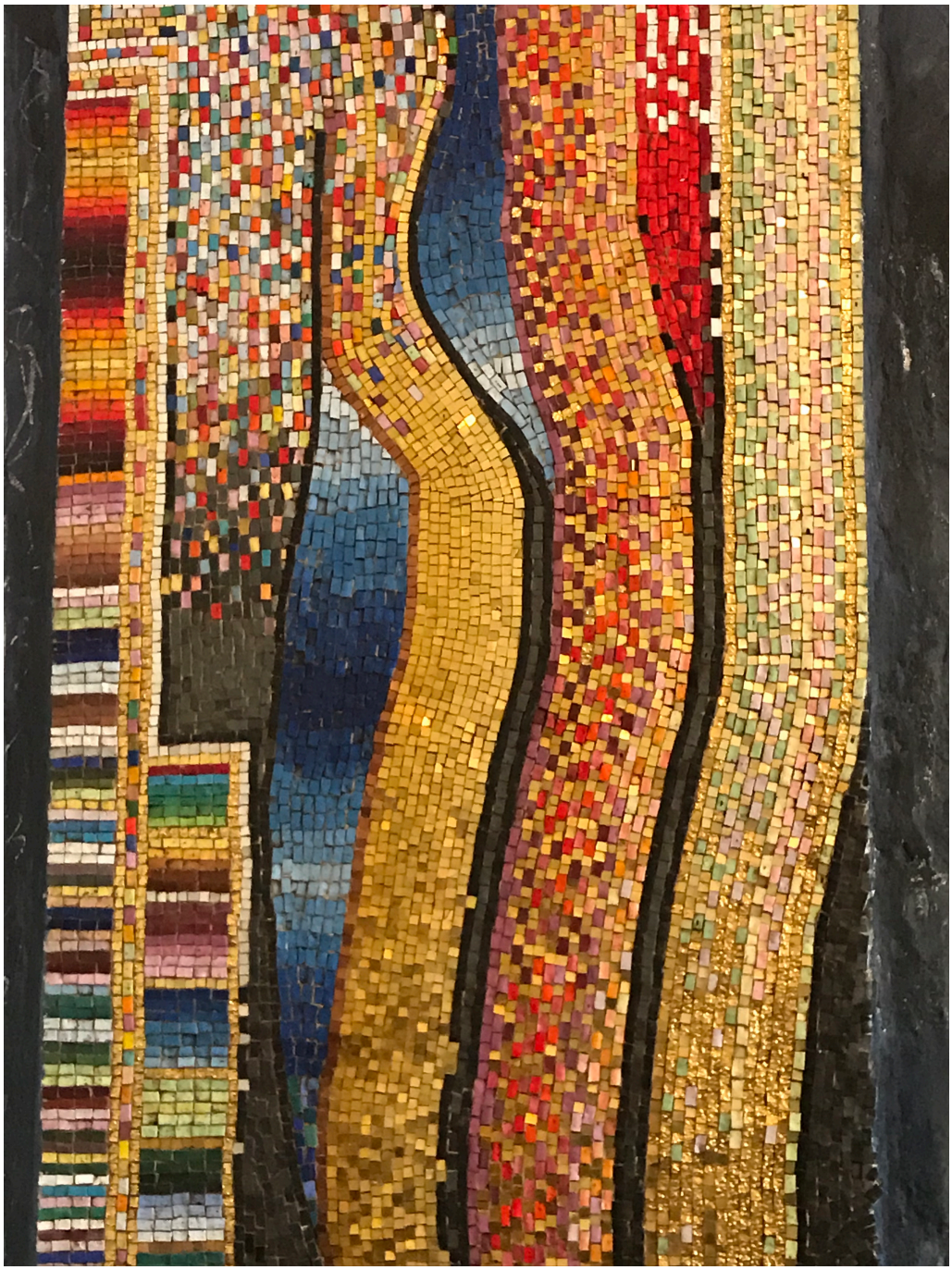
Digital Humanities in Practice

Winter 2020: Course Handbook



Dr. Sarah L. Ketchley
Near Eastern Languages & Civilization
University of Washington

A note about content:

This course book was developed for the 3-credit 'Digital Humanities in Practice' course offered at the University of Washington during the Winter quarter 2020. The course was taught through the Near Eastern Languages and Civilization Department, and the datasets used for teaching and analysis are therefore primarily related to Nile travel and archaeological excavation in the later 19th and early 20th centuries. This course could be adapted to use datasets from other disciplines without a great deal of difficulty.

PDF course content in this document was released on a week-by-week basis following the syllabus outline, and was intended to provide tool or method overviews, reading lists, written instructions, and sample projects for students to use as guidance. The class met twice a week for 1hr 20mins each session. In person work included discussions of reading, live demos and hands on work. Weekly assignments were primarily discussion posts and replies; students provided reading summaries, answered topic questions, and kept a full written record of their work to underscore the importance of maintaining documentation.

