

Computational Text Analyses Bootcamp

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

In-Person Workshop

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:Waterdown_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?

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

Morphological analysis

Stemming, Lemmatization,
Part of Speech Tagging

Syntactic analysis

Parsing, Sentence breaking

Lexical semantics

Named entity recognition, Sentiment analysis, word sense disambiguation

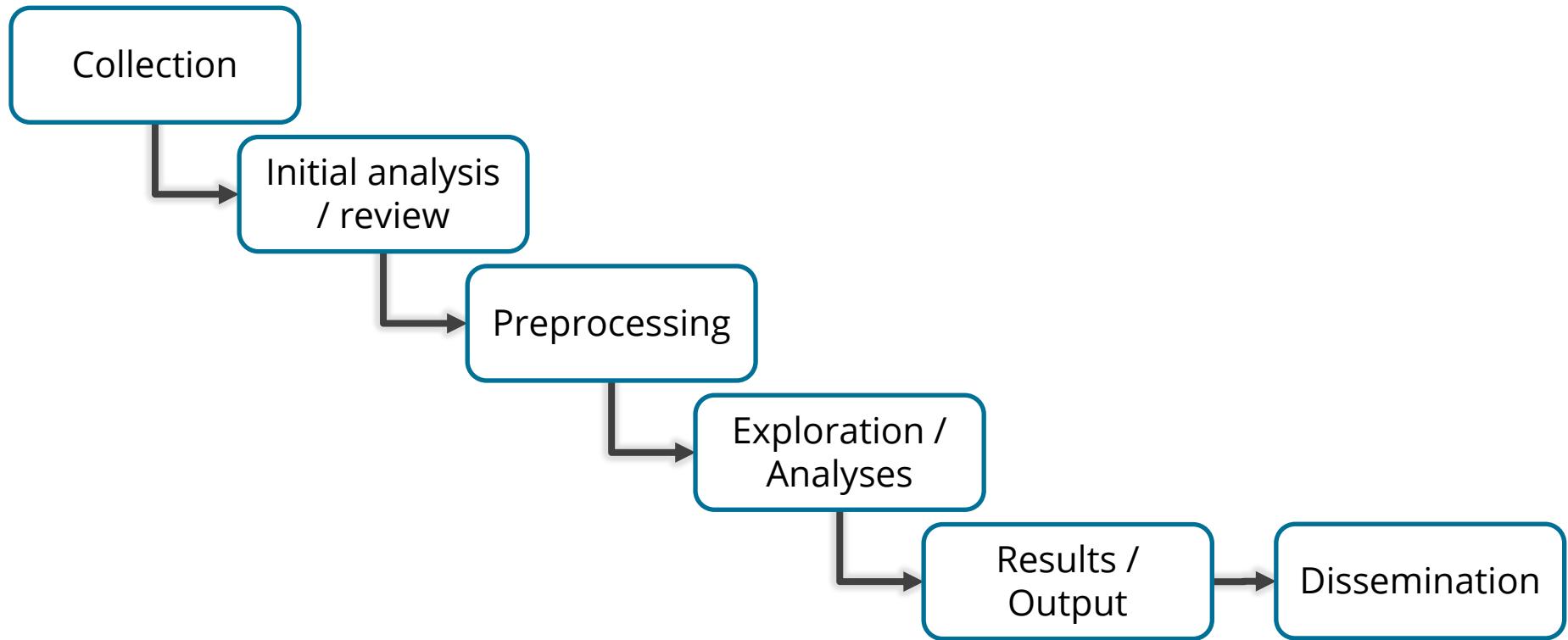
Relational semantics

Relationship extraction, Semantic parsing

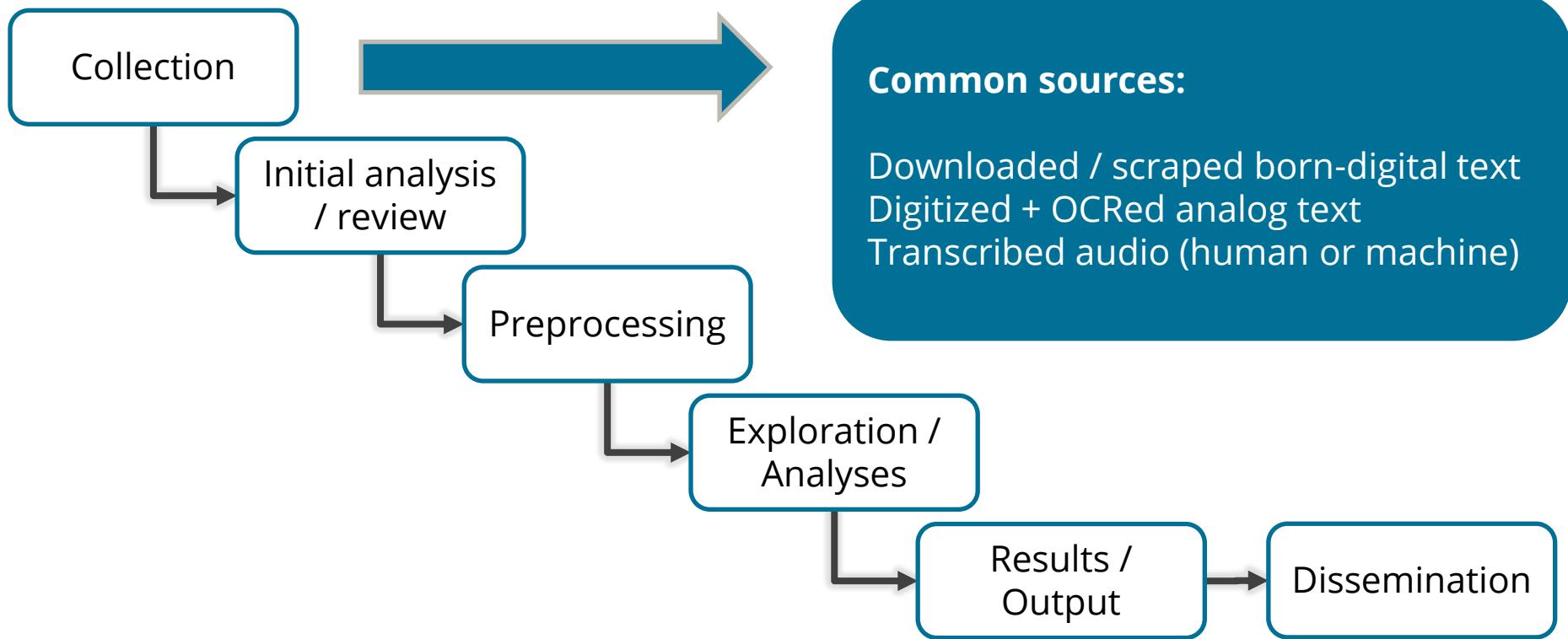
Discourse semantics

Discourse analysis, Topic segmentation,
Argument mining

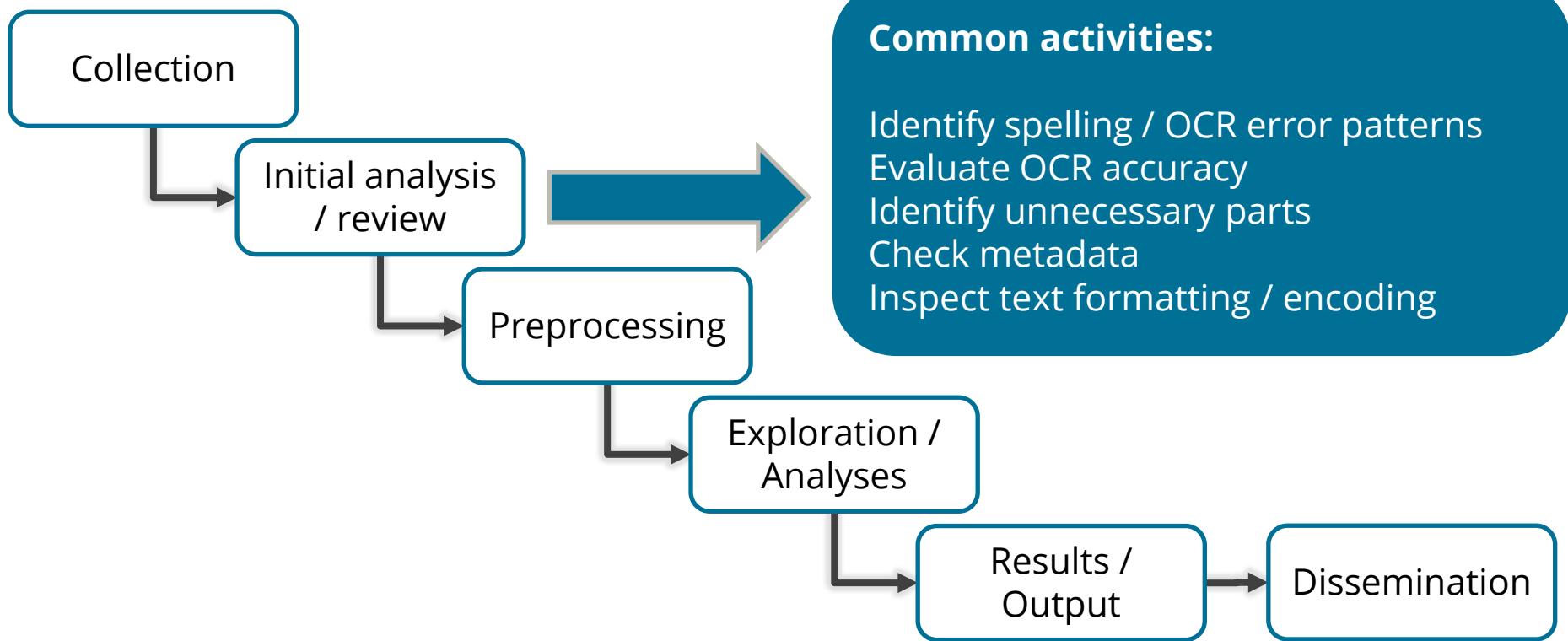
NLP Workflows



