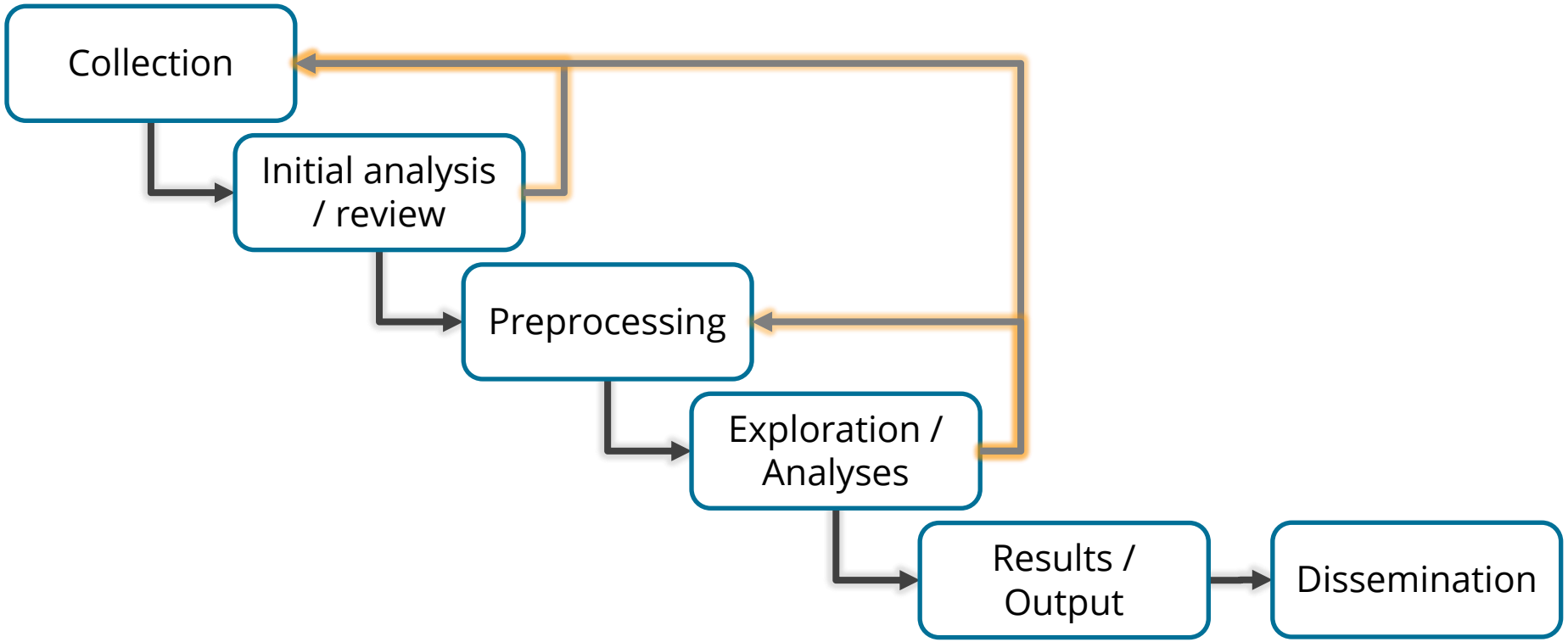
NLP Workflows

… are iterative



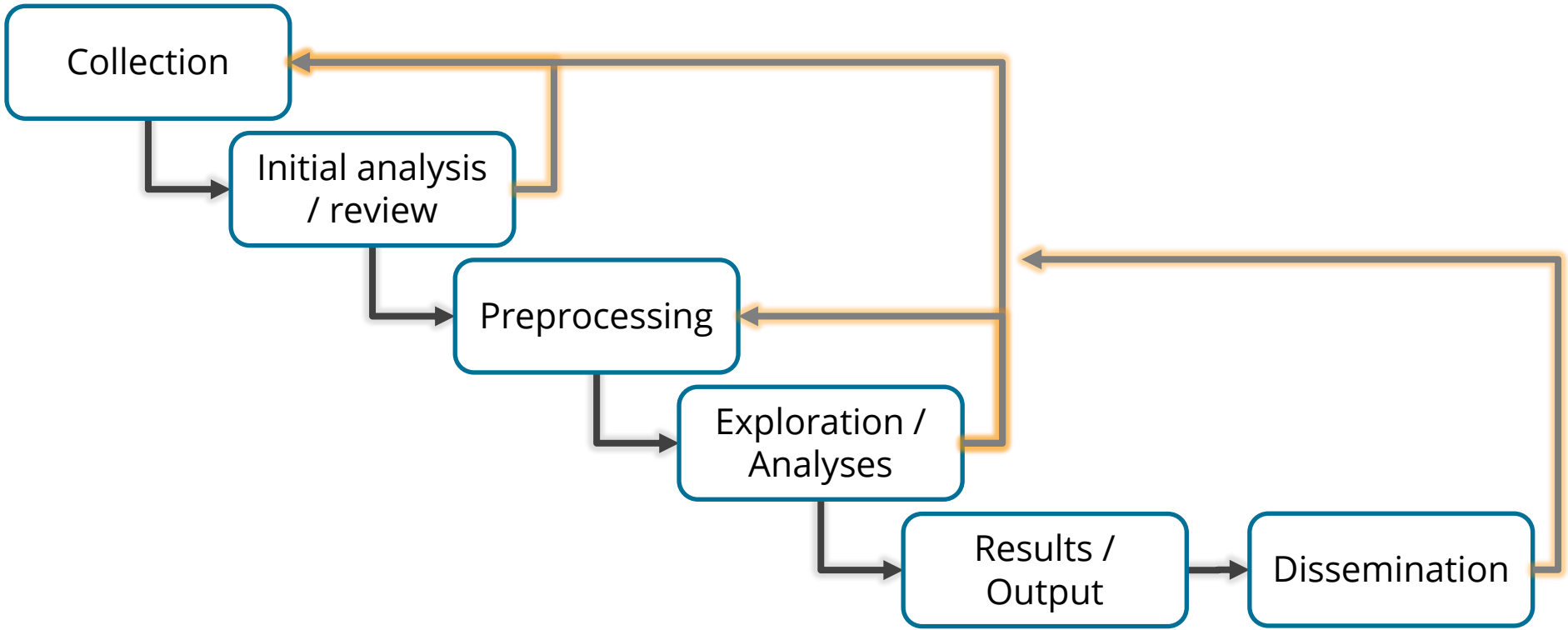
NLP Workflows

… are iterative

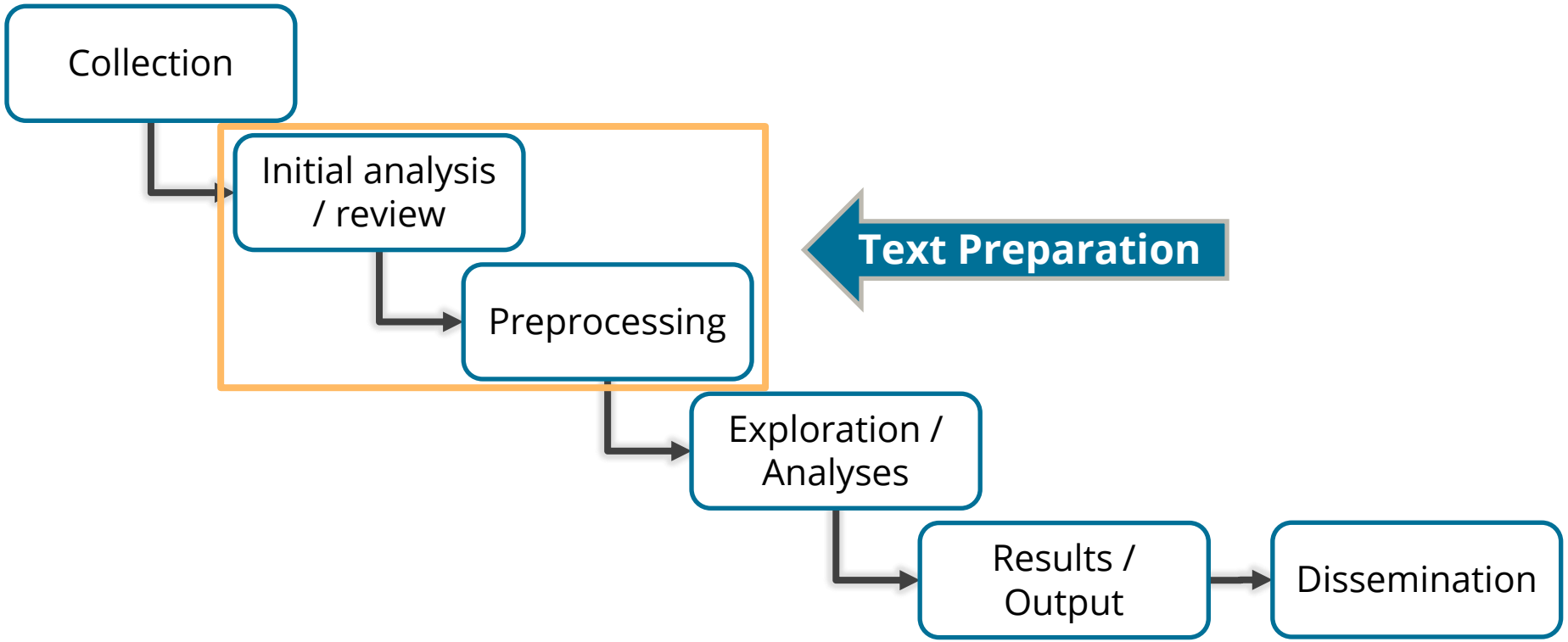


NLP Workflows

… are iterative



NLP Workflows



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

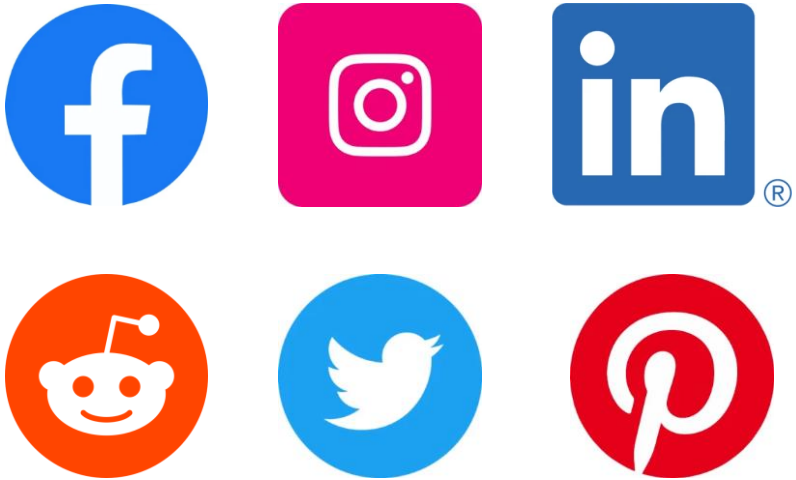
Common Transcription Issues

```
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
- contain non-textual data like markup and embedded scripts