Acknowledgements and thanks:

Course content has been developed and has evolved over a number of years and across multiple sessions of an 'Introduction to Digital Humanities'. Many people have generously contributed expertise, advice, code, content, tutorials, transcriptions and support. I am profoundly grateful to everyone.

Special thanks to Walter G. Andrews, Elisa Beshero-Bondar, Marc Cormier, Lindsey Gervais, Brad Holland, Selim Kuru, Matt Milner, Paige Morgan, Wendy Perla Kurtz, Calvin Scott Paulson, Margaret Waligora, Helene Williams & all the Emma B. Andrews Diary Project student interns who have worked diligently to transcribe reams of primary source documents over the past 9 years.

Archival resources include material from the following institutions and individuals: American Philosophical Society; Griffith Institute, Oxford; Massachusetts Historical Society; Metropolitan Museum of Art; Oregon Historical Society; Lady Ellen Strathnaver and the Newberry family.

Sarah Ketchley, March 2020

Table of Contents

WEEK 1a: Introduction

Overview

We'll take the first session to work through various class housekeeping tasks.

1. Introductions

2. Syllabus review

WEEK	DATE	TOPIC
1	Monday January 6	Welcome! Syllabus review & orientation
	Wednesday January 8	What is Digital Humanities? Exploring DH projects
2	Monday January 13	Introducing the primary sources; sourcing content
	Wednesday January 15	Palaeography, transcribing and text as data
3	Monday January 20	NO CLASS: MLK DAY
	Wednesday January 22	What is text mining? Practical approaches to distant reading
4	Monday January 27	Planning and managing digital projects
	Wednesday January 29	Working with digital archives, creating OCR text, intro to Digital Scholar Lab
5	Monday February 3	Corpus building/preparing texts for analysis: Digital Scholar Lab & Lexos
	Wednesday February 5	Corpus building/preparing texts for analysis: Digital Scholar Lab, Lexos & other options.
6	Monday February 10	On microhistory; introducing the Text Encoding Initiative (TEI)
	Wednesday February 12	Quantitative or qualitative? Considering which digital tools to use through the lens of sample projects.
7	Monday February 17	NO CLASS: PRESIDENTS' DAY
	Wednesday February 19	Ngrams & research interviews
8	Monday February 24	Sentiment Analysis
	Wednesday February 26	Named Entity Recognition & mapping
9	Monday March 2	Topic modeling, clustering & a sample project; Timelines & StoryMaps
	Wednesday March 4	Hands on Lab work
10	Monday March 9	Hands on Lab work
	Wednesday March 11	Presentations
	Friday March 13	Final Projects Due

Digital Humanities in Practice (NELC 296 B)—a no-prerequisite course to introduce students to concepts and methodologies of using digital humanities tools for dataset creation, analysis and presentation. These skills are essential for humanities and social science majors to develop in an increasingly competitive job market. Students will explore primary source material related to the lives and achievements of early pioneers in Near Eastern archaeology, focusing specifically on the period known as the 'Golden Age' of Egyptology at the end of the 19th and early 20th centuries.

Students will analyze primary source documents using text mining methodologies, build digital maps and timelines, and ultimately present research results on an online platform.

3. Evaluation

Assignments in the course will receive these weights. For collaborative work, your grade will be determined both by your individual contribution and by the overall quality of the project.

- 10% In Class Participation

- Attendance recorded by instructor in 'Roll Call Attendance' in Canvas.

- Participation in discussion and group work, reference to course readings in discussion. Grade assigned based on observations of instructor.

- 30% Weekly Worklog/Discussion Post

    - 9 weekly worklogs posted on the class Discussion Board, plus three replies to classmates' post.

A detailed weekly worklog is due by 11.59pm on Sunday of each week, prior to our class meeting on Mondays. Replies to three of your classmates is due by Tuesday of each week by 11.59pm.

Rubric:

Initial Post: submitted by Sunday deadline	15
Initial Post: addresses all parts of the question	15
Initial Post: 200-300 words in length	10
Initial Post: includes references to at least 3 reading or video resources	15
Initial Post: has been proofread and uses Standard English appropriate for college-level writing with few or no errors.	10
Reply Posts: three replies made by Tuesday deadline	15

Reply Posts: each reply further develops classmates' posts and ideas or offers additional analysis.	10
Reply Posts: concrete examples are used when appropriate to support statements.	10
Total points possible	**100**

Late weekly worklogs will lose 3 points/day.

- **60% Final Project & Presentation**
  - completion and presentation of final project with documentation.

