Penn Libraries
University *of* Pennsylvania

Introduction to Computational Text Analysis

Jajwalya Karajgikar
Applied Data Science Librarian
Research Data and Digital Scholarship
Penn Libraries

Andy Janco
Digital Scholarship Programmer
Research Data and Digital Scholarship
Penn Libraries

RDDS
RESEARCH DATA
& DIGITAL
SCHOLARSHIP

Fall 2022

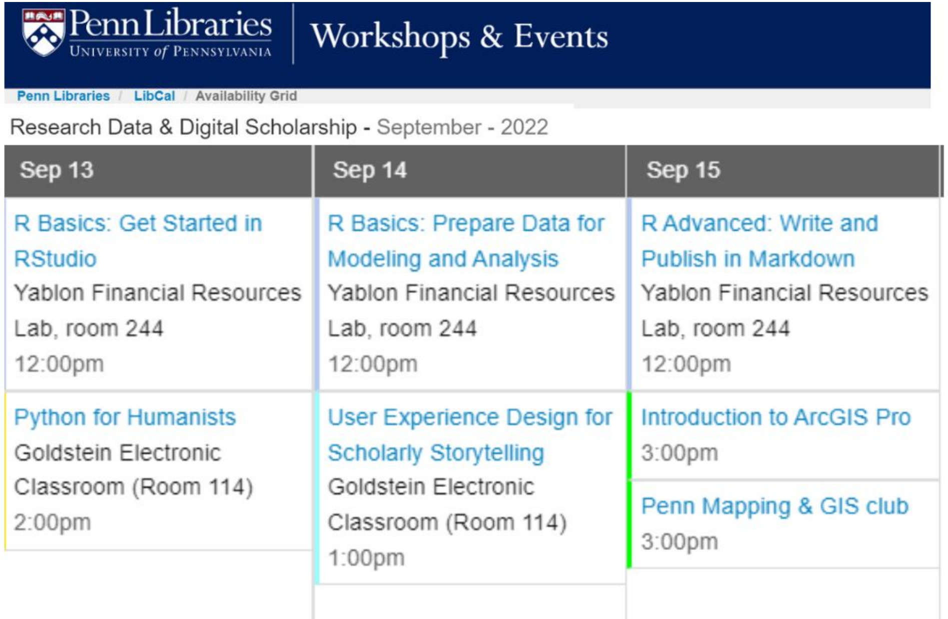
Contact libraryrdds@pobox.upenn.edu for queries on Data / Digital Projects

RDDS Clubs and User Groups

PUG@Penn

R Penn Group

Penn MGIS



Penn Libraries University of Pennsylvania | Workshops & Events

Penn Libraries / LibCal / Availability Grid

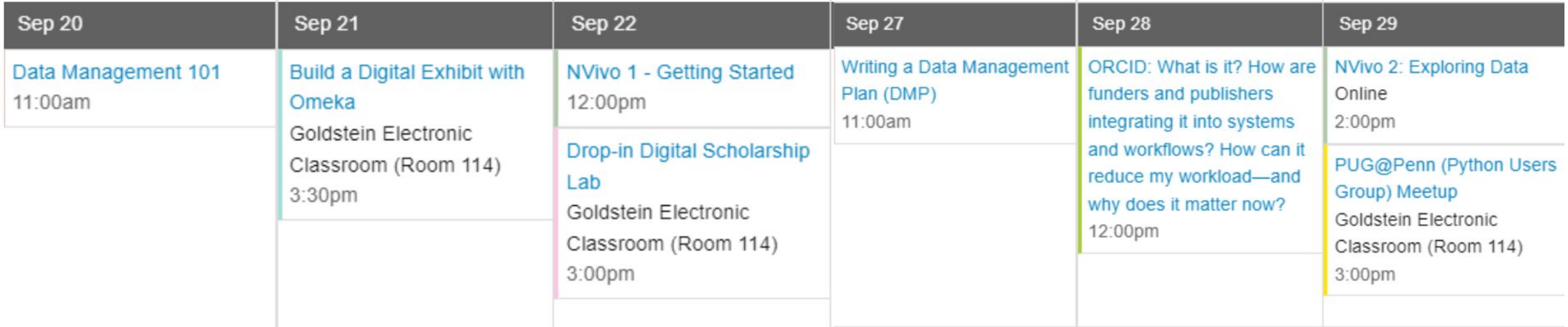
Research Data & Digital Scholarship - September - 2022

Sep 13	Sep 14	Sep 15
R Basics: Get Started in RStudio Yablon Financial Resources Lab, room 244 12:00pm	R Basics: Prepare Data for Modeling and Analysis Yablon Financial Resources Lab, room 244 12:00pm	R Advanced: Write and Publish in Markdown Yablon Financial Resources Lab, room 244 12:00pm
Python for Humanists Goldstein Electronic Classroom (Room 114) 2:00pm	User Experience Design for Scholarly Storytelling Goldstein Electronic Classroom (Room 114) 1:00pm	Introduction to ArcGIS Pro 3:00pm
		Penn Mapping & GIS club 3:00pm