Some text prep 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?

Hands-on text prep with OpenRefine

OpenRefine – for text preparation???

- Graphical interface (GUI)
- Non-destructive editing
- Self-documenting
- Reproducibility of steps

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.

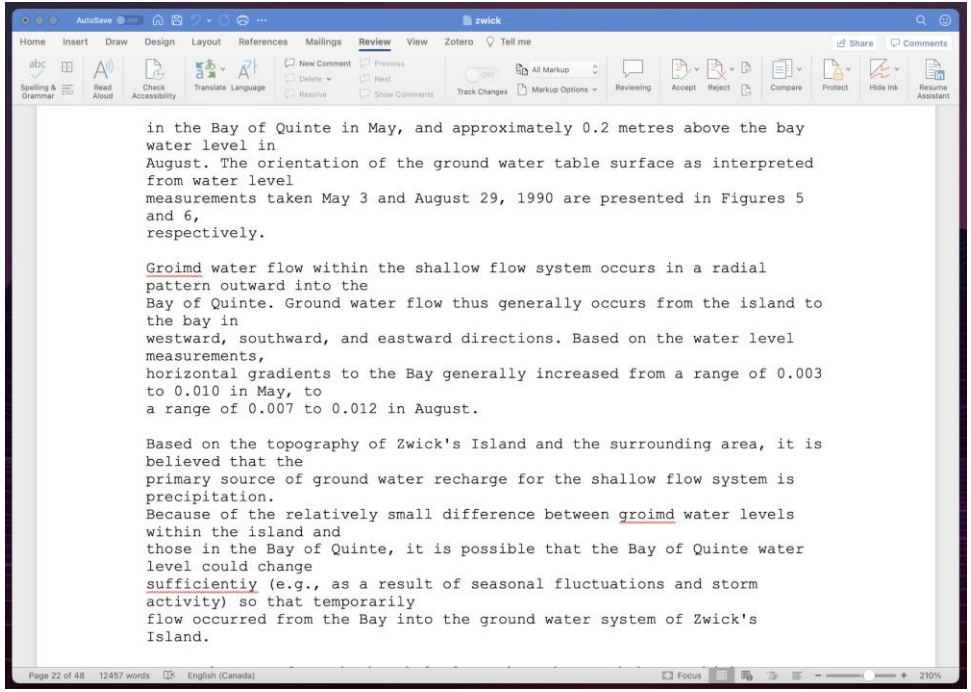
During the site visit surface water drainage and pathways were observed and noted. Conductivity measurements of surface water were taken at several