Rubric:

Title slide: a high-level overview of topic and describes research question.	15
1.-6. Narrative - in at least SIX slides describe your research topic, highlighting the main points, or sequences of events. Over the six slides (at minimum), summarize the main points of your research narrative. For example, for a topic related to 'sea serpents' in history, I might describe earliest recorded sightings, descriptions, reactions etc. Include images taken from Gale Primary Source material or other open source resources, with appropriate citation. Supplemental material can include video/website content. Your tone should be engaging and informative.	40
7. Archival and research resources This is where you will give an overview of the material that has informed your research. This will include original primary source documents, Gale Primary Source resources and any open source material you have used in building your project.	15

8. Data curation and cleaning You'll include a summary of the steps you took to curate and clean your data here, using your work log and discussion posts as reference points. Include: 1. A summary of the quality of the OCR you worked with, including OCR confidence levels in the DSL, and discussions of any specific challenges you encountered. 2. A description of the strategies you used to curate your content set - what search terms did you use? A description of the other material you sourced, as relevant. 3. A description of your cleaning process. How effective was it? What specific steps did you take to clean your data? How did cleaning in the DSL differ from using, say, Lexos?	20
9. Analysis summary Describe the analyses you carried out: a) in the Digital Scholar Lab b) using external tools. In both cases explain which specific tool(s) you identified as being most useful for answering your research question, and why this was. You'll go into more detail in the next slides.	20
10.-11. Visualization and Analysis Result 1-2 You should have at least one, preferably two visualizations and associated discussion. Give a detailed description of the specific analysis tool you used to answer your research question, and the types of questions the tool is best suited to answer (based on the reading you've completed for this class). Include your downloaded visualization. 1. If using a DSL visualization, what configuration options did you apply? If using an external tool, describe the setup you used in generating the visualization. 2. Discuss the analysis results in your visualization. How well do they answer your research question? Are the useful and/or intelligible? How much additional data curation did you have to do to add meaning to your analysis?	40

12. Learning Summary What have found most surprising, significant or unexpected in your text mining investigations? What have your learned in this class? What additional questions do you have?	10
13. Bibliography	10
Images including credits and a caption	10
Standard English and proofreading	10
Map background - some attempt has been made to include an appropriate map background with relevant points marked in each slide. This may not be a traditional location point since not all projects are well-suited to mapping place.	10
Total points possible	**200**

4. Reading and viewing

No book purchase is necessary. Our readings will mostly be from online, open source documents. Links to these will be posted in the weekly assignment list; feel free to read them online, download them to your computer or print them off so you can annotate them.

Each week, you will be assigned a set of articles to read, and one or two short videos to view. These may be lectures or tutorials recorded for this class, or material sourced from elsewhere which is appropriate for this introductory course. The expectation is that you are an active reader and viewer, taking notes on points that you find most salient and informative. You will be posting your notes and insights in your weekly work log, with appropriate citation, and this material is an important opportunity to demonstrate your understanding and generate active discussions with your peers and instructor.

5. Student outcomes/learning goals

Humanities and social sciences students will become familiar with a range of tools and technologies for text mining and text analysis that will enhance their abilities to succeed both as undergraduate researchers and in their lives after graduation. Students in technology disciplines will be able to explore the applications of digital tools to humanistic endeavors.

Students in this course will:

- Learn the basic vocabulary of concepts and tools in digital humanities and become

acquainted with a range of projects, best practices and resources in the field.

- Gain hands-on experience of humanities dataset creation, curation, analysis and presentation.
- Gain an introductory knowledge of many open source digital tools or methods useful to broad humanities disciplines.
- Create a digital narrative to present the results of work in (2) .

6. <u>Student and instructor expectations</u>

Readings/Lectures: it's important to keep up with course reading and video material, as you will be discussing them with your classmates each week. They will provide a valuable theoretical and practical framework as you begin to work with digital humanities tools.

Other expectations:

- **Creativity:** there's a lot of room for curiosity and creativity in digital humanities, and there is no one "right" answer. This course is a place to explore connections between content and technology.
- **Failure:** You will be working with disciplinary content and technology that is unfamiliar to you, and at times you may struggle with analysis tools or your research material. Often the most valuable learning happens during this time! I will score your work in this course primarily on process rather than on the final product. Showing us something that doesn't work quite like you want/expect, and explaining your steps and what your goal is, indicates a level of engagement and curiosity we are all striving towards.
- **Respect:** You'll be interacting with your classmates in person and in the online environment, primarily on the discussion board in posts and replies. The expectation is that interactions will be respectful, kind and constructive at all times.
- **Feedback:** I expect to hear from you if something is going well, or less than well, at any point in the quarter. You can expect me to provide feedback on submitted work in a timely manner (within two weeks of submission). I am available to talk via Skype/Zoom/phone throughout the week, including weekends outside of class, so feel free to make an appointment to chat about course content, processes, concerns etc.

Precarity: Any student facing housing, food, or health challenges which they believe will affect their performance in this course is urged to contact me or the NELC Advisor's office for support and accommodation. One of the best strategies you can have is to tackle issues before they become a crisis: it's OK to ask for assistance!

WEEK 1b: What is DH?

We'll examine what the term 'digital humanities' means (hint: there's no single definition!), and you'll carry these thoughts with you as you work through the process of creating a corpus of data, curating it and then analyzing it. You'll present the results of your work by building a digital narrative as your final project.