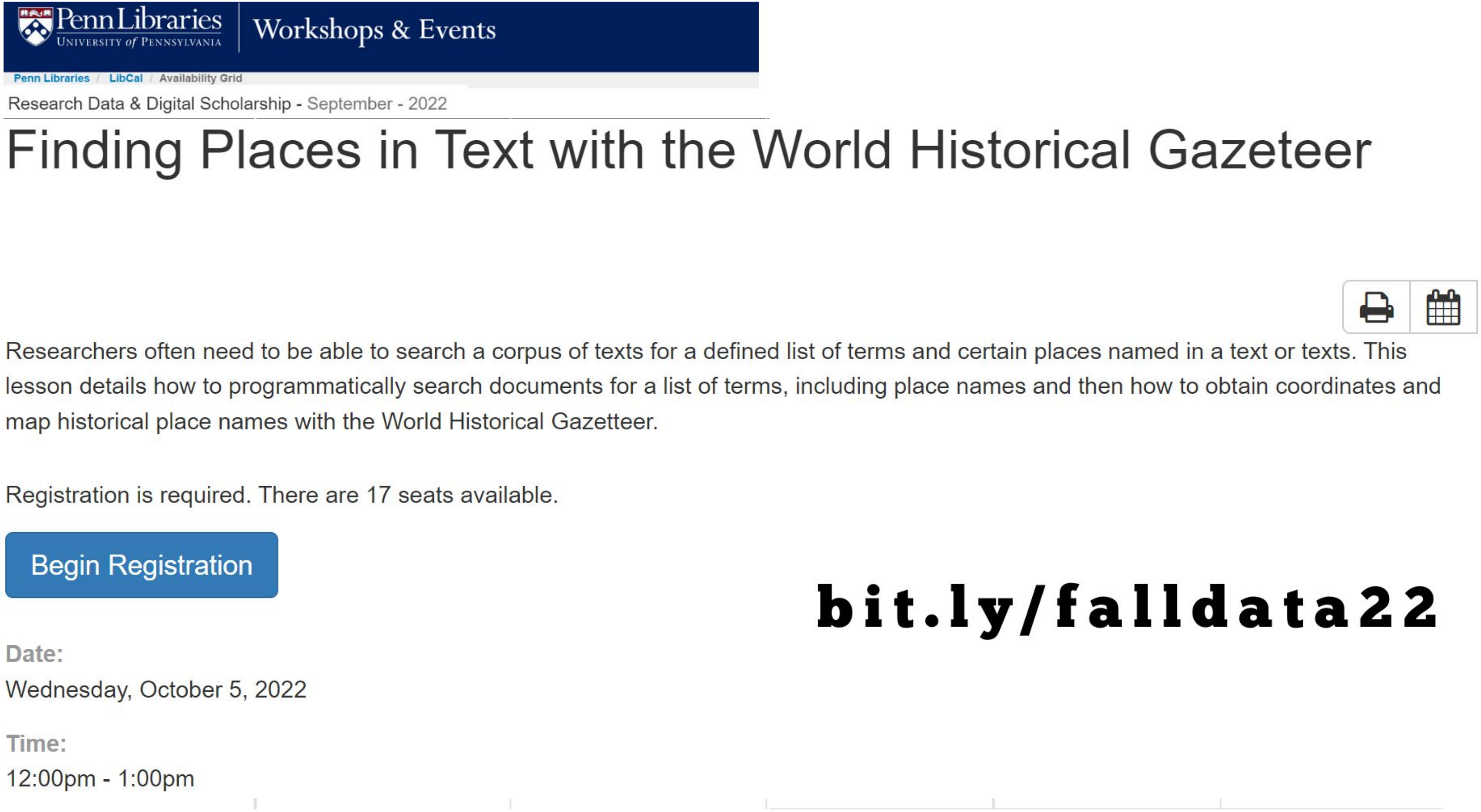
bit.ly/falldata22

Feedback form:

upenn.libwizard.com/f/rdds-survey



Sep 20	Sep 21	Sep 22	Sep 27	Sep 28	Sep 29
Data Management 101 11:00am	Build a Digital Exhibit with Omeka Goldstein Electronic Classroom (Room 114) 3:30pm	NVivo 1 - Getting Started 12:00pm	Writing a Data Management Plan (DMP) 11:00am	ORCID: What is it? How are funders and publishers integrating it into systems and workflows? How can it reduce my workload—and why does it matter now? 12:00pm	NVivo 2: Exploring Data Online 2:00pm
		Drop-in Digital Scholarship Lab Goldstein Electronic Classroom (Room 114) 3:00pm			PUG@Penn (Python Users Group) Meetup Goldstein Electronic Classroom (Room 114) 3:00pm



Penn Libraries
University of Pennsylvania
Workshops & Events
Penn Libraries / LibCal / Availability Grid
Begin Registration

Research Data & Digital Scholarship - September - 2022

Finding Places in Text with the World Historical Gazeteer

Researchers often need to be able to search a corpus of texts for a defined list of terms and certain places named in a text or texts. This lesson details how to programmatically search documents for a list of terms, including place names and then how to obtain coordinates and map historical place names with the World Historical Gazetteer.

Registration is required. There are 17 seats available.

Date:
Wednesday, October 5, 2022

Time:
12:00pm - 1:00pm

bit.ly/falldata22

Goals for the Workshop - Part 1

What is Text Analysis? Why use Text Analysis?

When to use Text Analysis Techniques?

What are the Text Analysis terms to know?

What are sources for text data (specific to Penn researchers)?

Different platforms that don't require coding (Voyant, MALLET, Pinpoint, AntConc)

Different platforms that require Intermediate understanding of Python / R coding (Proquest TDM Studio, Constellate)

When to Use Computational Text Analysis Techniques like Topic Modeling, Authorship Attribution, Clustering, Classification, Sentiment Analysis?