locations around the site along with observations of iron staining and vegetation loss in roadside and drainage ditches in the northeast corner of the Island. 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

undenake

concepmal

smdy

smdy

smdy

backgroimd

acmal

coUect

Mimicipal

fmalizing

Dtiring

aroimd

stammg

landfiUed

moimted

Figiu

Stratigraphie

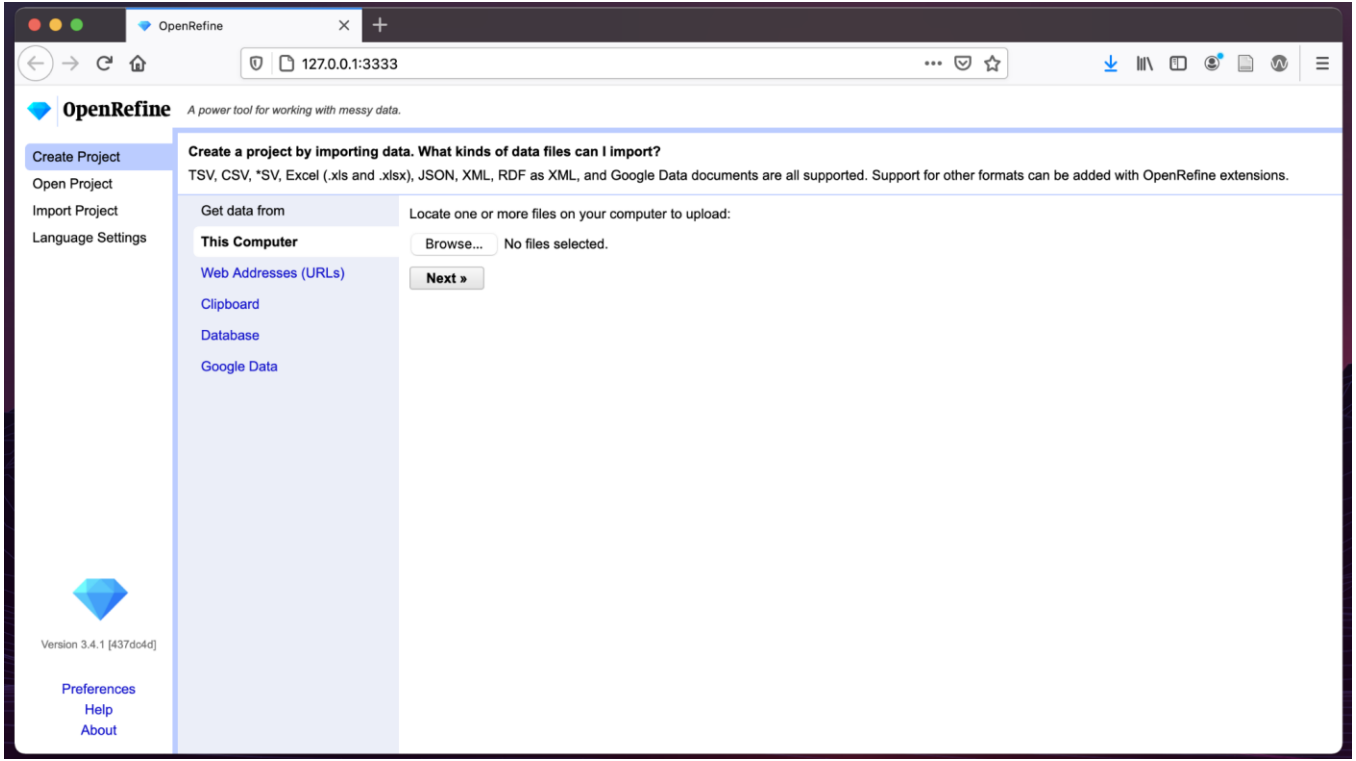
concurrenLly

Ruoride

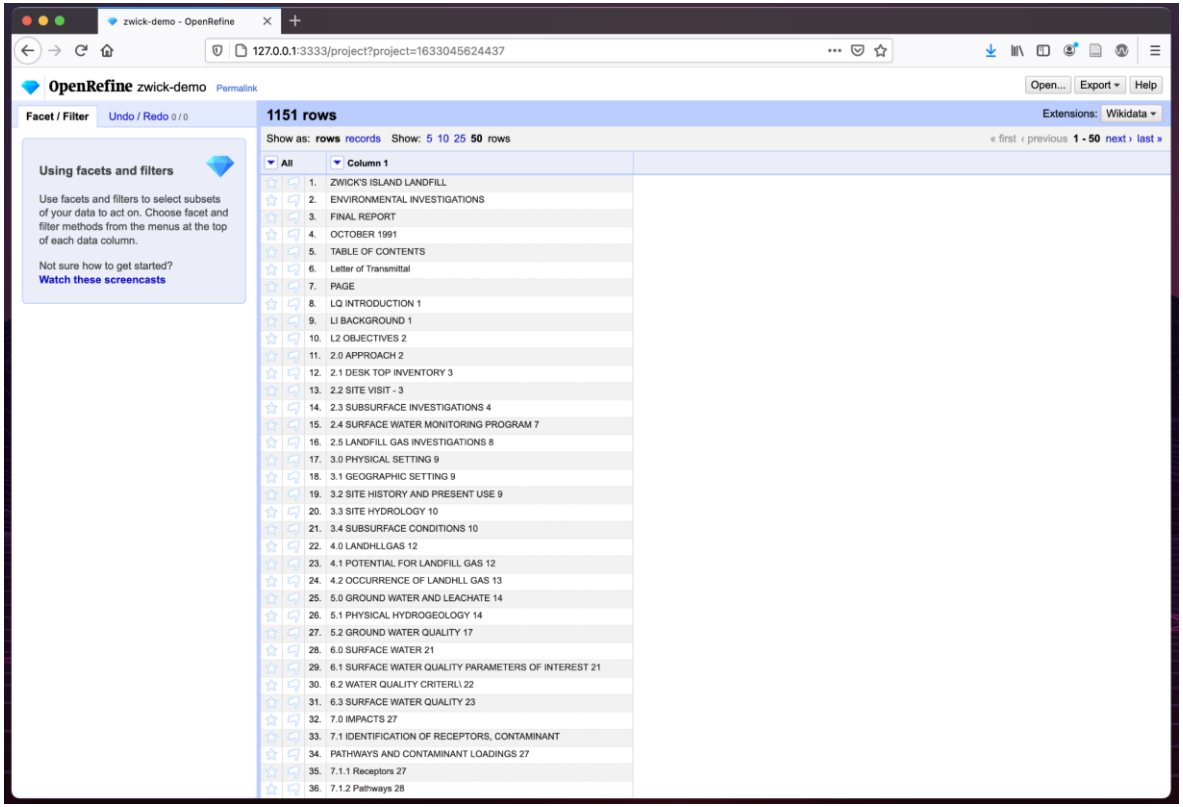
Qiloride