Reading/Viewing

Paige Morgan, 'What Digital Humanists Do', Demystifying Digital Humanities workshops, 2013

Maria Popova, Digital Humanities Spotlight: 7 important digitization projects, 2011

Melissa Dinsman, The Digital in the Humanities, LA Review of Books (read a selection of interviews)

Refresh https://whatisdigitalhumanities.com/to see a range of definitions.

Lecture One

www.youtube.com/embed/Xu6Z1SoEZcc

Panel at Columbia University, 2011.

Panelists:

Daniel J. Cohen, Assoc. Professor of History and Director of the Center for History and New Media (CHNM) at George Mason University.

Federica Frabetti, Senior Lecturer in the Communication, Media, and Culture Program at Oxford Brookes University.

Dino Buzzetti recently retired from the Dept. of Philosophy at the University of Bologna.

The transcript is available on YouTube. Please focus on Cohen's section of the presentation.

Lecture Two

www.youtube.com/embed/LF8duSp2geo

A brief history and overview of the discipline of digital humanities.

Evaluating DH Projects

Before class on Wednesday, review the following:

Miriam Posner, 'How did they they make that?', blog post, 2013

Watch Miriam Posner's lecture How did they make that?

With the proliferation of digital humanities projects, tool building and research in recent years, it’s often hard to establish a simple yet comprehensive set of criteria for evaluating and assessing digital scholarly work. You've now read a number of articles exploring potential methods and standards for evaluation. You've also watched Miriam Posner's video 'How did they make that?'

Now it's your turn to assess a selection of projects, and post your findings to the Discussion Board by the Sunday of Week 1 at 11.59pm, and then respond to three of your classmates' posts by the Tuesday of the second week at 11.59pm.

You'll begin your evaluation by considering:

SOURCES what is the collection and purpose of site?

PROCESSING i.e. how are the sources made machine readable?

PRESENTATION

- Is there an intentional and appropriate organization of information?
- Does the project use accepted standards for web design, metadata, and encoding?
- Is there a thoughtful balance between design, content, and medium?
- Overall, how intuitive is it to navigate the website?

Extend your evaluation to consider the following:

- Issues of community, scholarship, digital infrastructure, values embodied in the language, practices, and organization of the component parts.
- Documentation describing the project's data management plan, and plans for **long term sustainability and digital preservation.**
- Who is the project team? Who are the funding partners? Is this a collaborative effort between institutions or individuals?
- What are the fields of expertise of its members? If you're technically-minded, maybe you're looking for information about the technical methods and standards applied. Can you find this? How about the code?

DOCUMENTATION

Remember you are evaluating a scholarly resource. The ideal digital resource should keep documentation and make it available from the project website, making clear the extent, provenance and selection methods of materials for the resource. For textual projects, there should be a description of the editing protocols, for example. **Why do you think it's important to have this type of information? Do you find it all websites?**

Here is a SAMPLE EVALUATION to give you an idea of the type of information to include:

1. **Project Name & URL**

Google Ancient Places https://googleancientplaces.wordpress.com/ (Links to an external site.)

2. **Purpose of Site**

GAP uses ancient world places as target information to find and visualize while solving the problem of discovery and usability.

3. **Sources** *The ideal digital resource should keep documentation and make it available from the project website, making clear the extent, provenance and selection methods of materials for the resource*

Hyperlinks of references/sources are attached to the relevant blog posts. The books which are geotagged are listed on the site that the data visualization is found on. The extent of their usage is clear (scanned for geographical references) and while unstated, the likely method of selection was the likelihood of the books to contain useful geographical references.

4. **Processing** *i.e. how are the sources made machine readable?*

Raw text from various texts that reference historical places is fed into Edinburgh Geoparser to categorize and georesolve.

5. **Presentation** *Is there an intentional and appropriate organization of information? Does the project use accepted standards for web design, metadata, and encoding? Is there a thoughtful balance between design, content, and medium? Overall, how intuitive is it to navigate the website?*

The organization of the information as described in text seems to be intentional and effective, with clear meaning and accessibility to tags. However, the presentation itself does not publicly function.

6. **Design/Functionality/Accessibility** *The ideal digital resource should be designed for a wide variety of users, and include information to help the non-expert to understand the resource and use its contents*

The GapVis interface was their medium for visualization, but it's not functional as it doesn't load. It may need different tests on browsers. The tabs (Home, About, GapVis, etc.) were helpful towards the design though. Overall, it did seem like a lot of information to digest and read through, so design could be improved.

7. **Documentation** *The ideal digital resource should keep documentation and make it available from the project website, making clear the extent, provenance and selection methods of materials*

for the resource. For textual projects, there should be a description of the editing protocols, for example.