Goals for the Workshop - Part 2

Natural Language Processing

Entity Extraction

Entity Linking

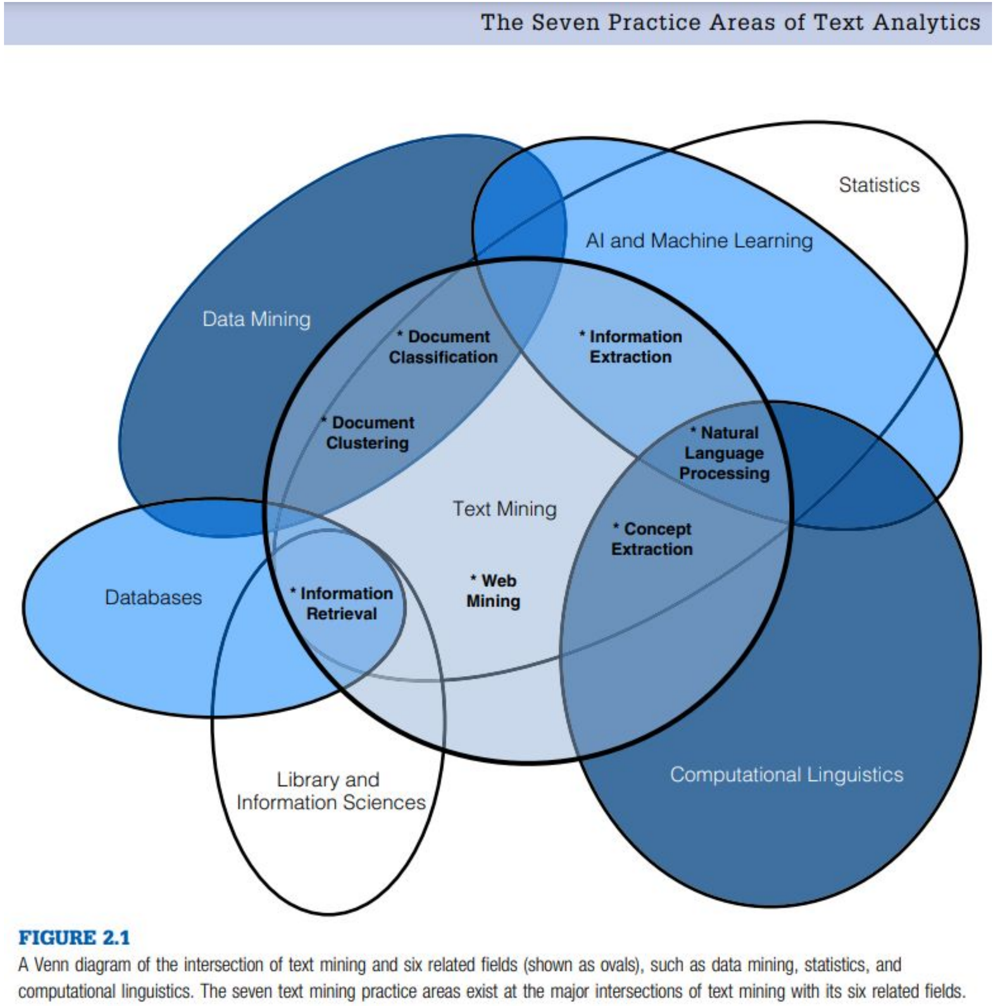
Dependency Parsing

Categorization

Not Goals for This Workshop

- ✖ How to do Text Mining?
- ✖ How to do **every** Text Analysis method?
- ✖ What is the statistical background of Text Analysis methods?
- ✖ What is the comprehensive overview of the state-of-the-art of TDM methods & applications?
- ✖ How to code?

If you have specific TDM questions, please Contact LibraryRDDS@pobox.upenn.edu



The Seven Practice Areas of Text Analytics

FIGURE 2.1
A Venn diagram of the intersection of text mining and six related fields (shown as ovals), such as data mining, statistics, and computational linguistics. The seven text mining practice areas exist at the major intersections of text mining with its six related fields.

Miner, Gary D. et al. “Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.” (2012).

Computational Text Analysis - What and Why?

- Extracting information from texts, such as novels, monographs, articles, web pages, research papers, newspapers, gazetteers, etc
- **Computational Text Analysis, Computer-aided Text Analysis, Text Mining**, and the abbreviation **TDM** are broad terms for searching, organizing, and analyzing large amounts of text data.
- Detecting patterns
  - Word frequency
  - Associative links between words
  - Co-occurances
- Solution to data abundance
- Combines with Close Reading Techniques = Distant Reading
- Qualitative + Quantitative Methodology
- Making sense of Unstructured Data (text-heavy data that is not organized in a pre-defined manner)

Text Analysis Terms

Corpus: Pl. Corpora, a collection of written texts, particularly the entire body of work on a subject or by a specific creator; a collection of written or spoken material in machine-readable form, assembled for the purpose of studying linguistic structures, frequencies, etc.

Text preprocessing: Cleaning, normalizing, and preparing text data for analysis by removing stopwords, punctuations, wide spaces, etc.

Stopwords: words filtered out before or after processing of natural language data (text), usually words with little meaning such as “and,” “the,” “a,” “an”

-From Folgerpedia’s Glossary of Digital Humanities Terms

Text Data Sources

- Library Databases (Contact your Subjects Specialists)
- Linguistic Data Consortium
- Open Source Search:
  - re3data.org
  - datasetsearch.research.google.com
  - huggingface.co
  - kaggle.com/datasets
  - github.com/awesomedata/awesome-public-datasets
- Web Scraping Data: Copying website information in order to extract large amounts of data and saving to a local file is web crawling or spider or scraping.
  - **Please note that not all online resources allow text mining and that there are** legal and ethical limitations **to consider.**
- Social Media (Twitter) Data Mining: Social media data access and availability depends on the platform(s) and time period of your research. There are legal and ethical considerations when it comes to social media data mining.
- Extracting Texts:
  - Perform Optical Character Recognition (OCR) using ABBYY Finereader at the Butler Assistive Technology Room.
  - Textract using Python extracts text from docx, pdfs, images and sound.
  - Tesseract is a free software to OCR your documents. Using Tesseract requires experience using the command line.