Organo

Open OpenRefine

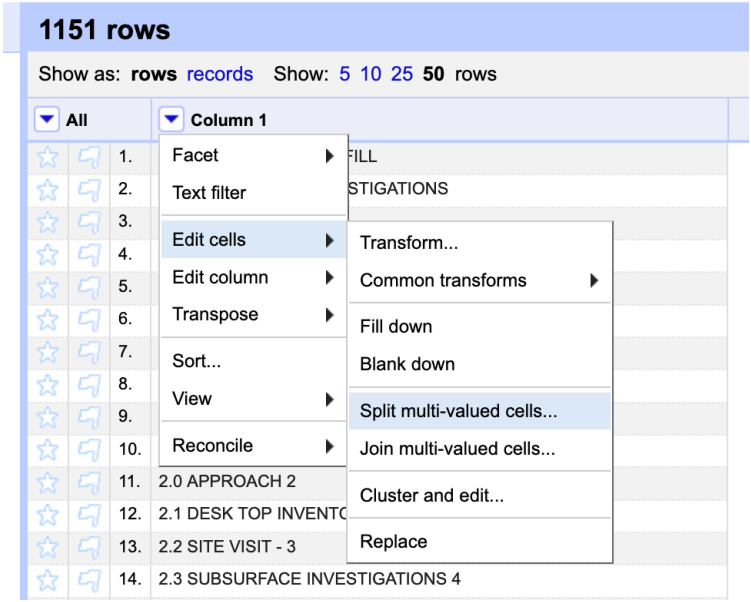


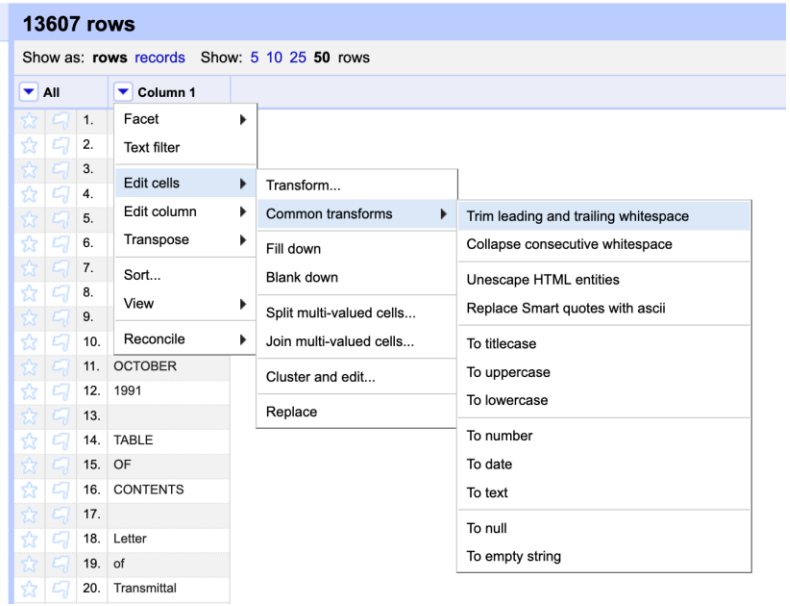
Initial Data Analysis: in OpenRefine



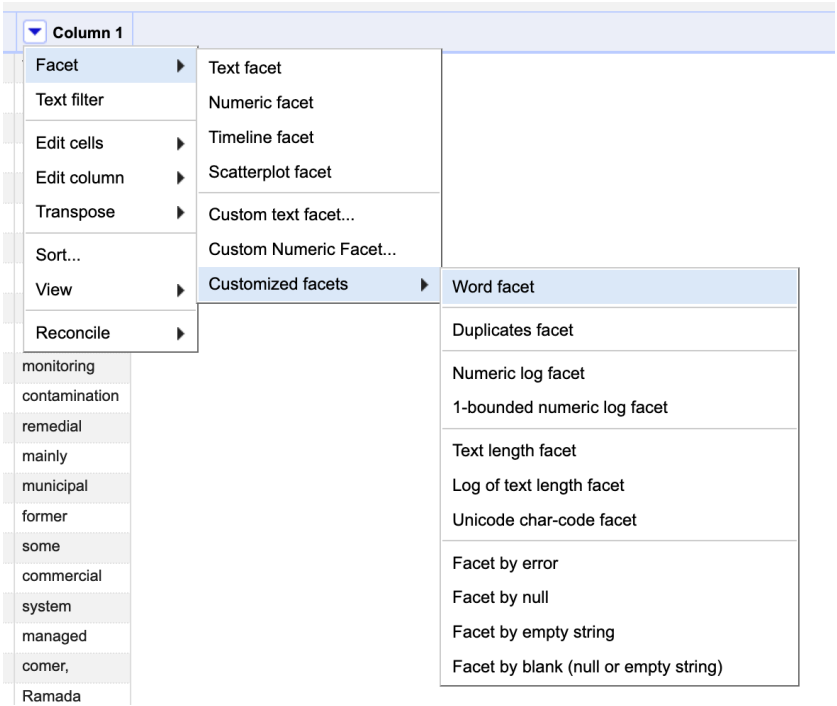
Prepare Dataset for Text Analysis

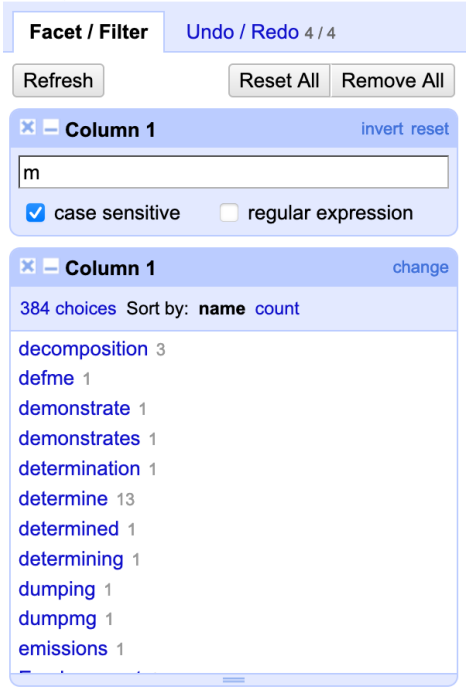
Tokenize, trim and remove blank rows



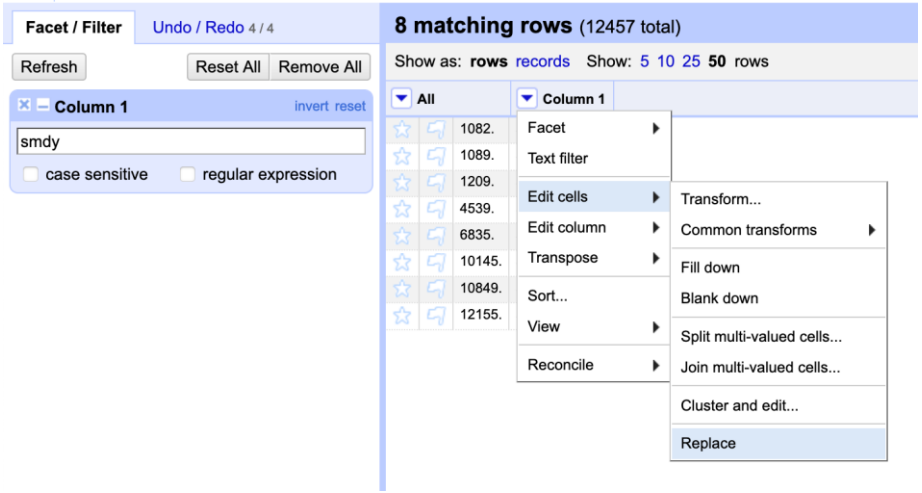


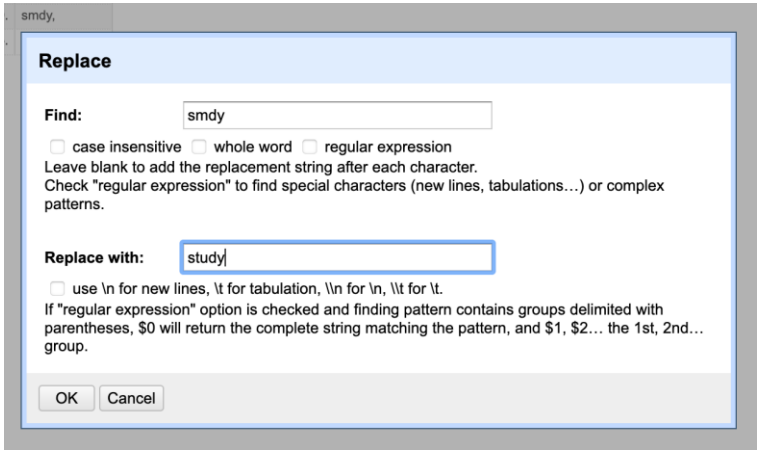
Filter and Facet to Find Errors





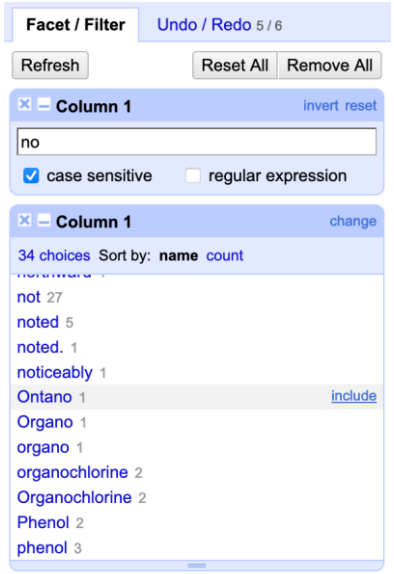
Find and Replace with Text Filter