The sources are listed and thoroughly documented, as is the extent and reason of their usage. Some details of the data usage is not present (such as whether input into the database is programmatic or manual). While the preliminary sources are documented, the data after being filtered to any degree is no longer available, so details are unavailable on documentation within the visualization.

8. **Data Management** *Is there a plan describing the project's data management plan?*

They used the Pelagios system, which describes machine identification of certain tokens in 9 books. They plan on having data being public domain.

9. **Sustainability** *What are the project's plans for long term sustainability and digital preservation?*

GAP has formally ensured that the digital information of continuing value remains accessible and usable. This has been achieved by using a public domain and taking references from Google books etc.

10. **Project and Funding Partners** *Who is the project team? Who are the funding partners? Is this a collaborative effort between institutions or individuals? What are the fields of expertise of its members?*

GAP is run by a variety of researchers from a range of institutions (Independent study, University of Southampton and so on).

Discussion Post

Due Week 1 Sunday 11.59pm

1. Referring to at least 3 items from the introductory reading and viewing you have completed, describe 2-3 interesting or unexpected aspects of digital humanities that you would like to explore further:

Write at least 200-300 words, and include appropriate citation for the material you reference.

2. The 'Evaluating DH Projects' overview and in-class work provide the framework for this Discussion Post.

Choose THREE projects to evaluate from this list:

Mark Twain Project

Pleiades

Perseus Digital Library

Nomisma

Regnum Francorum Online

Willa Cather Archive

Inscriptions of Israel/Palestine

Papyri.info

Digital Mitford

Darwin Online

Moore Archive

CLAROS

The Griffith Institute

Arachne

Railways and the making of modern America

TIMEA

Walt Whitman Archive

Melville's Marginalia

Meketre

OCRE

Europeana

ORBIS

MJBC

ISAW Papers

Shelley Godwin Archive

Archives Unleashed

SAWS

Trismegistos

Ancient World Mapping Center

You can cut and paste the blank template below into your discussion post and complete the details.

1. Project URL

2. Purpose of Site

3. Sources *The ideal digital resource should keep documentation and make it available from the project website, making clear the extent, provenance and selection methods of materials for the resource*

4. Processing *i.e. how are the sources made machine readable?*

5. Presentation *Is there an intentional and appropriate organization of information? Does the project use accepted standards for web design, metadata, and encoding? Is there a thoughtful balance between design, content, and medium? Overall, how intuitive is it to navigate the website?*

6. **Design/Functionality/Accessibility** *The ideal digital resource should be designed for a wide variety of users, and include information to help the non-expert to understand the resource and use its contents*

7. **Documentation** *The ideal digital resource should keep documentation and make it available from the project website, making clear the extent, provenance and selection methods of materials for the resource. For textual projects, there should be a description of the editing protocols, for example.*

8. **Data Management** *Is there a plan describing the project's data management plan?*

9. **Sustainability** *What are the project's plans for long term sustainability and digital preservation?*

10. **Project and Funding Partners** *Who is the project team? Who are the funding partners? Is this a collaborative effort between institutions or individuals? What are the fields of expertise of its members?*

REPLIES

Due Tuesday of Week 2 11.59pm

Aim to write 100-150 words per reply.

Choose THREE of your classmates' evaluations, and explore the digital project they have referenced in their discussion. Do you agree with their assessments? What additional analysis can you contribute in each case?

WEEK 2a: Primary Sources & Research Data

Primary Sources

This is an overview of the archival material we are currently working on. Our workflow to date has been to transcribe from the original into a plaintext document. These transcriptions are checked for accuracy before text encoding begins.

Additional datasets can be found on the Griffith Institute website (start with Minnie Burton's diaries), and the Egypt Exploration Society's main website and Flickr pages.

PRELIMINARY DATA FOR THIS DH CLASS:

Diaries of Mrs. Emma B. Andrews

19 volumes of Nile travel journal documenting travel, society and excavation in the Valley of the Kings.

PDF	TEXT

Correspondence of Helen Winlock & family

Archive of the wife of Herbert Winlock, Director of the Metropolitan Museum's Egyptian Expedition. Descriptions of society, life and archaeology.

JPG	TEXT 1 | 2

Mary Buttles Newberry journal & letters

A family relative of Theodore Davis's, Mary B. Newberry joined Davis and Emma Andrews during the 1912-1913 season. Her handwritten letters were transcribed by a family member a

number of years ago. The TXT file shows an example of OCR created from the PDF with many messy characters, as yet not cleaned up.

PDF	TEXT

Diaries of Joseph Lindon Smith

Artist who travelled with his family in Egypt for many years. We are working two diary volumes 1904-05 and 1906-07 which cover the time he recorded Theodore Davis's discoveries in the Valley of the Kings.

PDF & TEXT 1905-6 | 1906-7

Joseph Lindon Smith Correspondence & Misc. Material

A selection of Egypt-related material, mostly letters, written by JLS and some members of his family. We are currently (January 2020) working on material between 1898-1909. The original archive is in the American Archives of Art, DC.