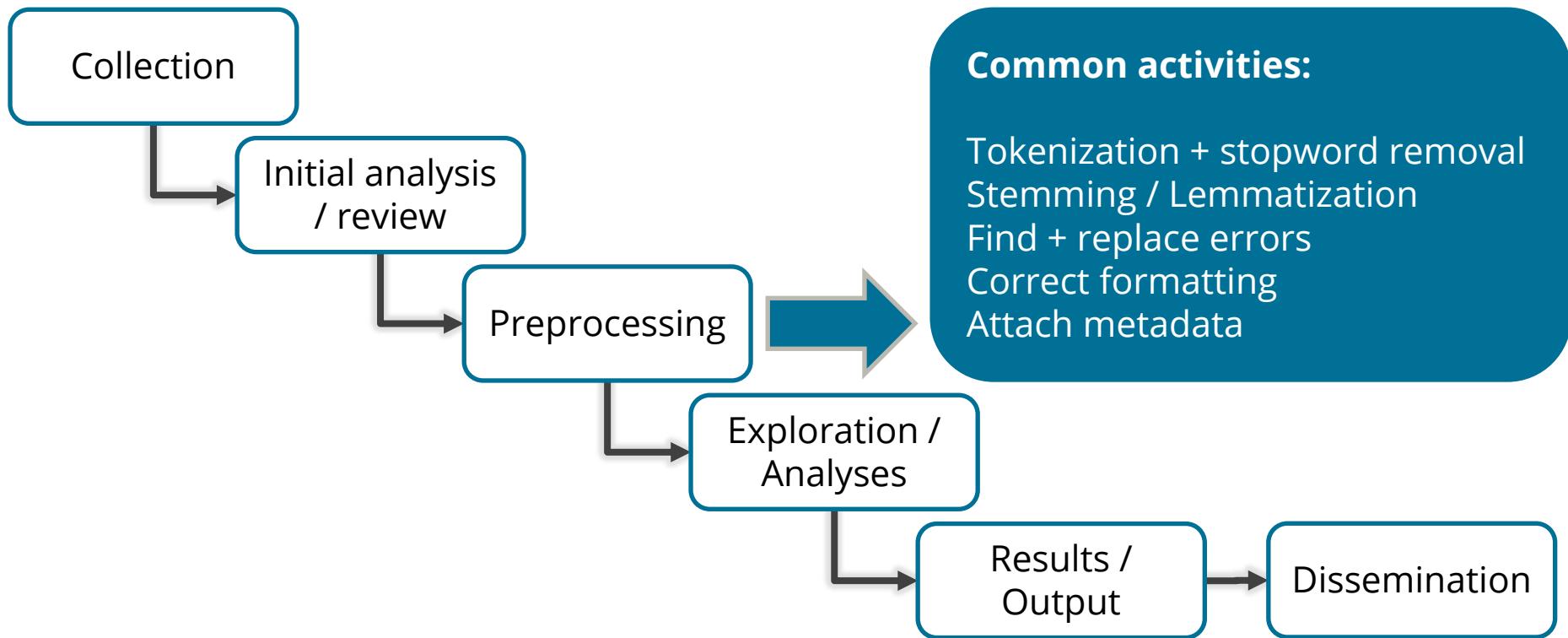
NLP Workflows



NLP Workflows

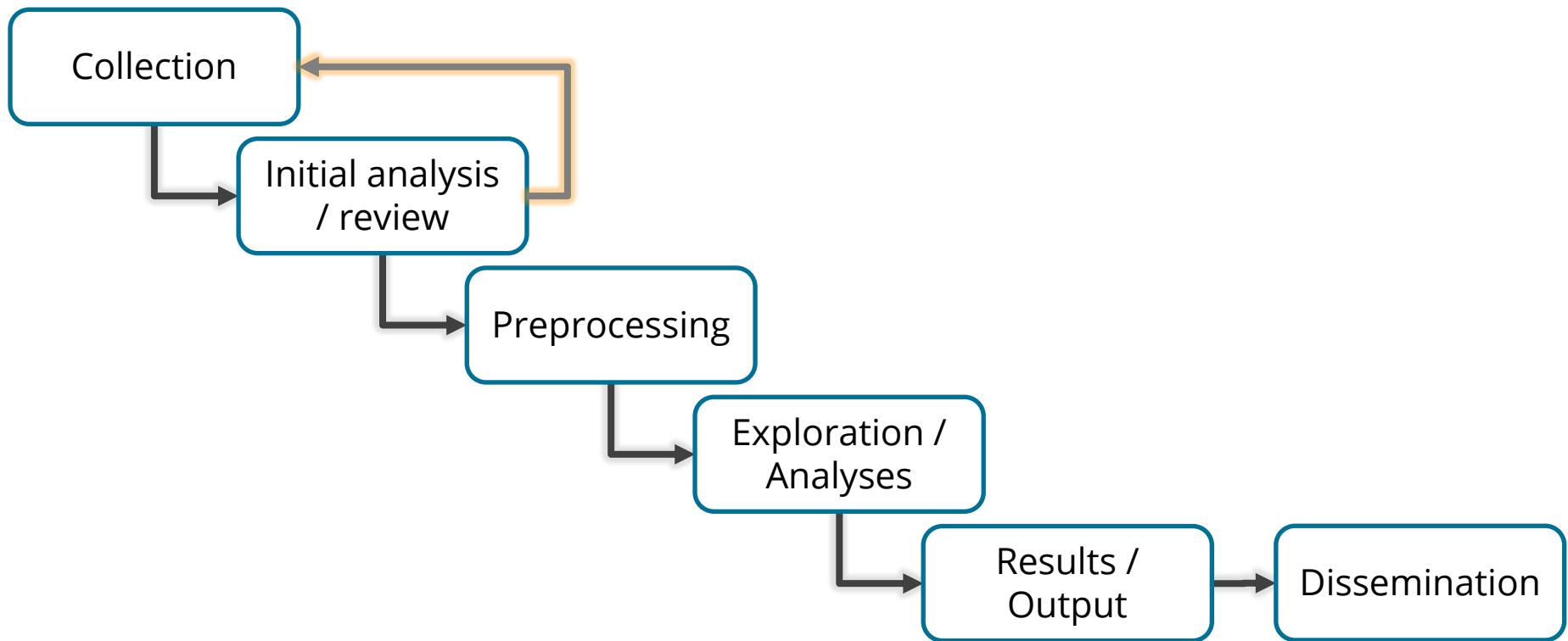


NLP Workflows



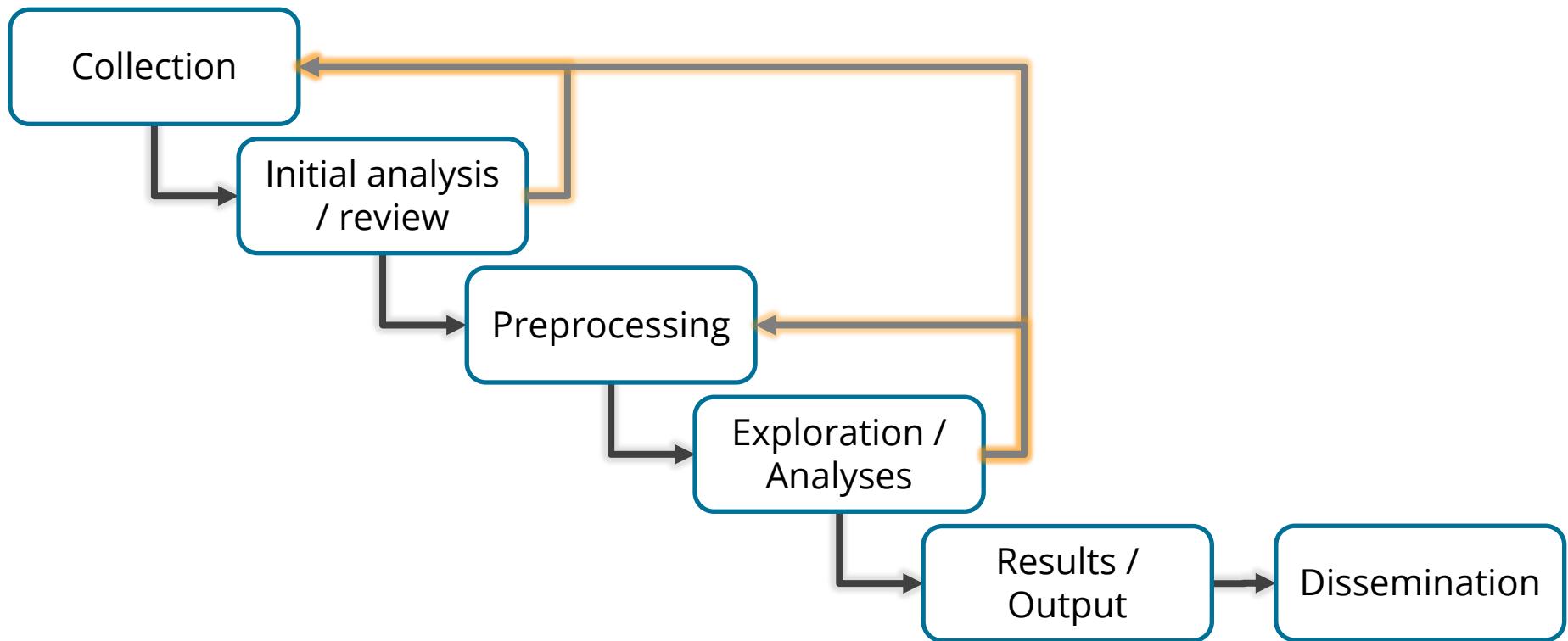
NLP Workflows

... are iterative



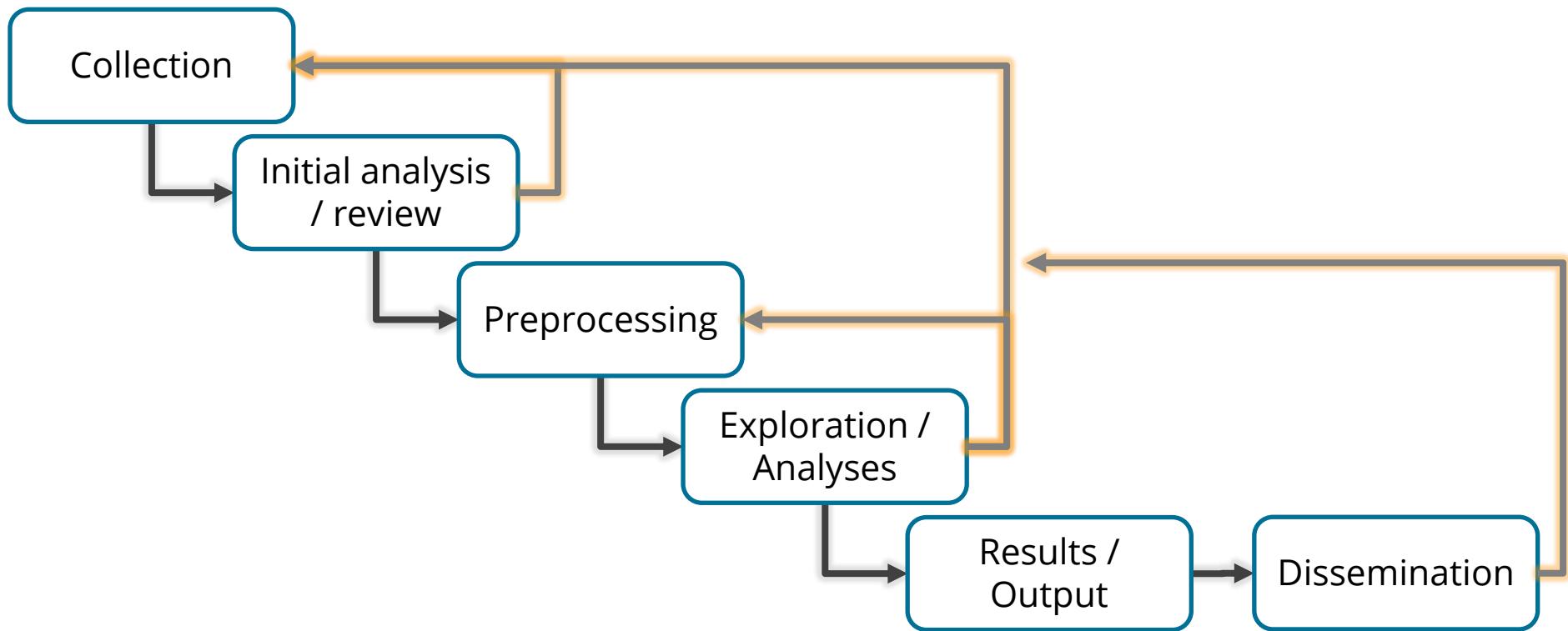
NLP Workflows

... are iterative



NLP Workflows

... are iterative

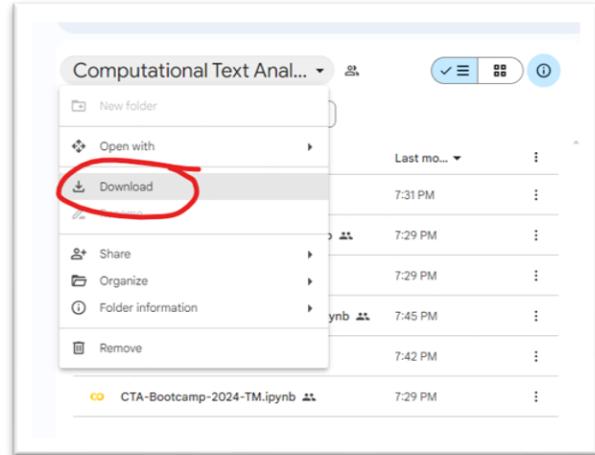


Getting started: Google Colab & Jupyter notebooks Get your data

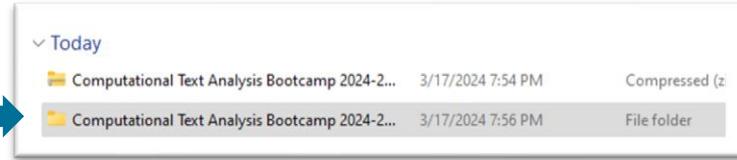
Go to u.mcmaster.ca/cta-bootcamp to download your data and view the workshops for today's workshop.
Follow along with Devon's instructions

Copying materials to your Google Drive

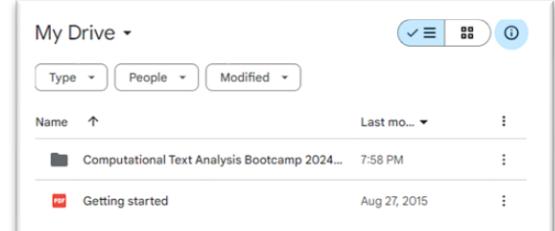
1. Go to u.mcmaster.ca/cta-bootcamp. Download the entire contents to your local computer (as a zipfile).



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



OME 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 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 pictures then known the earliest date from his seventeenth year, and it was not until nearly 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. 120. The Nativity, called the "Night," No. 152. The Madonna with St. Sebastian, No. 153.

73



i IPS^S^ffcls OME comments on correg ncRftJH GIO IN CONNECTION with 1^58^^^ HIS PICTURES IN DRESDEN. SpI^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 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, 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

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

The screenshot shows a Microsoft Word document window titled "zwick". The ribbon menu is visible at the top, with the "Review" tab selected. The main content area contains three paragraphs of text. The first paragraph discusses water level measurements taken in May and August 1990. The second paragraph describes the flow pattern of ground water within the shallow flow system, mentioning a radial outward flow from the island into the Bay of Quinte. The third paragraph discusses the topography of Zwick's Island and the surrounding area, noting that precipitation is the primary source of ground water recharge for the shallow flow system. It also mentions the possibility of seasonal fluctuations and storm activity causing temporary changes in water levels.

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 sufficiently (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.