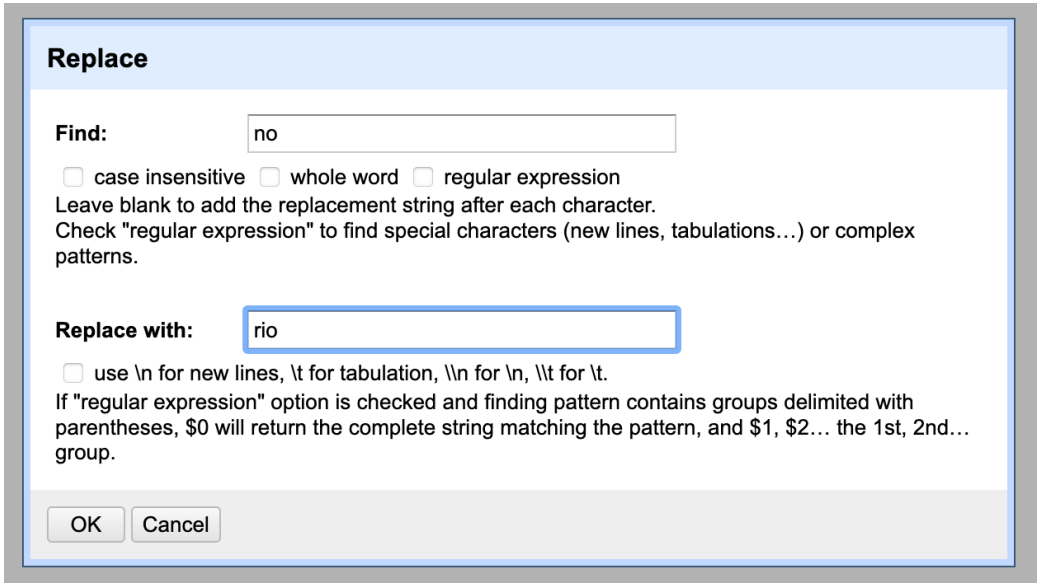




Find and Replace with Text Filter

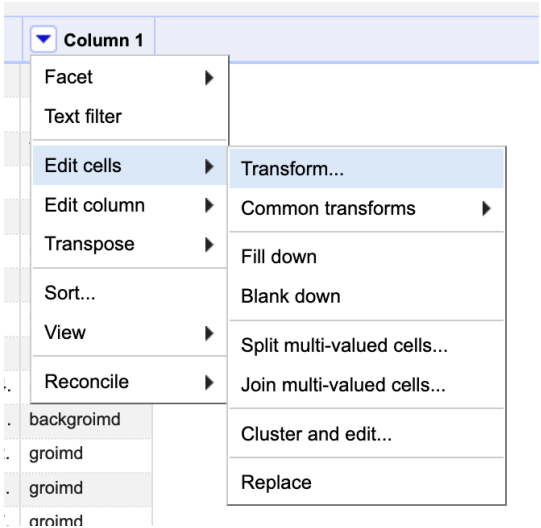
Filter by “no” as case sensitive, then word facet & include

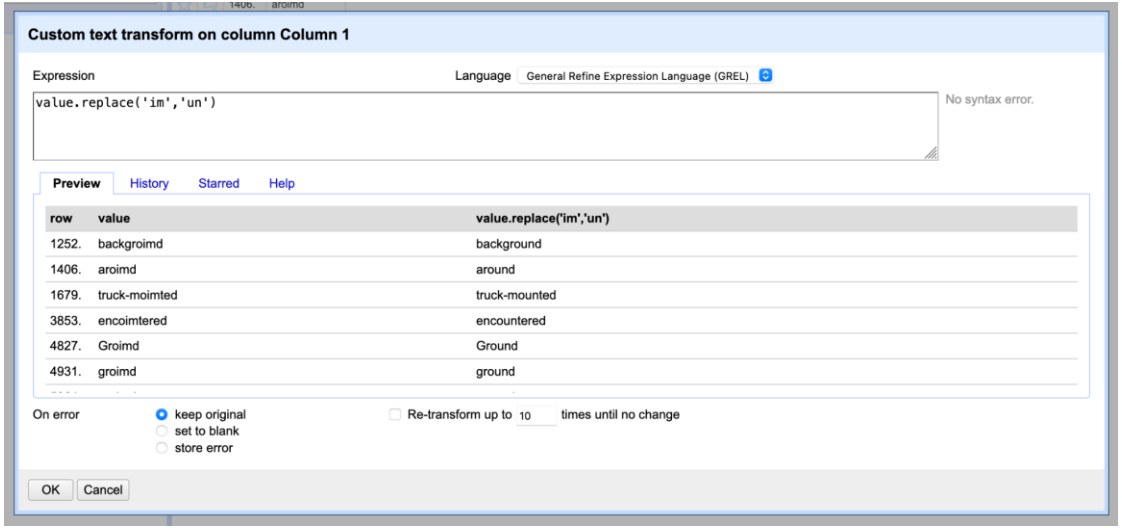




Find and Replace with GREL

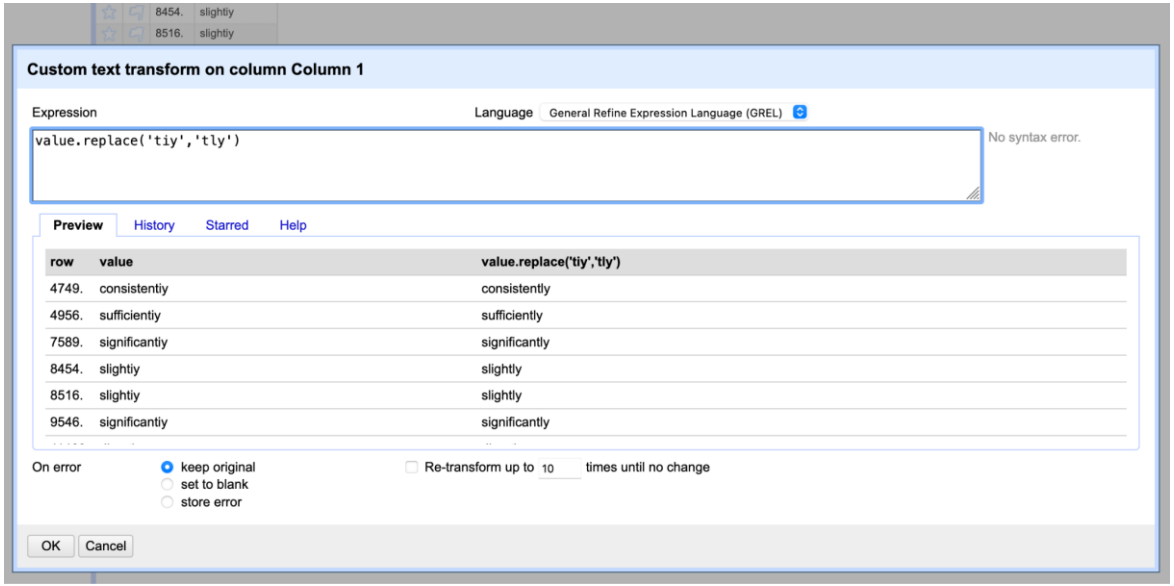
Filter by “oim” (oun) → value.replace(‘im’, ‘un’)





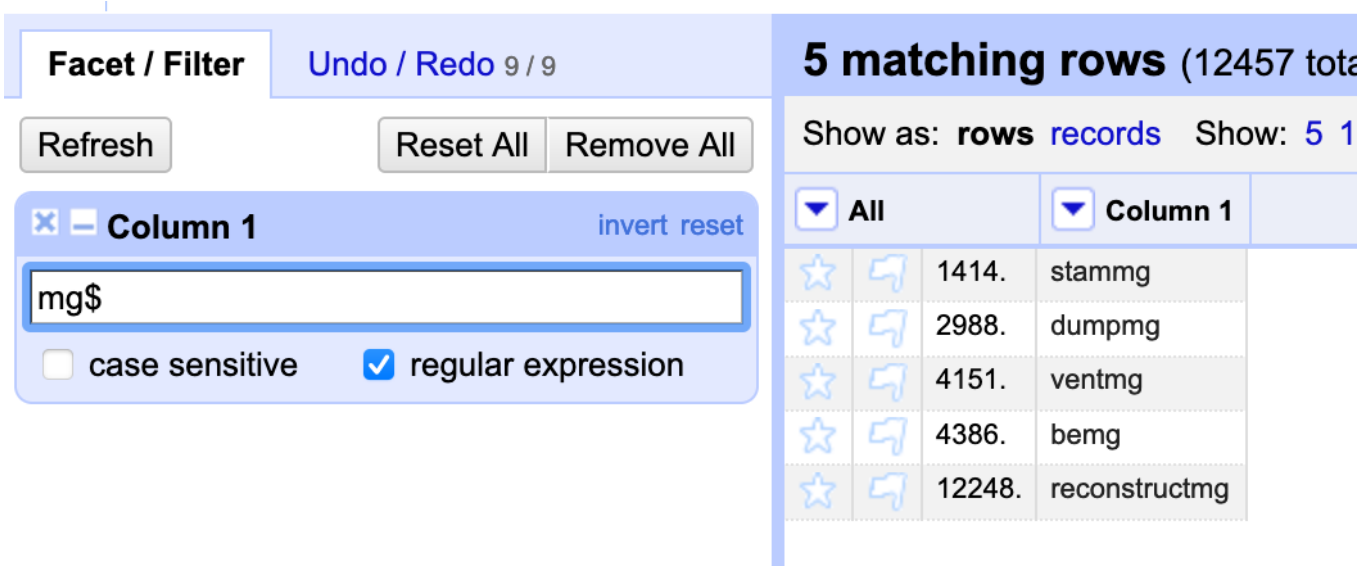
Find and Replace with GREL

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.