View our current list of letters and research tracker here.

PDF	TEXT

Archive of Lindsley Foote Hall

This is my most recent find, in Portland, OR. Hall was an MIT-trained draughtsman who worked on many project in Egypt including Tutankhamun's tomb. He was initially interviewed and hired to work in Egypt by Professor Chandler, who is Helen Winlock's father. He worked in Egypt under the supervision of Herbert Winlock. Scanned diaries range from 1901-1922.

PDF	TEXT 1919 |1920 |1921 | 1922

ADDITIONAL ARCHIVAL MATERIAL:

Percy Newberry Correspondence

Percy Newberry (no relation to Mary) is one of the 'giants' of British Egyptology. Some of his early work in Egypt was funded by Emma Andrews and Theodore Davis. Newberry wrote extensively about Egyptology, and his archive is now in the Griffith Institute in Oxford. the TIFF/JPEG files are of typewritten and handwritten material.

TIFF/JPG	TEXT

Howard Carter Notes

Egyptologist who discovered the tomb of Tutankhamun in 1922. Carter was the first Egyptologist/ archaeologist to work with Theodore Davis when he was granted the concession to work in the Valley of the Kings at the turn of the century. This material comprises typewritten notecards, drawings and handwritten material with Carter's notes about these excavations. Carter intended to compile them into a book, but never completed it.

TIFF/JPG	TEXT

Thomas Cook Travel Guides

Thomas Cook & Son were granted the right to run Egyptian government steamers from Cairo to Aswan in 1870, and into the Sudan by 1874, giving the company exclusive rights to carry mail along the Nile. These Cooks steamers also carried tourists along the Nile - a cheap and speedy way to see the sights. The European middle classes started swarming to Egypt, which was advertised as the ideal destination for travelers looking to escape the dreary winters or to improve their health. The archival material in this collection comes from the company archive in Peterborough, U.K. It comprises sets of travel guides written for Cook's tourists, as well as a collection of passenger lists. A more general set of guides is also in the folder ('Tourist Guides') including the world-renowned Baedeker Guides,

PDF	TEXT

Published Egyptian Travel Memoirs

An ongoing compilation of Egypt travel memoirs written in the mid- to late-19th and early 20th centuries. The starting point for this work is Kalfatovic's Nile Notes of a Howadji. A Bibliography of Travelers' Tales from Egypt, from the Earliest Time to 1918 (Metutchen, N.J. & London, 1992). Books have been sourced primarily from Google Books, the Hathi Trust and the Internet Archive.

PDF	TEXT

The Egyptian Gazette

Newspaper published in Alexandria between. I have copies from October - December 1900. Originals are in a moth-eaten archive in Cairo, and in the British Library. Future work will include working from microfiche to make PDF copies of the remaining run of newspapers.

PDF	TEXT

Research Databases

Ancestry

http://offcampus.lib.washington.edu/login?url=http://ancestrylibrary.proquest.com/aleweb/ale/do/login

Genealogical records including census, military, immigration, marriage, yearbooks, biographical dictionaries, birth records.

American National Biography

http://offcampus.lib.washington.edu/login?url=https://www.anb.org/

Standard biographical dictionary for Americans in all fields. Only deceased are included.

Periodicals Archive Collection

http://offcampus.lib.washington.edu/login?url=https://search.proquest.com/pao?accountid=14784

Collection of back issues of American, British and European magazines and journals in the arts, humanities & social sciences.

Gale Digital Scholar Lab (with Gale Primary Sources/GPS Archives - more details below)

https://infotrac.gale.com/itweb/dslabwa?db=DSLAB

Pw: uwash

- create your account with Google or Microsoft credentials. NOTE: this is a unique instance for this class, with all Gale Primary Sources activated.

Gale Primary Sources: An Overview

The data you have access to comes from a selection of Gale Primary Sources archives detailed below. For a brief summary of each Archive's content, you can download the 2019 catalogue here: https://www.gale.com/intl/primary-sources

17th and 18th Century Burney Collection
17th and 18th Century Nichols Newspapers Collection
19th Century UK Periodicals
American Amateur Newspapers
American Civil Liberties Union Papers, 1912-1990
American Fiction
American Historical Periodicals
Archives Unbound
Archives of Sexuality & Gender
Associated Press Collections Online
Brazilian and Portuguese History and Culture