When to Use Computational Text Analysis Techniques?

Data Availability

Time constraints

Coding Background

Text Analysis Platforms that don't require coding

- Voyant Tools: Voyant tool is an open-source, web-based text reading and analysis environment.
- MALLET: Mallet is a Java programming language-based software for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text.
- AntConc: (Tutorial) A freeware corpus analysis toolkit for concordancing and finding clusters (frequency patterns of word sequences) or n-grams (sequences of n words within your corpus or document).
- Google Pinpoint: Part of Google's Journalist Studio, search keywords and identify entities in large amounts of text

Data is Available + Less Time + Beginner to Coding

Voyant Tools

- Developed by Stéfan Sinclair (McGill University) and Geoffrey Rockwell (University of Alberta)
  - Citation: Sinclair, Stéfan and Geoffrey Rockwell, 2016. *Voyant Tools*. Web. http://voyant-tools.org/.
  - Documentation: https://voyant-tools.org/docs/#!/guide
- Free, Open-source project
- Web-based text reading and analysis environment
- Lower the barrier of entry for text analysis
  - No coding required
  - No login required
- Large, robust user community
- Consistently upgraded and supported infrastructure
- Balances user-friendliness with powerful functionality
  - 25+ visualization tools
- Hermeneutical principle
  - Combines Distant Reading + Close Reading Techniques

VOYANT

see through your text

Possible Research Questions

What do these articles have in common? How are they different? Choose several (2-3) longer-form articles on a specific topic and have students read them in advance of a class. Summarize each article in class, discuss what they understood about the articles. Then put articles in Voyant Tools and discuss what changes when you read them with a bird's eye view. What is obvious? What is less obvious about these articles in this format? Do they all use specific vocabulary the same way? What makes them similar, what makes them different?

How do we understand the work we've been doing from a different perspective?If you and your students have been focusing a lot on a specific text during the semester, upload a copy of it to Voyant Tools. Discuss where your focal point(s) have been, and use these as "ins" to the corpus. Can we find those words in the word cloud? Are there any words in there that surprise you? Why or why not? Explore context: any patterns or recurrent themes in surrounding terms or concepts of interest that become more easily identifiable with the key word in context viewer? Does one of them take off after a certain point in the text and supercede another?

What does my writing look like?Classes with regular writing assignments (such as Canvas posts or short response essays) often discuss the editorial process. Having students upload their own writing and observe what their own language looks like in a non-linear way offers a twist on the classic "review your essay and think about what you say". Are there specific phrases or words you like to use a lot? What other ways are there to present the same idea but use some different language?

From Penn State LibGuide

Possible Use Cases

Examples of use-cases:

- Computers-assisted analysis works for academic tasks.
- Add functionality to online collections, journals, blogs or websites so others can see through your texts with analytical tools.
- Add interactive evidence to your essays that you publish online.
- Add interactive panels right into your research essays (if they can be published online) so your readers can recapitulate your results.
- Develop your own tools using our functionality and code.

https://voyant-tools.org/

Voyant Tools

Resources
Examples
Tutorials

Download Voyant Server, Documentation

RDDS Blogpost and Workshop Materials (Audience: undergrads)

De Caro W, Mitello L, Marucci AR, Lancia L, Sansoni J. Textual Analysis and Data Mining: An Interpreting Research on Nursing. Stud Health Technol Inform. 2016;225:948. PMID: 27332424.

Chen, Jingfeng & Wei, Wei & Guo, Chonghui & Lin, Tang & Sun, Leilei. (2017). Textual analysis and visualization of research trends in data mining for electronic health records. Health Policy and Technology. 6. 10.1016/j.hlpt.2017.10.003.

Medical Text Analytic Techniques And Its Applications

Chapter-wise Tutorials: Welcome to Dialogica: Thinking-Through Voyant!

Spanish Tutorial:
https://programminghistorian.org/es/lecciones/analisis-voyant-tools

Text Data Sources - Platforms that require some understanding of Code

TDM Platforms for Historic Newspaper & Text Data

- Constellate

  Full-text and metadata from JSTOR resources is now available for visualizing and analyzing within JSTOR's platform.

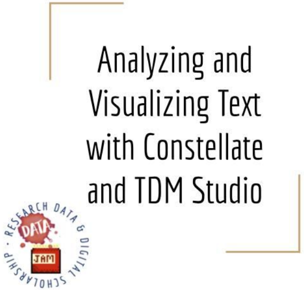
- Gale Text Data

  Penn Libraries can provide access to full text data from numerous Gale research databases including ECCO, the Gale NewsVault, and Archives Unbound.

- Proquest TDM Studio

  Full-text data from ProQuest databases is available for visualizing and analyzing within the ProQuest platform