Page 22 of 48 12457 words English (Canada) Focus 210%

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

backgroimd

acmal

coUect

Mimicipal

fmalizing

Dtiring

aroimd

stammg

landfiUed

moimted

Figu

Stratigraphie

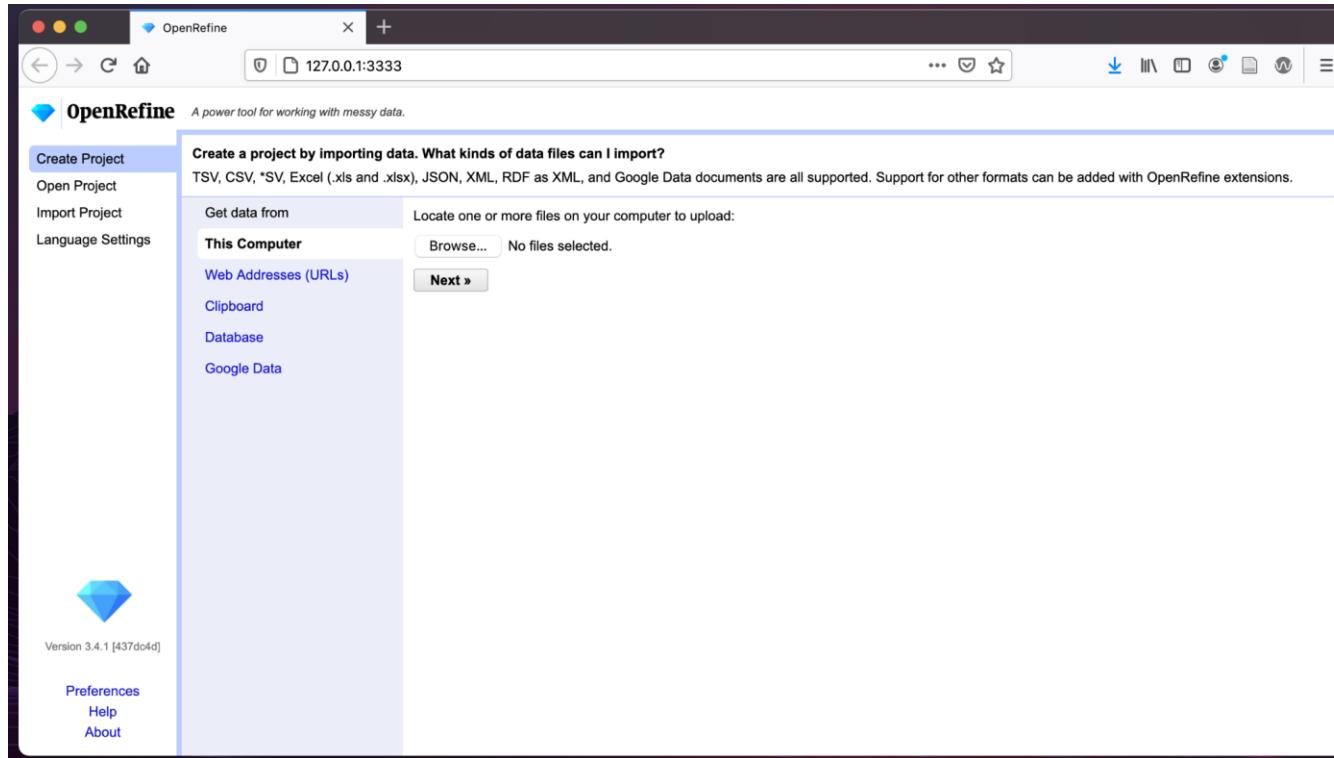
concurrenLly

Ruoride

Qilorde

Organo

Open OpenRefine



Initial Data Analysis: in OpenRefine

The screenshot shows the OpenRefine interface with a dataset titled "zwick-demo". The main pane displays 1151 rows of data, with the first few rows listed below:

Row	Content
1.	ZWICK'S ISLAND LANDFILL
2.	ENVIRONMENTAL INVESTIGATIONS
3.	FINAL REPORT
4.	OCTOBER 1991
5.	TABLE OF CONTENTS
6.	Letter of Transmittal
7.	PAGE
8.	L0 INTRODUCTION 1
9.	LI BACKGROUND 1
10.	L2 OBJECTIVES 2
11.	2.0 APPROACH 2
12.	2.1 DESK TOP INVENTORY 3
13.	2.2 SITE VISIT - 3
14.	2.3 SUBSURFACE INVESTIGATIONS 4
15.	2.4 SURFACE WATER MONITORING PROGRAM 7
16.	2.5 LANDFILL GAS INVESTIGATIONS 8
17.	3.0 PHYSICAL SETTING 9
18.	3.1 GEOGRAPHIC SETTING 9
19.	3.2 SITE HISTORY AND PRESENT USE 9
20.	3.3 SITE HYDROLOGY 10
21.	3.4 SUBSURFACE CONDITIONS 10
22.	4.0 LANDHILL GAS 12
23.	4.1 POTENTIAL FOR LANDFILL GAS 12
24.	4.2 OCCURRENCE OF LANDHILL GAS 13
25.	5.0 GROUND WATER AND LEACHATE 14
26.	5.1 PHYSICAL HYDROGEOLOGY 14
27.	5.2 GROUND WATER QUALITY 17
28.	6.0 SURFACE WATER 21
29.	6.1 SURFACE WATER QUALITY PARAMETERS OF INTEREST 21
30.	6.2 WATER QUALITY CRITERIA 22
31.	6.3 SURFACE WATER QUALITY 23
32.	7.0 IMPACTS 27
33.	7.1 IDENTIFICATION OF RECEPTORS, CONTAMINANT
34.	PATHWAYS AND CONTAMINANT LOADINGS 27
35.	7.1.1 Receptors 27
36.	7.1.2 Pathways 28

The interface includes a sidebar with a "Using facets and filters" guide and a "Watch these screencasts" link. The top right corner shows "Extensions: Wikidata".

Prepare Dataset for Text Analysis

Tokenize, trim and remove blank rows

1151 rows

Show as: rows records Show: 5 10 25 50 rows

All Column 1

- 1. Facet ► FILL
- 2. Text filter
- 3.
- 4.
- 5.
- 6.
- 7.
- 8.
- 9.
- 10. Edit cells ► Transform...
- 11. Edit column ► Common transforms ► Fill down
- 12. Transpose ► Blank down
- 13. Sort... ► Split multi-valued cells...
- 14. View ► Join multi-valued cells...
- 15. Reconcile ► Cluster and edit...
- 16. 2.0 APPROACH 2
- 17. 2.1 DESK TOP INVENTC
- 18. 2.2 SITE VISIT - 3
- 19. 2.3 SUBSURFACE INVESTIGATIONS 4
- 20. Replace

13607 rows

Show as: rows records Show: 5 10 25 50 rows

All Column 1

- 1. Facet ►
- 2. Text filter
- 3.
- 4. Edit cells ► Transform...
- 5.
- 6. Edit column ► Common transforms ► Trim leading and trailing whitespace
- 7. Transpose ► Fill down
- 8. Sort... ► Collapse consecutive whitespace
- 9.
- 10. View ► Blank down
- 11. Reconcile ► Unescape HTML entities
- 12. OCTOBER ► Replace Smart quotes with ascii
- 13.
- 14. Sort... ► To titlecase
- 15. TABLE ► To uppercase
- 16. OF ► To lowercase
- 17. CONTENTS ► To number
- 18. LETTER ► To date
- 19. OF ► To text
- 20. TRANSMITTAL ► To null
- 21. OF ► To empty string

Filter and Facet to Find Errors

Column 1

- Facet
 - Text facet
 - Numeric facet
- Edit cells
- Edit column
- Transpose
- Sort...
- View
 - Customized facets
 - Word facet
 - Duplicates facet
 - Numeric log facet
 - 1-bounded numeric log facet
 - Text length facet
 - Log of text length facet
 - Unicode char-code facet
- Reconcile
 - monitoring
 - contamination
 - remedial
 - mainly
 - municipal
 - former
 - some
 - commercial
 - system
 - managed
 - comer,
 - Ramada

Facet / Filter Undo / Redo 4 / 4

Refresh Reset All Remove All

Column 1 invert reset

m

case sensitive regular expression

Column 1 change

384 choices Sort by: name count

choice	count
decomposition	3
defme	1
demonstrate	1
demonstrates	1
determination	1
determine	13
determined	1
determining	1
dumping	1
dumpmg	1
emissions	1