British Library Newspapers
China and the Modern World
Crime, Punishment, and Popular Culture 1790-1920
Daily Mail Historical Archive, 1896-2004
The Economist Historical Archive
Eighteenth Century Collections Online
The Illustrated London News Historical Archive, 1842-2003
The Independent Digital Archive
Indigenous Peoples: North America
International Herald Tribune Historical Archive 1887-2013
Liberty Magazine Historical Archive, 1924-1950
The Listener Historical Archive, 1929-1991
The Making of Modern Law: Foreign Primary Sources
The Making of Modern Law: Foreign, Comparative, and International Law, 1600-1926
The Making of Modern Law: Legal Treatises, 1800-1926
The Making of Modern Law: Primary Sources
The Making of Modern Law: Trials, 1600-1926
The Making of the Modern World
Nineteenth Century Collections Online
Nineteenth Century U.S. Newspapers
Picture Post Historical Archive
Political Extremism & Radicalism
Punch Historical Archive, 1841-1992
Sabin Americana, 1500-1926
Smithsonian Collections Online
The Sunday Times Digital Archive
The Telegraph Historical Archive
The Times Digital Archive
The Times Literary Supplement Historical Archive
U.S. Declassified Documents Online
U.S. Supreme Court Records and Briefs, 1832-1978
Women's Studies Archive

Copyright-free Research Material

Your work this session will primarily draw on material from scanned archival material, open source data you find, and selected Gale Primary Sources, including digitized material, its OCR text

and associated metadata. Beyond this, you are welcome to source additional information related to the research topic you decide on and should you do so, and opt to include it in your final StoryMap, it is important that you choose material with a license that allow for re-use, and that you include appropriate citation.

For an introductory overview of Fair Use, Licensing and other copyright considerations, begin with the UW Libraries Copyright Guide.

Further guides applicable to a wide variety of situations include Stanford University's Copyright Centre.

Digital Humanities, Copyright and Open Resources

• NYU Workshop on DH and copyright (summary)

• NYU Workshop with links and documents

• Cornell's chart on what's in the public domain

• Visual Resources Association Intellectual Property Rights and Copyright Law resources

More Reading

• Laura Quilter Copyright Futures in the Digital Humanities

• DH Best Practices (NYU)

• Student collaborators' Bill of Rights

• Media Commons DH Collaborators' Bill of Rights

• Primary Sources on Copyright, 1540-1900

• NYU Guide to copyright

• Columbia: Copyright and DH

Specific Full-Text and Image Databases (UW Libraries)

• Search in research guides

• EEBO (Early English Books Online)

• ECCO (Eighteenth Century Collection)

• ArtStor

Public Domain Material

• HathiTrust (general)

– Datasets info

– Research Center info

• Internet Archive

• Project Gutenberg

• BYU Corpora

• American National Corpus

• Directory of Open Access Journals

• GitHub Datasets

• OER Commons

format/subject-based

• Open Culture (cultural/educational media)

• Data for Research from JSTOR (yes, open!)

• Chronicling America (newspapers)

• Open Greek and Latin Project

• CELT (Corpus of Electronic Texts)

• Open Access English/American Lit sources

• Wikimedia Commons

• British Library images

Open Access Metadata

• Digital Public Library of America

• Europeana

– Terms of use

And disciplinary librarians: they know subjects, tools, and resources

http://guides.lib.uw.edu/research/subject-librarians

(*thanks to Helene Williams/Information School for compiling the preceding list of copyright resources. Questions? helenew@uw.edu)*

WEEKS 2b-3b: Paleography, Transcribing & Text as Data

These sessions will be a combination of reading, and hands on work in class. Here's what you will cover:

- Working with primary source material and referring to Newbook intern manuals (see below), you will practice transcribing handwritten 19th century material into plaintext. In doing so, you'll get some insight into some of the challenges of working with 19th century handwriting.
- You'll consider some of the editorial considerations transcribers should be aware of. You'll note these in your transcriptions.
- After this hands on work, you will consider the process of 'distant reading', or exploring a body of texts for recurrent thematic patterns. You'll have a chance to put your reading into practice when we distant read some of the datasets (in this case, collections of txt files) to determine which words occur most frequently.