Data is Available Within the Platform + More Time + Some experience with Coding

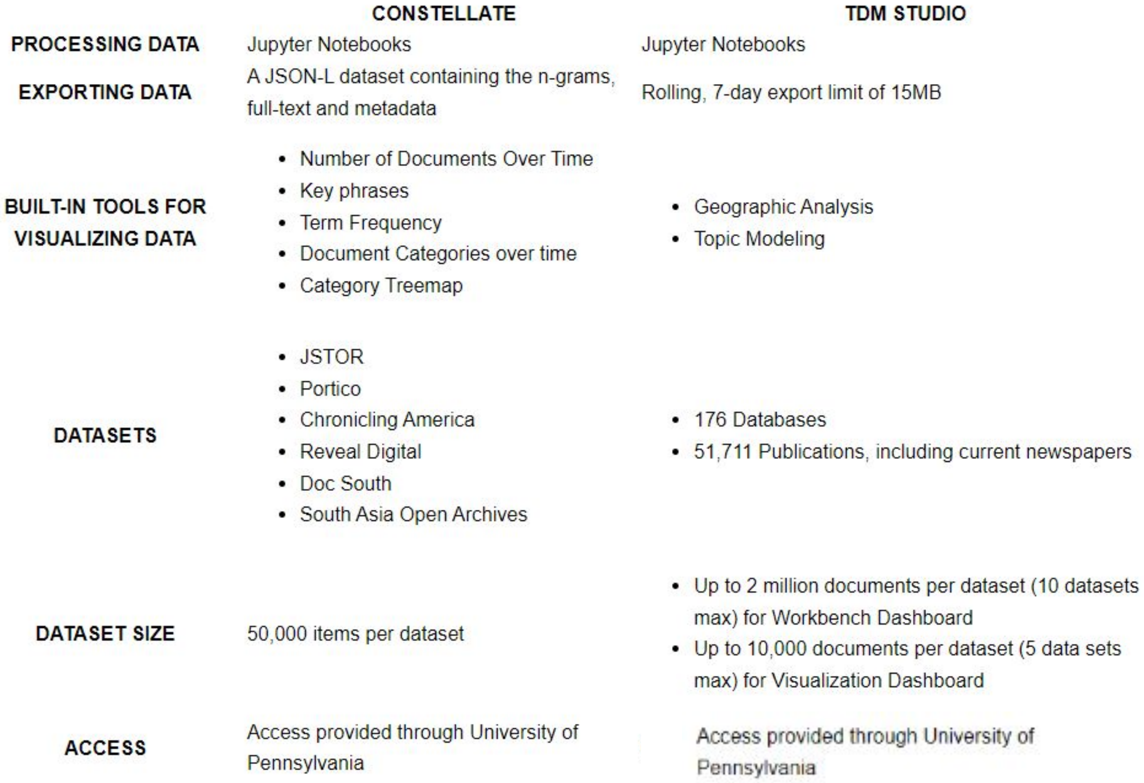


Analyzing and Visualizing Text with Constellate and TDM Studio

RESEARCH DATA & DIGITAL SCHOLARSHIP
DATA
JAM

https://old.library.upenn.edu/blogs/rdds/tools/analyzing-and-visualizing-text-constellate-and-proquest-tdm-studio

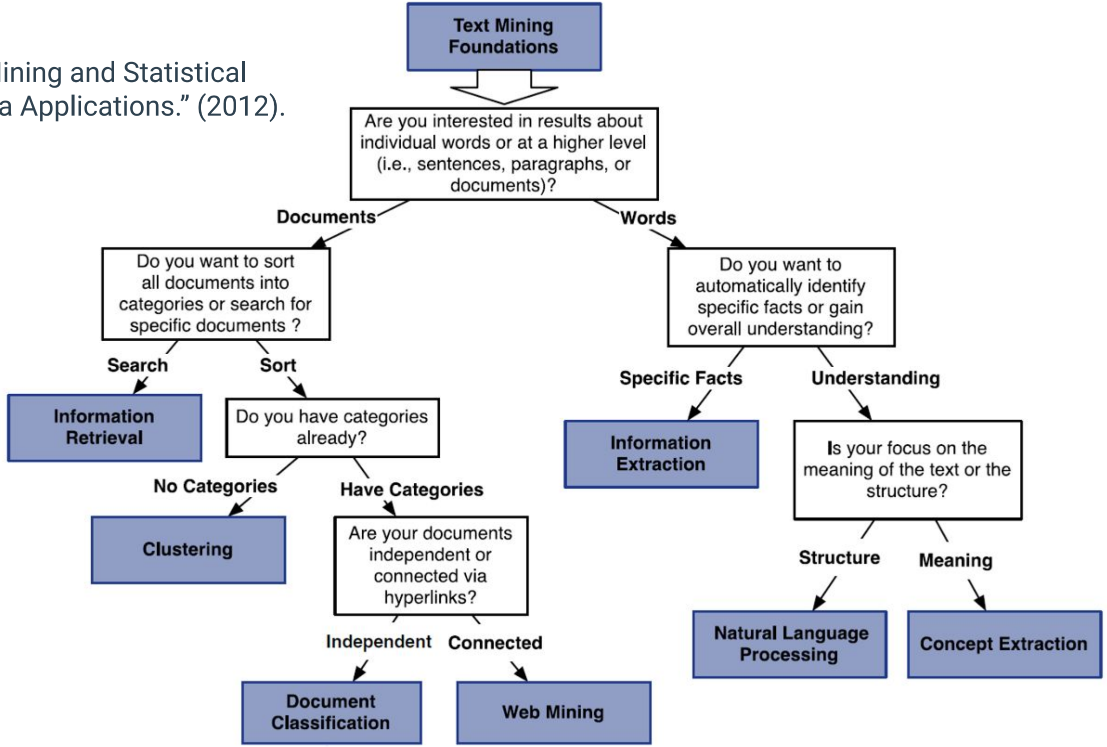
Text Data Sources - TDM Platforms that require some coding



	CONSTELLATE	TDM STUDIO
PROCESSING DATA	Jupyter Notebooks	Jupyter Notebooks
EXPORTING DATA	A JSON-L dataset containing the n-grams, full-text and metadata	Rolling, 7-day export limit of 15MB
BUILT-IN TOOLS FOR VISUALIZING DATA	• Number of Documents Over Time • Key phrases • Term Frequency • Document Categories over time • Category Treemap	• Geographic Analysis • Topic Modeling
DATASETS	• JSTOR • Portico • Chronicling America • Reveal Digital • Doc South • South Asia Open Archives	• 176 Databases • 51,711 Publications, including current newspapers
DATASET SIZE	50,000 items per dataset	• Up to 2 million documents per dataset (10 datasets max) for Workbench Dashboard • Up to 10,000 documents per dataset (5 data sets max) for Visualization Dashboard
ACCESS	Access provided through University of Pennsylvania	Access provided through University of Pennsylvania

When to Use Computational Text Analysis Techniques?