Find and Replace with Text Filter

Facet / Filter Undo / Redo 4 / 4

Refresh Reset All Remove All

Column 1

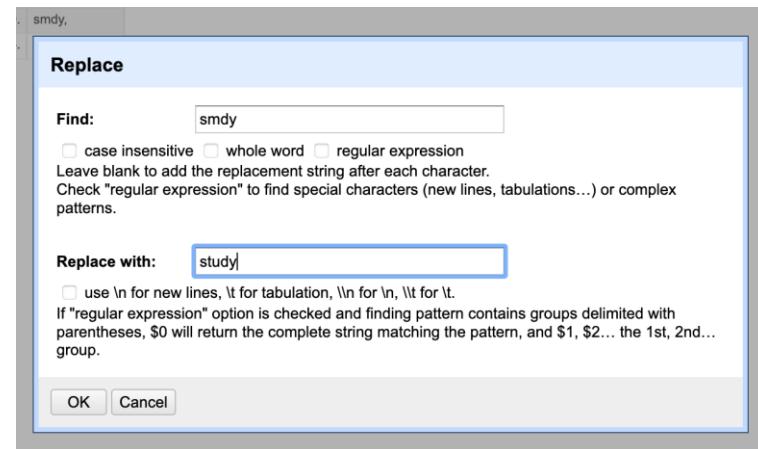
smdy

case sensitive regular expression

8 matching rows (12457 total)

Show as: rows records Show: 5 10 25 50 rows

All	Column 1
1082.	Facet
1089.	Text filter
1209.	
4539.	
6835.	
10145.	Edit cells
10849.	Transform...
12155.	Edit column
	Common transforms
	Transpose
	Fill down
	Blank down
	Sort...
	View
	Split multi-valued cells...
	Join multi-valued cells...
	Reconcile
	Cluster and edit...
	Replace



Find and Replace with Text Filter

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

The image shows a user interface for a search or filtering application. On the left, there is a 'Facet / Filter' panel with a blue header. It contains several sections:

- Column 1**: A list with 'no' selected. Below it are checkboxes for 'case sensitive' (checked) and 'regular expression'.
- Column 1**: A list with 'not 27', 'noted 5', 'noted. 1', 'noticeably 1', 'Ontano 1', 'Organo 1', 'organo 1', 'organochlorine 2', 'Organochlorine 2', 'Phenol 2', and 'phenol 3'. There is also an 'include' link next to 'Ontano 1'.

On the right, a 'Replace' dialog box is open:

Replace

Find: no

case insensitive whole word regular expression

Leave blank to add the replacement string after each character.
Check "regular expression" to find special characters (new lines, tabulations...) or complex patterns.

Replace with: rio

use \n for new lines, \t for tabulation, \\n for \n, \\t for \t.
If "regular expression" option is checked and finding pattern contains groups delimited with parentheses, \$0 will return the complete string matching the pattern, and \$1, \$2... the 1st, 2nd...

OK **Cancel**

Find and Replace with GREL

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

The screenshot shows a data processing interface with a context menu open for 'Column 1'. The menu includes options like 'Facet', 'Text filter', 'Edit cells' (which is currently selected), 'Edit column', 'Transpose', 'Sort...', 'View', 'Reconcile', and 'Replace'. The 'Edit cells' option has a submenu with 'Transform...' selected.

A 'Custom text transform' dialog box is displayed over the interface. It shows the expression 'value.replace('im','un')' in the 'Expression' field, with the 'Language' set to 'General Refine Expression Language (GREL)'. Below the expression, a preview table shows the transformation of several rows:

row	value	value.replace('im','un')
1252.	backgroundd	background
1406.	aroimd	around
1679.	truck-moimted	truck-mounted
3853.	encointered	encountered
4827.	Groimd	Ground
4931.	groimd	ground

At the bottom of the dialog, there are settings for 'On error': 'keep original' (selected), 'set to blank', and 'store error'. There is also an option to 'Re-transform up to 10 times until no change'. The 'OK' and 'Cancel' buttons are at the bottom left.

Find and Replace with GREL

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

The screenshot shows a 'Custom text transform' dialog box. At the top, there's a preview of two rows from a dataset: '8454. slightly' and '8516. slightly'. Below this, the main area has a title 'Custom text transform on column Column 1'. The 'Expression' field contains the GREL code: `value.replace('tiy','tly')`. The 'Language' dropdown is set to 'General Refine Expression Language (GREL)'. A status message 'No syntax error.' is displayed next to the expression. Below the expression, there's a 'Preview' tab which is selected, showing a table with the original values and the transformed values using the GREL expression. The table includes rows for 4749, 4956, 7589, 8454, 8516, and 9546. At the bottom of the dialog, there are options for handling errors: 'On error' with radio buttons for 'keep original', 'set to blank', or 'store error', and a checkbox for 'Re-transform up to 10 times until no change'. Finally, there are 'OK' and 'Cancel' buttons at the bottom.

row	value	value.replace('tiy','tly')
4749.	consistently	consistently
4956.	sufficiently	sufficiently
7589.	significantly	significantly
8454.	slightly	slightly
8516.	slightly	slightly
9546.	significantly	significantly

Find and Replace with Regular Expression (Regex)

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

The screenshot shows a user interface for searching and filtering data. On the left, there's a 'Facet / Filter' panel with a search bar containing 'mg\$', a 'case sensitive' checkbox (unchecked), and a 'regular expression' checkbox (checked). Above the search bar are 'Refresh', 'Reset All', and 'Remove All' buttons. To the right of the search bar is an 'invert reset' link. The main area displays '5 matching rows (12457 total)' and a table with columns for row number, ID, and name. The table shows five rows where the name ends in 'mg'.

	All	Column 1
1414.	stammg	
2988.	dumpmg	
4151.	ventmg	
4386.	bemg	
12248.	reconstructmg	

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

Custom Tabular Exporter

[Content](#) [Download](#) [Upload](#) [Option Code](#)

Line-based text formats

Tab-separated values (TSV)
 Comma-separated values (CSV)
 Custom separator
Line separator
Character encoding UTF-8
Always quote text

Other formats

Excel (.xls)
 Excel in XML (.xlsx)
 HTML table

[Preview](#) [Download](#)

[Cancel](#)

Export your “Recipe” of Tasks

Extract Operation History

Extract and save parts of your operation history as JSON that you can apply to this or other projects in the future.

Split multi-valued cells in column Column 1

Text transform on cells in column Column 1 using expression value.trim()
Star row 4

Remove rows

Text transform on cells in column Column 1 using expression value.replace("smdy","study")

Text transform on cells in column Column 1 using expression value.replace("no","rio")

Text transform on cells in column Column 1 using expression value.replace("no","nic")

Text transform on cells in column Column 1 using expression grel:value.replace('im','un')

Text transform on cells in column Column 1 using expression grel:value.replace('iy','ly')

Text transform on cells in column Column 1 using expression grel:value.replace('mg','ing')

Mass edit cells in column Column 1
Star row 11473

```
{
  "facets": [
    {
      "type": "text",
      "name": "Column 1",
      "columnName": "Column 1",
      "query": "staming",
      "mode": "regex",
      "caseSensitive": false,
      "invert": false
    }
  ],
  "mode": "row-based",
  "columnName": "Column 1",
  "expression": "value",
  "edits": [
    {
      "from": [
        "staming"
      ],
      "fromBlank": false,
      "fromError": false,
      "to": "staining"
    }
  ],
  "description": "Mass edit cells in column C"
}
```