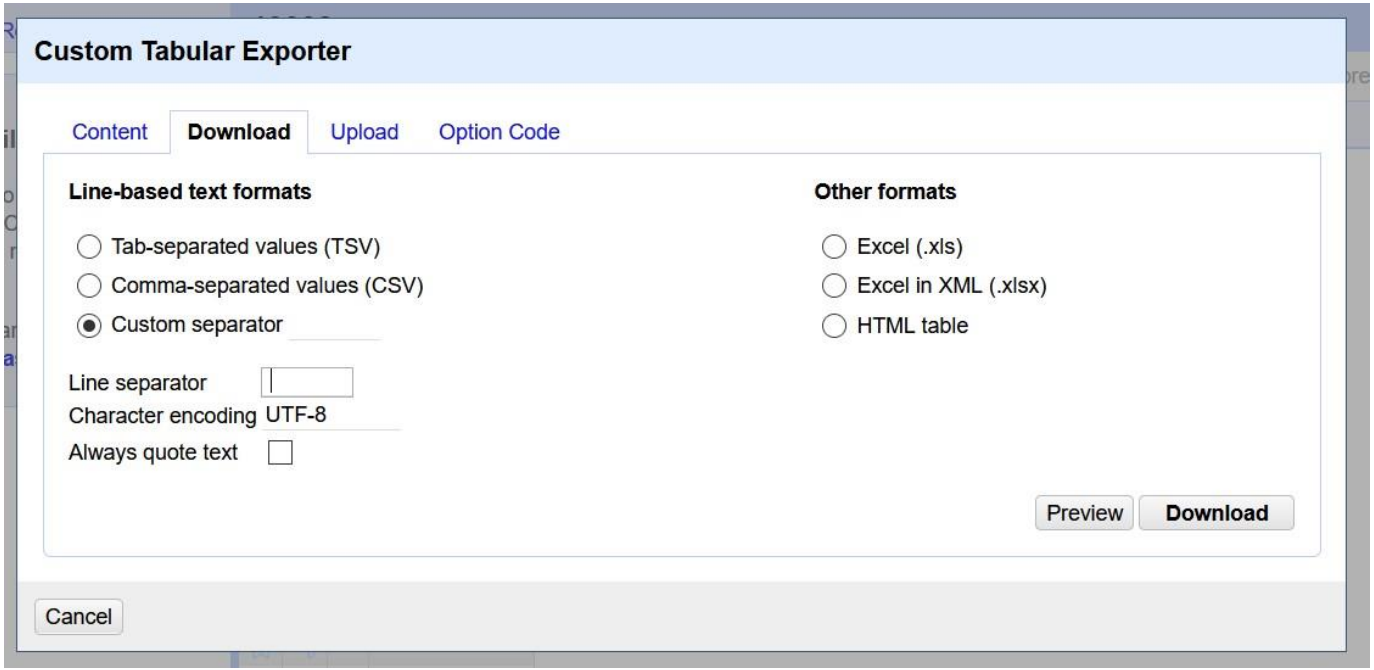
Try it out:

^mt → value.replace(‘m’ , ‘in’)

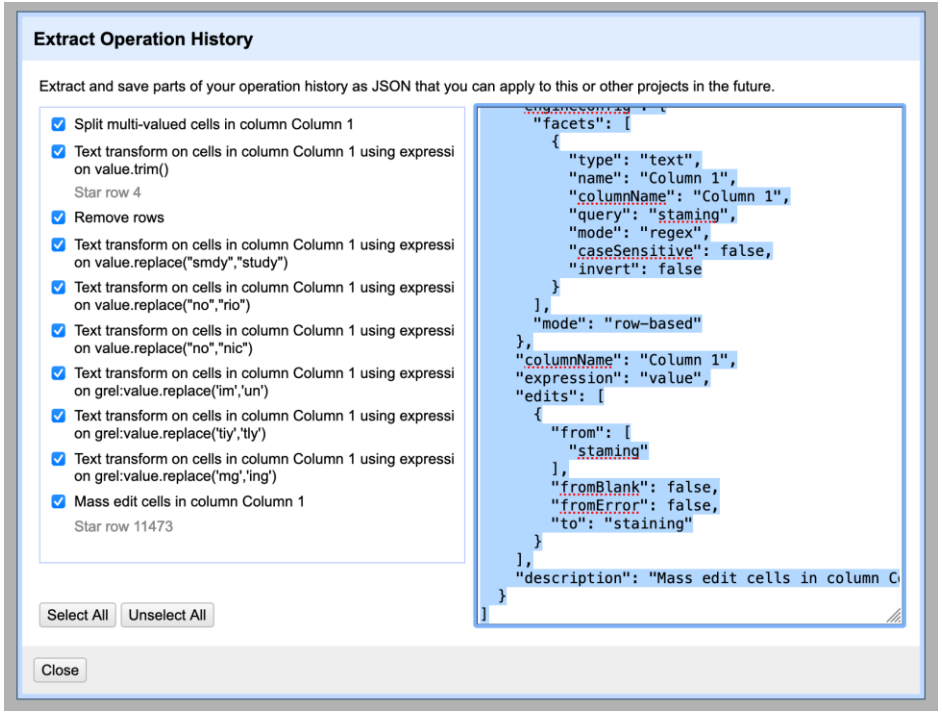
[a-z]U → value.replace(‘U’ , ‘ll’) [with case sensitive checked]

Others…?

Reconstitute your Document



Export your “Recipe” of Tasks



Break time!