Miner, Gary D. et al. “Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.” (2012).



Text Mining Foundations

Are you interested in results about individual words or at a higher level (i.e., sentences, paragraphs, or documents)?

Documents

Do you want to sort all documents into categories or search for specific documents ?

Search

Information Retrieval

Sort

Do you have categories already?

No Categories

Clustering

Have Categories

Are your documents independent or connected via hyperlinks?

Independent

Document Classification

Connected

Web Mining

Words

Do you want to automatically identify specific facts or gain overall understanding?

Specific Facts

Information Extraction

Understanding

Is your focus on the meaning of the text or the structure?

Structure

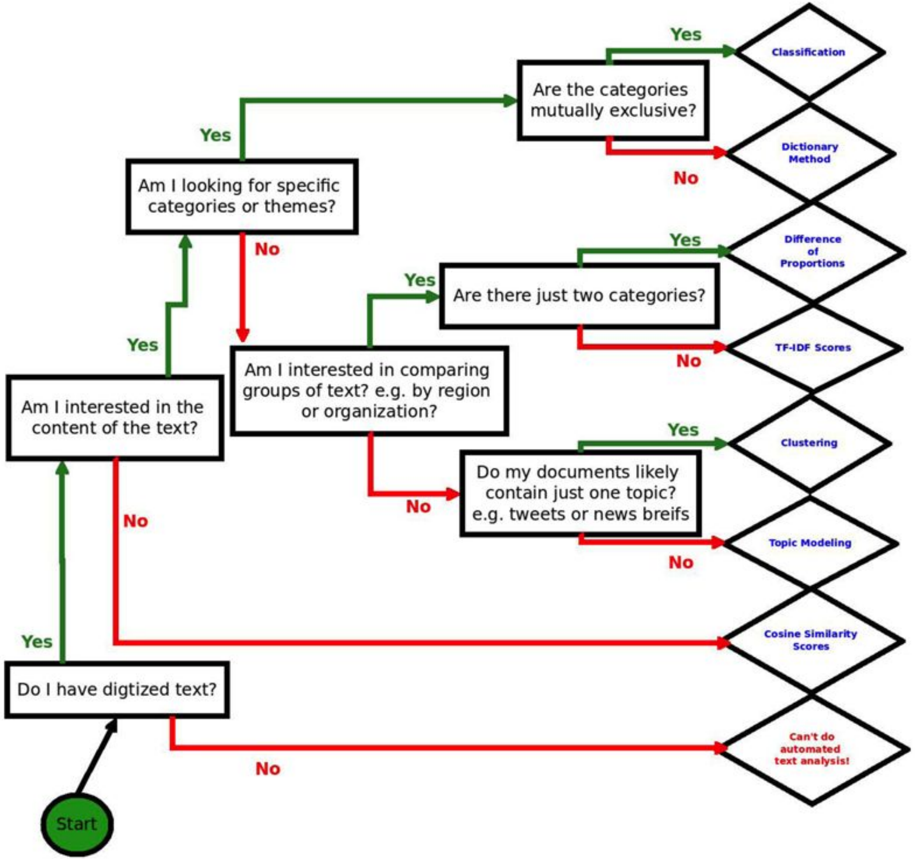
Natural Language Processing

Meaning

Concept Extraction

Data is Available, even if Messy + Longer Time + Experienced with Coding

When to Use Computational Text Analysis Techniques?

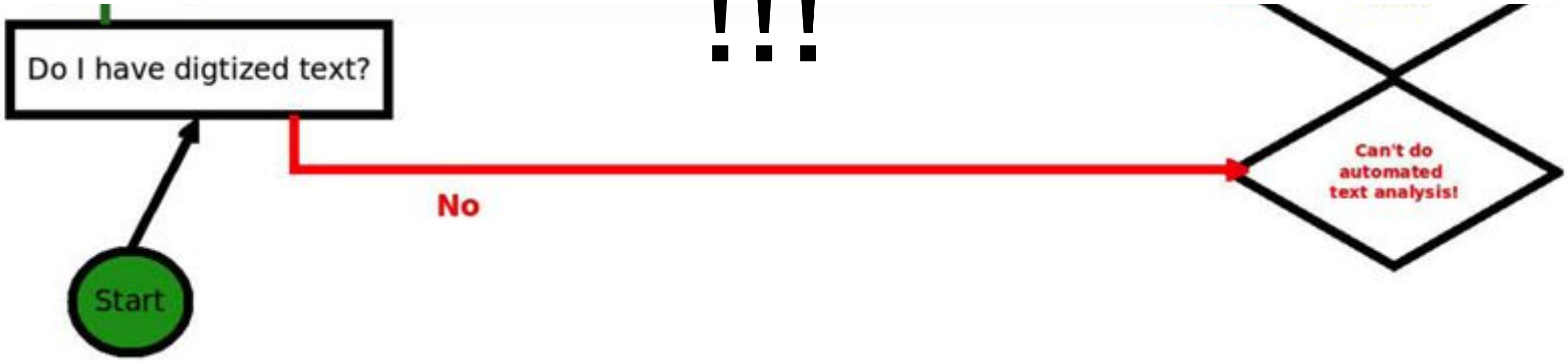

Start
Do I have digtized text?
Yes
No
Am I interested in the content of the text?
Yes
No
Am I looking for specific categories or themes?
Yes
No
Are the categories mutually exclusive?
Yes
No
Am I interested in comparing groups of text? e.g. by region or organization?
Yes
No
Are there just two categories?
Yes
No
Do my documents likely contain just one topic? e.g. tweets or news breifs
Yes
No
Classification
Dictionary Method
Difference of Proportions
TF-IDF Scores
Clustering
Topic Modeling
Cosine Similarity Scores
Can't do automated text analysis!