Programmatic approaches with Python

Text prep and analysis as a continuum of mediation

Completely
manual

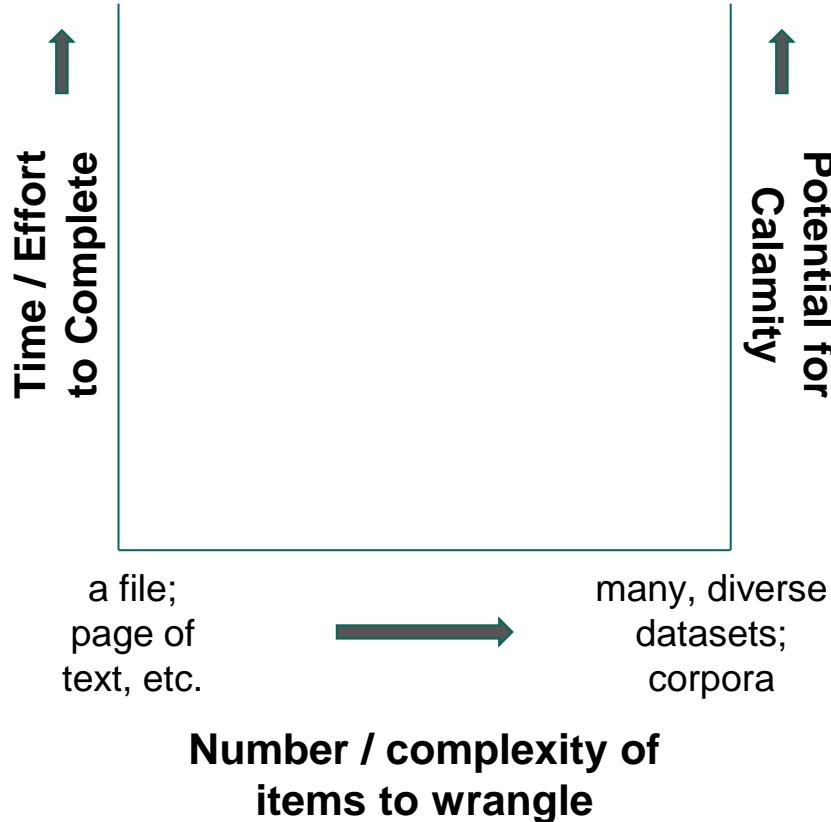


Completely
automated

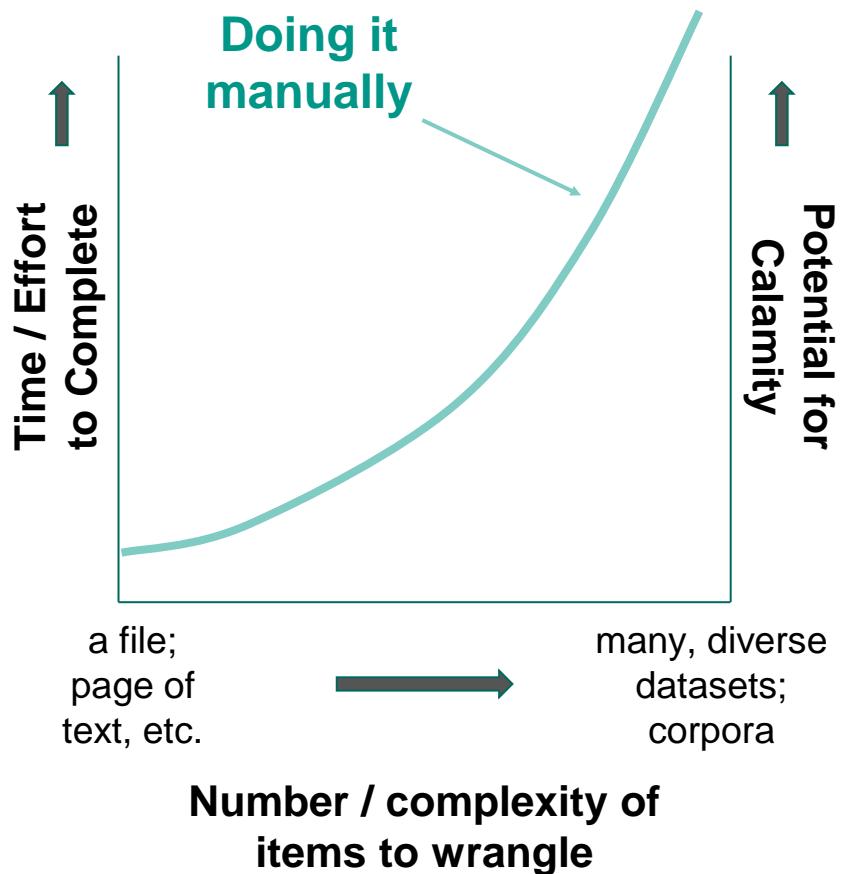
Text prep and analysis as a continuum of mediation



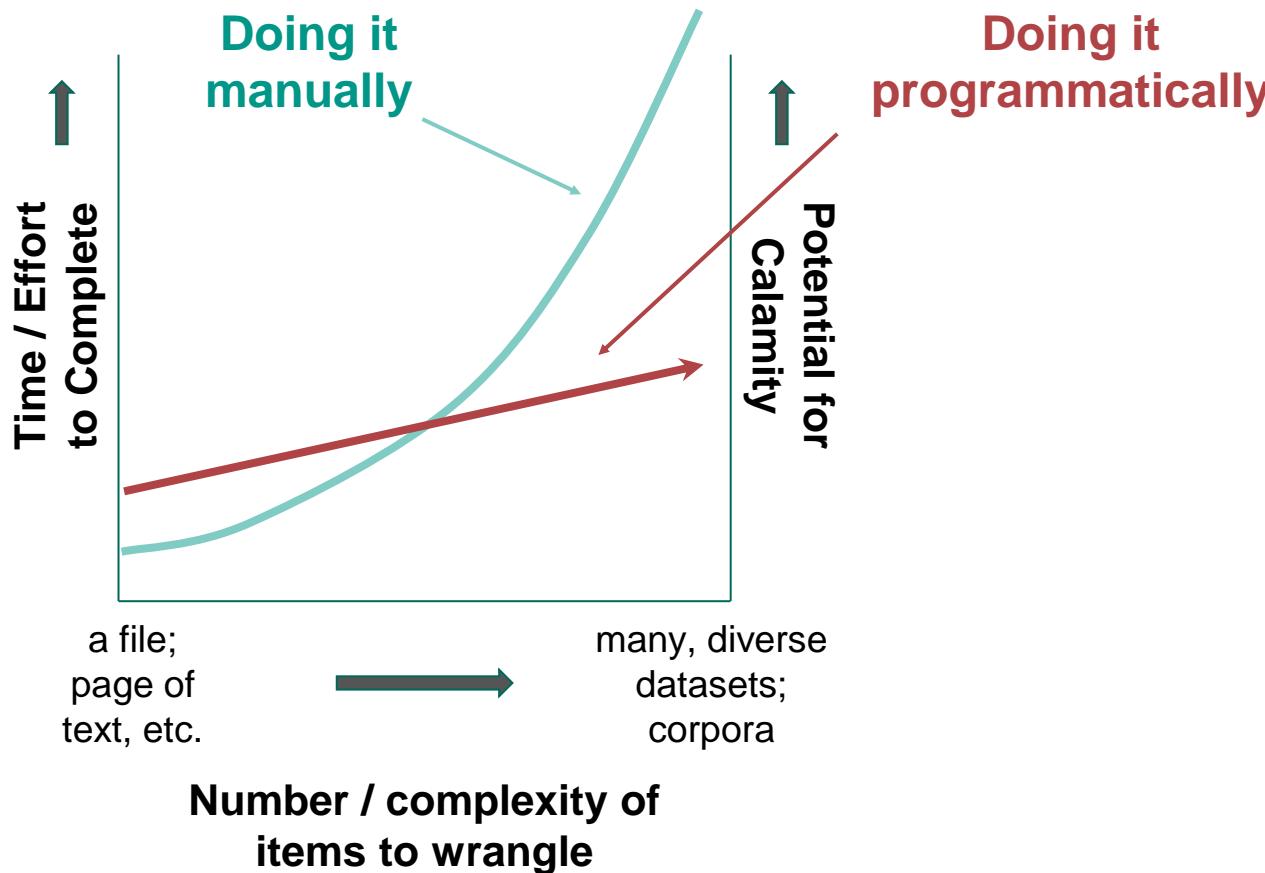
So, when to let the computer take over?