Let’s take 10

Programmatic approaches with Python

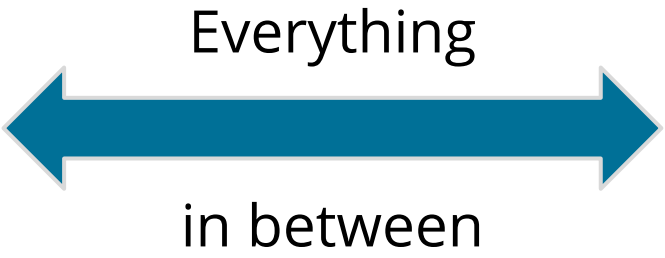
Text prep and analysis as a continuum of mediation

Completely manual

Completely automated

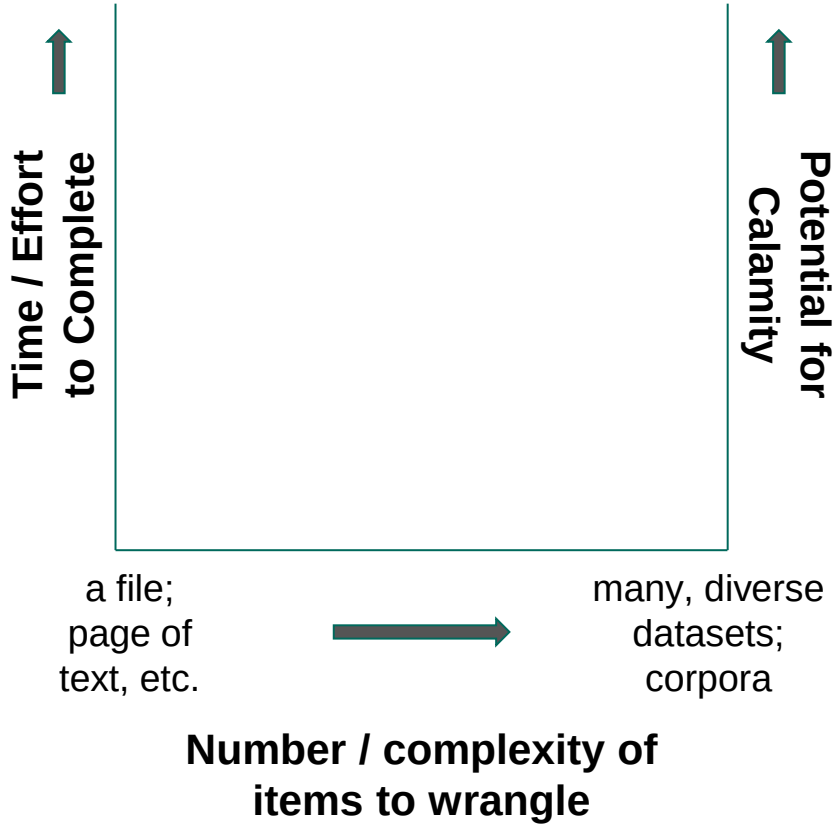
Text prep and analysis as a continuum of mediation

Completely manual

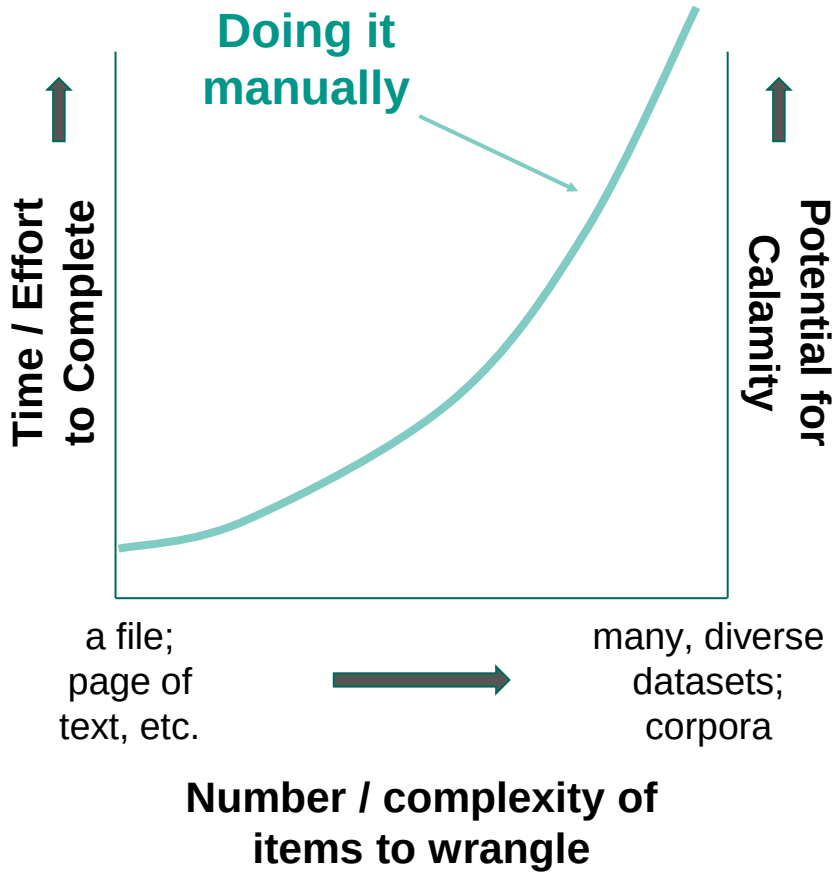


Completely automated

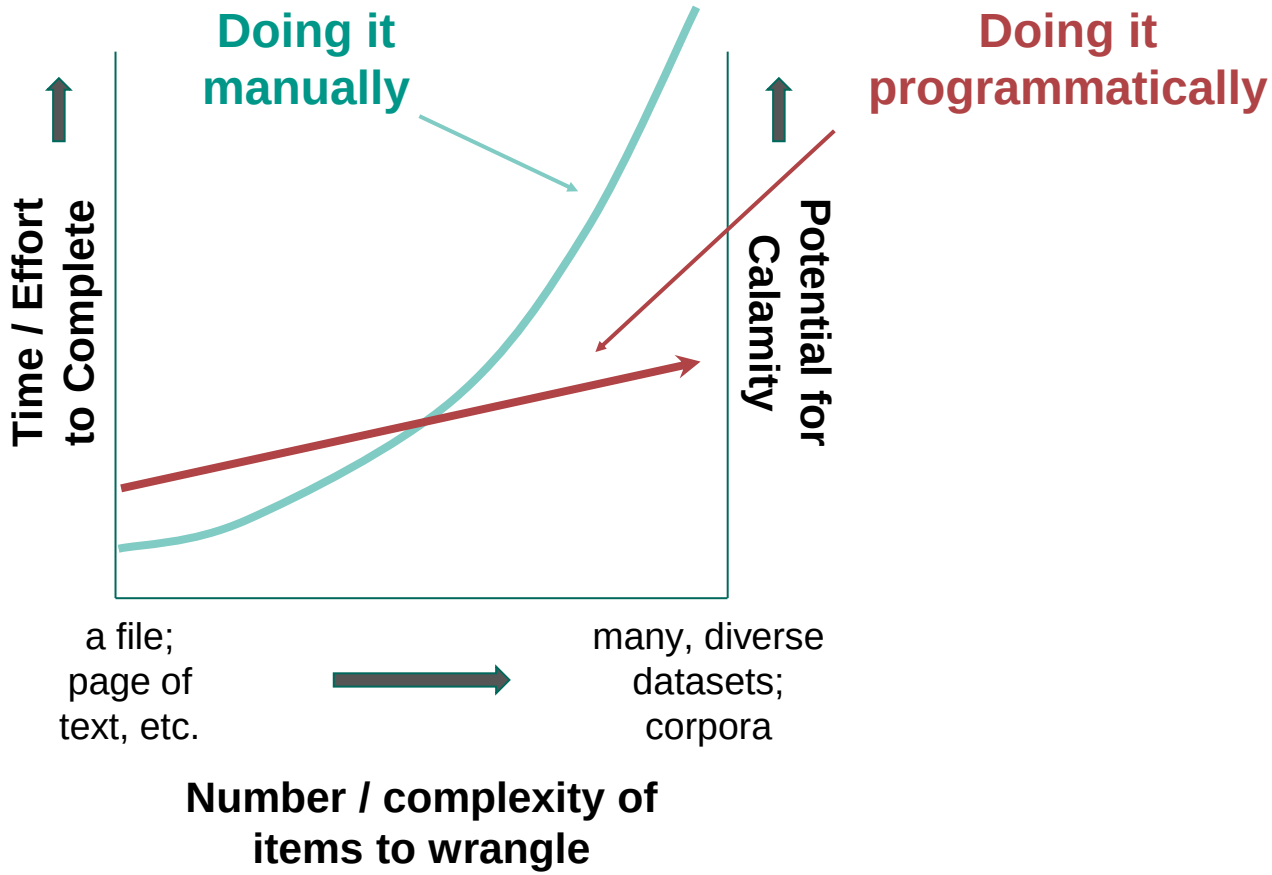
So, when to let the computer take over?