Method Used Must be **Specific** to your Research Question.

From DH@Berkeley

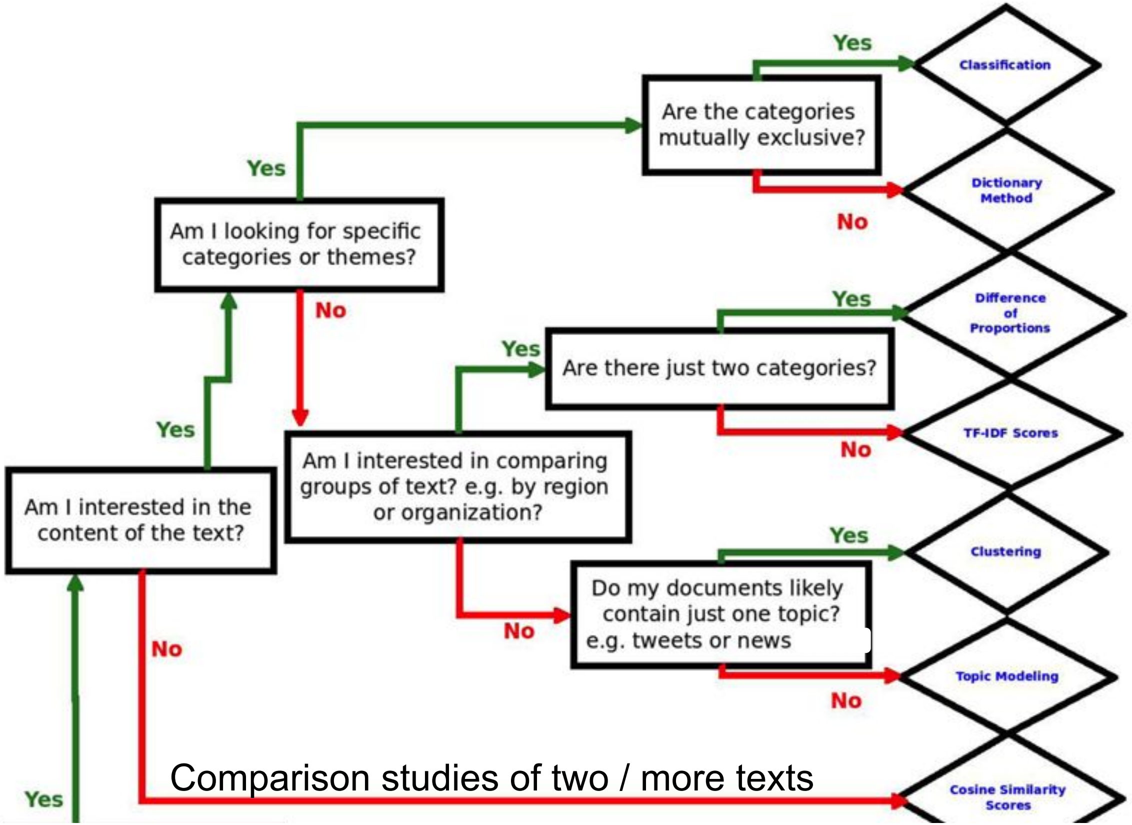
When to Use Computational Text Analysis Techniques?

!!!



From DH@Berkeley

When to Use Computational Text Analysis Techniques?


Am I interested in the content of the text?
Yes
No
Am I looking for specific categories or themes?
Yes
No
Are the categories mutually exclusive?
Yes
Classification
No
Dictionary Method
Am I interested in comparing groups of text? e.g. by region or organization?
Yes
Are there just two categories?
Yes
Difference of Proportions
No
TF-IDF Scores
No
Do my documents likely contain just one topic? e.g. tweets or news
Yes
Clustering
No
Topic Modeling
Cosine Similarity Scores
Yes

Classification for text documents

Comparison studies of two / more texts

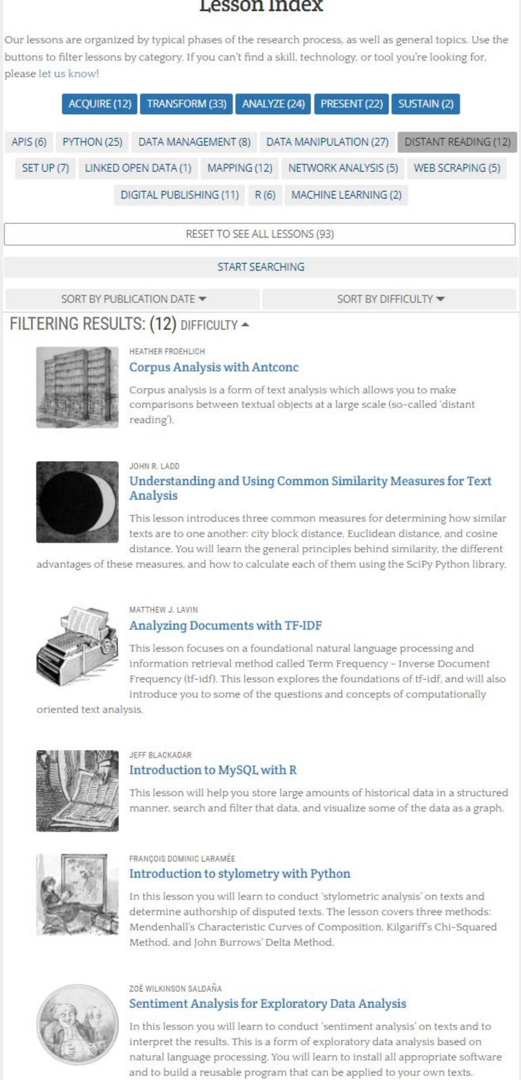
From DH@Berkeley

When to Use Computational Text Analysis Techniques?