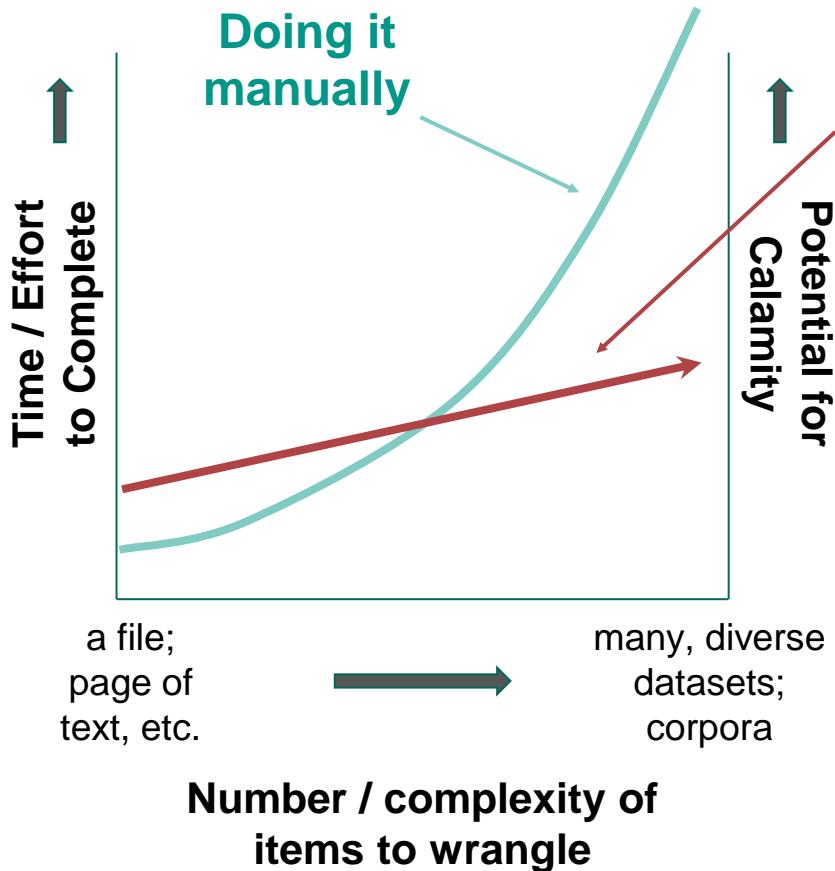
Spreadsheets: The frenemy of research



Spreadsheets: The frenemy of research



Spreadsheets: The frenemy of research



Doing it programmatically

HOW LONG CAN YOU WORK ON MAKING A ROUTINE TASK MORE EFFICIENT BEFORE YOU'RE SPENDING MORE TIME THAN YOU SAVE? (ACROSS FIVE YEARS)

		HOW OFTEN YOU DO THE TASK					
		50/DAY	5/DAY	DAILY	WEEKLY	MONTHLY	YEARLY
HOW MUCH TIME YOU SHAVE OFF	1 SECOND	1 DAY	2 HOURS	30 MINUTES	4 MINUTES	1 MINUTE	5 SECONDS
	5 SECONDS	5 DAYS	12 HOURS	2 HOURS	21 MINUTES	5 MINUTES	25 SECONDS
	30 SECONDS	4 WEEKS	3 DAYS	12 HOURS	2 HOURS	30 MINUTES	2 MINUTES
	1 MINUTE	8 WEEKS	6 DAYS	1 DAY	4 HOURS	1 HOUR	5 MINUTES
	5 MINUTES	9 MONTHS	4 WEEKS	6 DAYS	21 HOURS	5 HOURS	25 MINUTES
	30 MINUTES		6 MONTHS	5 WEEKS	5 DAYS	1 DAY	2 HOURS
	1 HOUR			10 MONTHS	2 MONTHS	10 DAYS	2 DAYS
	6 HOURS	1 DAY			2 MONTHS	2 WEEKS	1 DAY
						8 WEEKS	5 DAYS

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

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



— Text to annotate —

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.

— Annotations —

— Language — English Submit

named entities ×

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

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')

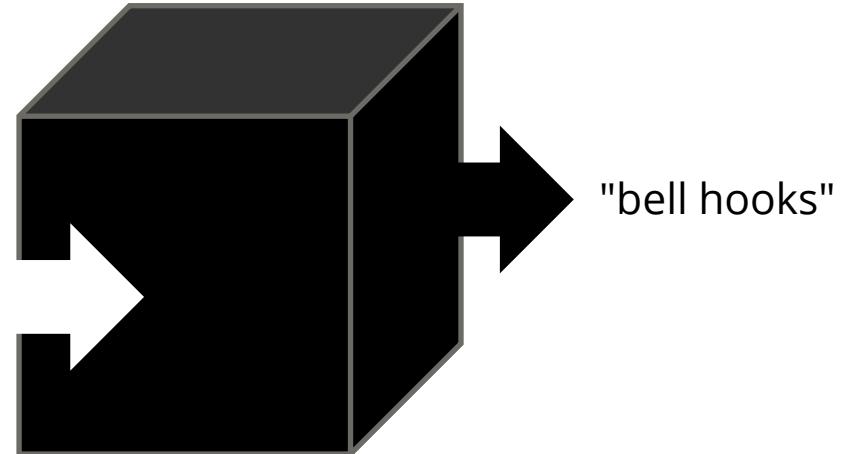
# Create the Doc object by passing it through the text pipeline (nlp)
doc = nlp(ner_text)

[ ] 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."



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

Run 50 iterations 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

Topic Documents Topic Correlations Time Series Vocabulary Downloads

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

Use a different collection:

Documents Browse... d-corp.txt

Stoplist Browse... No file selected.

Upload

[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 (art)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!

+ 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