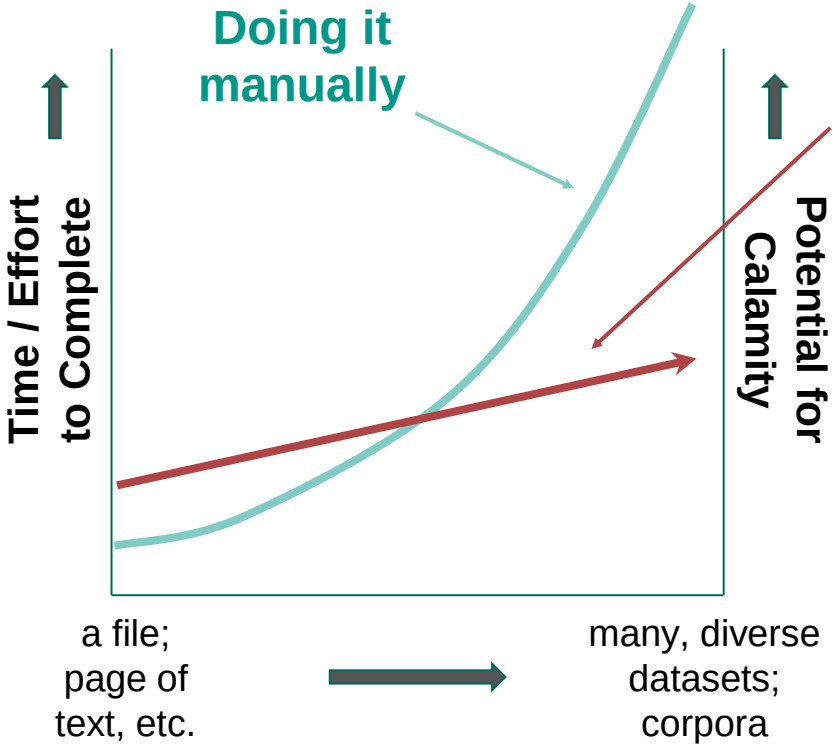
Spreadsheets: The frenemy of research

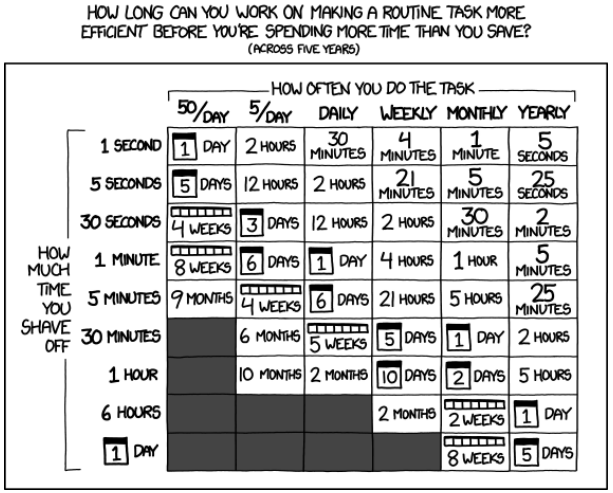


Spreadsheets: The frenemy of research



Spreadsheets: The frenemy of research





Relevant xkcd:
https://xkcd.com/1205/

Reasons to code some/all of your approach

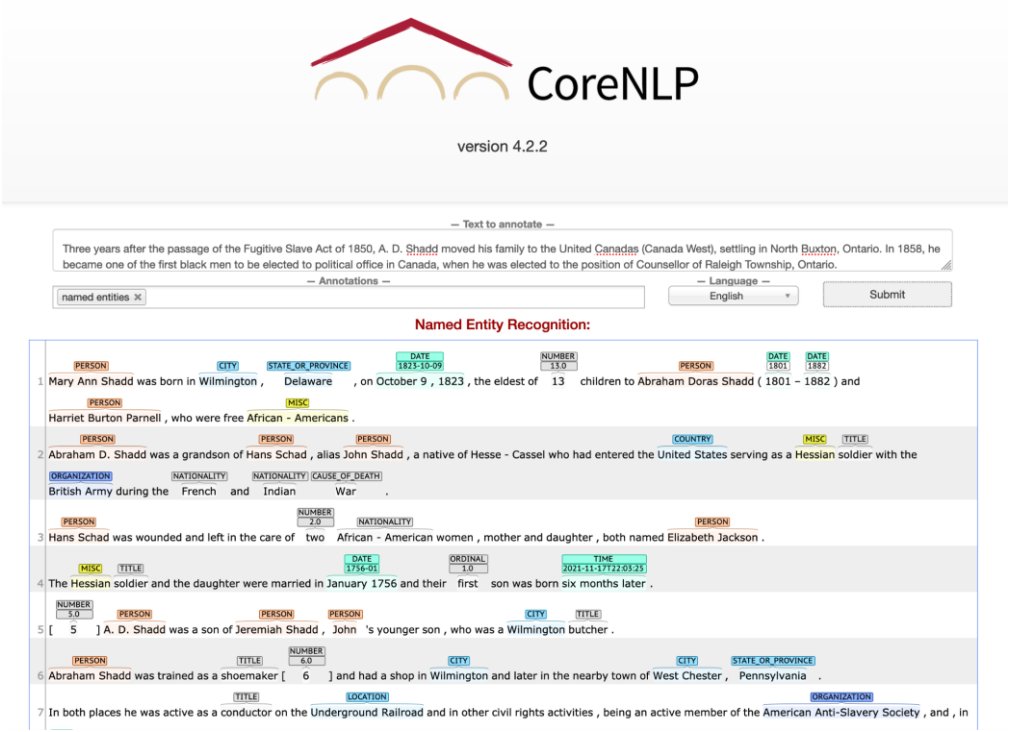
- To save you time
- To scale your approaches
- To reduce analytical toil
- Because (some people think) it is fun
- To build your own ‘toolkit’ of analytical scripts, functions, modules
- To enhance tractability, transparency, reproducibility, and reuse

To our Jupyter Notebook

Go to u.mcmaster.ca/dmds-text-prep and save a copy to your Google Drive.
Follow along with Jay's instructions

A hands-on sampler

How Named Entity Recognition (NER) Works



Try it out in Jupyter Notebooks...

+ 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()
```

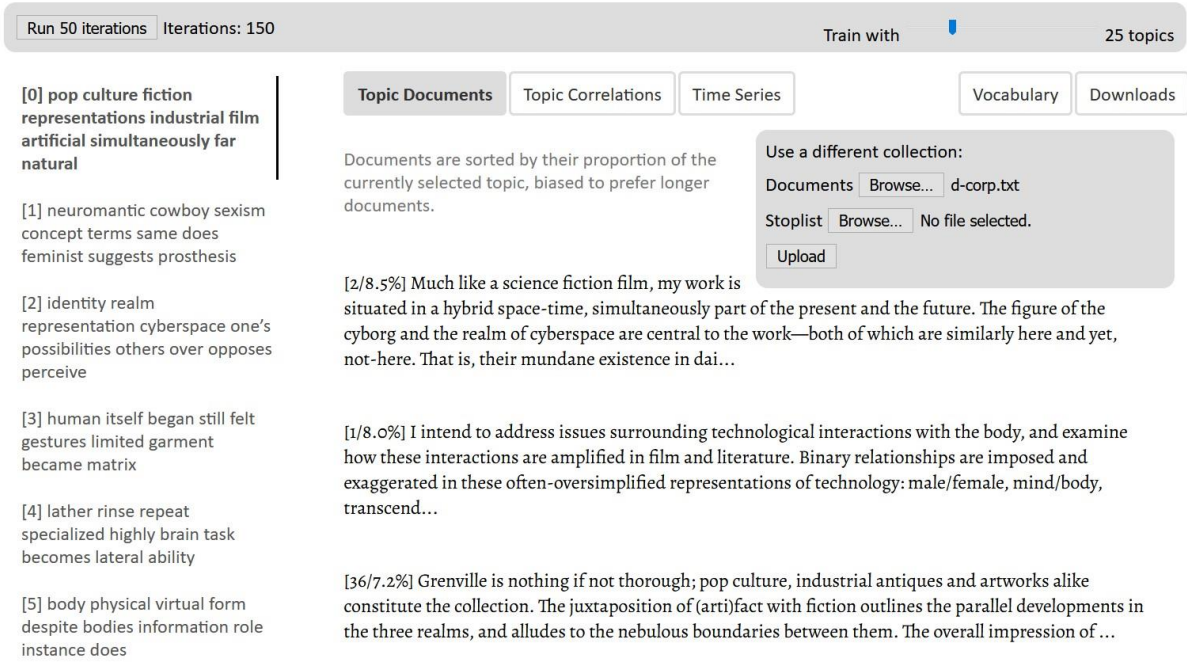
```
[ ]  # Instantiate NLP pipeline - load transformer corpus
     nlp = spacy.load('en_core_web_trf')

     # 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_))
```

Discerning Corpus "Topics" with Topic Modeling



Try it out in Jupyter Notebooks...

+ 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())
```

Sentiment Analysis – Jay

Go to u.mcmaster.ca/dmds-sentimental and save a copy to your Google Drive.

Follow along with Jay's instructions

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**