- Am I interested in the content of the text for stylistic choices made in the text?
  a. Sentiment Analysis on Twitter Data mining: find out how people think and feel on a macro scale
  b. Authorship Attribution / Stylometric Analysis techniques: resolve undecided authorship of literary texts computationally

ProgrammingHistorian

https://programminghistorian.org/en/lessons/?topic=distant-reading



Lesson Index

Our lessons are organized by typical phases of the research process, as well as general topics. Use the buttons to filter lessons by category. If you can't find a skill, technology, or tool you're looking for, please let us know!

FILTERING RESULTS: (12) DIFFICULTY ▴

HEATHER FROEHLICH

Corpus Analysis with Antconc

Corpus analysis is a form of text analysis which allows you to make comparisons between textual objects at a large scale (so-called 'distant reading').

JOHN R. LADD

Understanding and Using Common Similarity Measures for Text Analysis

This lesson introduces three common measures for determining how similar texts are to one another: city block distance, Euclidean distance, and cosine distance. You will learn the general principles behind similarity, the different advantages of these measures, and how to calculate each of them using the SciPy Python library.

MATTHEW J. LAVIN

Analyzing Documents with TF-IDF

This lesson focuses on a foundational natural language processing and information retrieval method called Term Frequency - Inverse Document Frequency (tf-idf). This lesson explores the foundations of tf-idf, and will also introduce you to some of the questions and concepts of computationally oriented text analysis.

JEFF BLACKADAR

Introduction to MySQL with R

This lesson will help you store large amounts of historical data in a structured manner, search and filter that data, and visualize some of the data as a graph.

FRANÇOIS DOMINIC LARAMÉE

Introduction to stylometry with Python

In this lesson you will learn to conduct 'stylometric analysis' on texts and determine authorship of disputed texts. The lesson covers three methods: Mendenhall's Characteristic Curves of Composition, Kilgariff's Chi-Squared Method, and John Burrows' Delta Method.

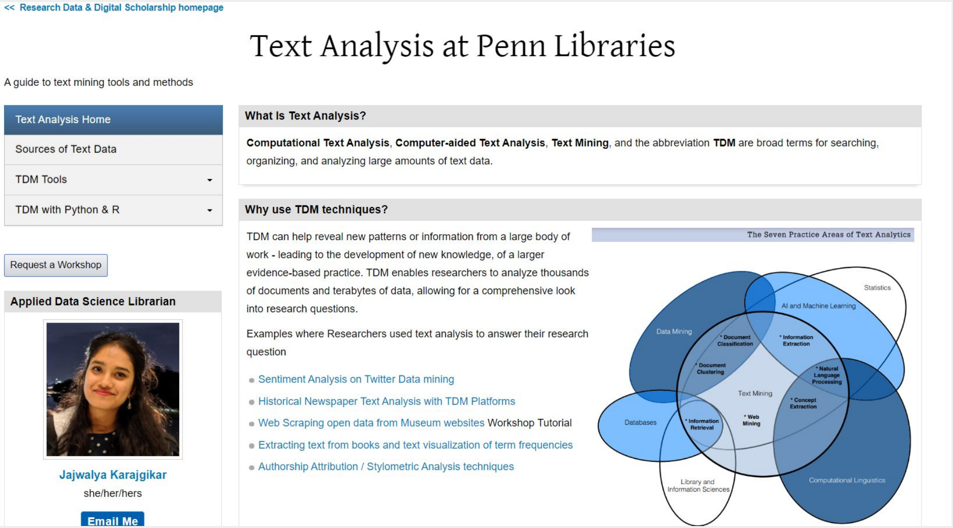
ZOË WILKINSON SALDAÑA

Sentiment Analysis for Exploratory Data Analysis

In this lesson you will learn to conduct 'sentiment analysis' on texts and to interpret the results. This is a form of exploratory data analysis based on natural language processing. You will learn to install all appropriate software and to build a reusable program that can be applied to your own texts.

Library Resources

https://guides.library.upenn.edu/penntdm



<< Research Data & Digital Scholarship homepage

Text Analysis at Penn Libraries

A guide to text mining tools and methods

- Text Analysis Home
- Sources of Text Data
- TDM Tools
- TDM with Python & R

Request a Workshop

Applied Data Science Librarian

Jajwalya Karajgikar

she/her/hers

Email Me

What Is Text Analysis?

Computational Text Analysis, **Computer-aided Text Analysis**, **Text Mining**, and the abbreviation **TDM** are broad terms for searching, organizing, and analyzing large amounts of text data.

Why use TDM techniques?

TDM can help reveal new patterns or information from a large body of work - leading to the development of new knowledge, of a larger evidence-based practice. TDM enables researchers to analyze thousands of documents and terabytes of data, allowing for a comprehensive look into research questions.

Examples where Researchers used text analysis to answer their research question

- Sentiment Analysis on Twitter Data mining
- Historical Newspaper Text Analysis with TDM Platforms
- Web Scraping open data from Museum websites Workshop Tutorial
- Extracting text from books and text visualization of term frequencies
- Authorship Attribution / Stylometric Analysis techniques

Feedback form:

upenn.libwizard.com/f/rdds-survey