Paleography & Transcribing

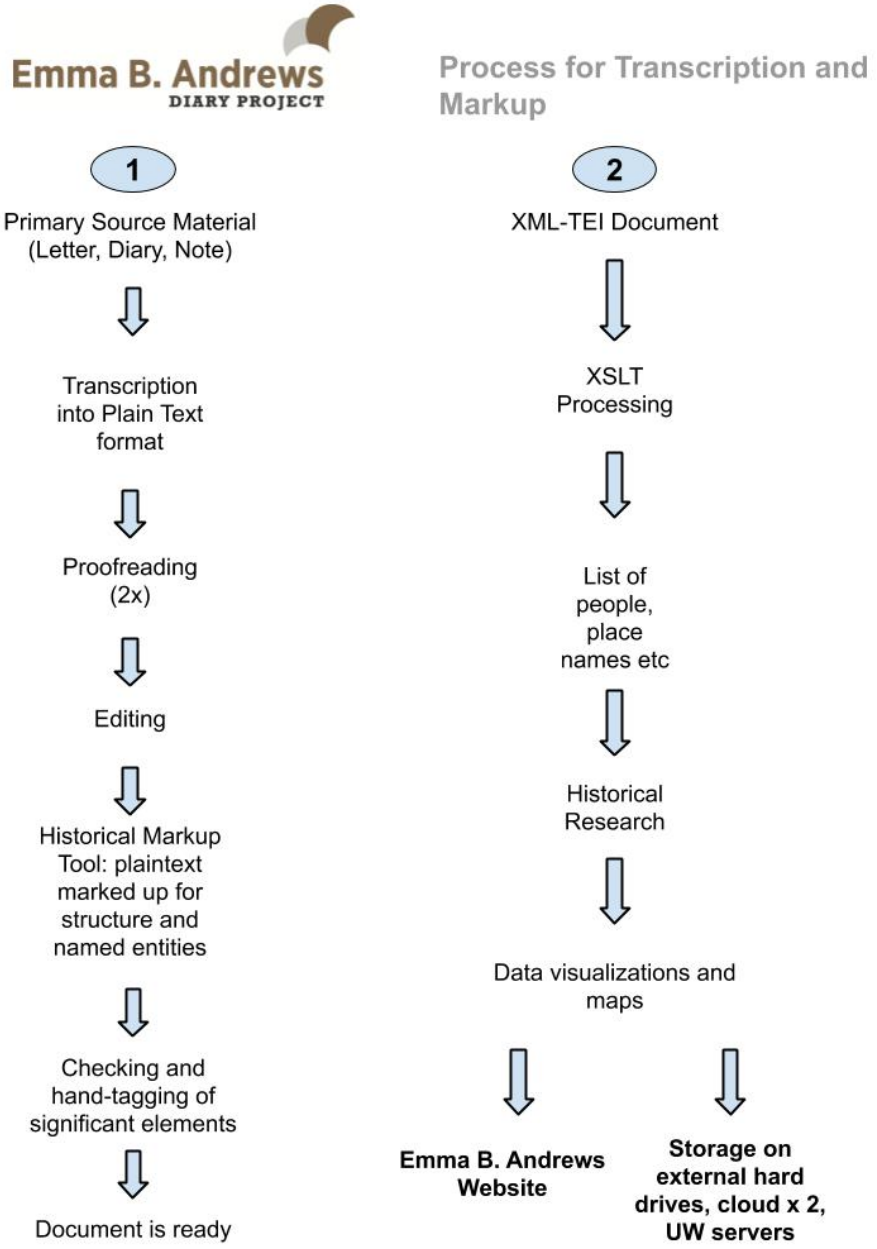
The notes and instructions for transcribing come directly from the Emma Andrews project guidebook for interns. The intent of including them is to give you some insight into the work process and considerations for a text-based DH project.

Emma B. Andrews Diary Project Transcription & Editing Process

Over the course of the project (2020 marks 8 years' work), we have developed processes for efficiently transcribing and encoding texts, and using the encoded data to extract information about 'named entities', notably people, places, hotels, boats, artwork and historic sites.

This is often slow, painstaking work, but we have developed a number of time-saving strategies over the years.

- initially we worked in Google Docs but found that it was very difficult to track edits and changes in a document when multiple editors were working on the same material. We switched to working in Github (https://github.com/) where a master copy of each text file is kept. An intern 'checks out' a copy of the master, completes their work, then submits a request to me to merge the two versions. I check their work, and once all looks good, I complete the merge. It is possible to track back through the changelog in Github, and roll back versions if necessary.



- Markup - we marked up text by hand in the early years, focusing on capturing document structure - page breaks, journal entries, paragraphs etc. We also marked up dates, people's names, place names, hotels, boats and hotels. We work with XML-TEI, using a fairly minimal schema we developed for our Project. Here is a sample of a marked up text:

```
<div xml:id="EBA19111021" type="Entry">
<p><title><name type="vessel" ref="#SS_Berlin">SS. Berlin</name>. <date
when="1911-10-21">Oct. 21st</date></title>.</p>
<p>Sailed this morning for <placeName ref="#Genoa">Genoa</placeName> - where we will
wait 15 days for our boat for <placeName ref="#Alexandria"
>Alexandria</placeName>. Have my old appartment which is the Captain's, and
very commodious and comfortable, and which I occupied two years ago. We know a
few of the passengers. The 4 days we passed in <placeName ref="#New_York">New
York</placeName> were dismal in the extreme - it rained all the time, I took
a severe cold, and did not go out of the house after the first day - and had to
cancel a long standing engagement to drive with <orgName
ref="#Fairfield_Osborns">the Fairfield Osborns</orgName>, who had invited a
lot of pleasant people to meet us. It was altogether a great piece of
disappointment, that visit in <placeName ref="#New_York">New
York</placeName>.</p>
</div>
<div xml:id="EBA19111002" type="Entry">
<p><title><placeName ref="#Genoa">Genoa</placeName> - <date when="1911-10-23">
<sic>Oct. 2</sic>.
<!-- need to calculate which Wednesday this was likely to be (options: October 25th, November 1, 8, 15).
Research when the SS Berlin arrived in Genoa after departing NY on October 21st 1911 -->
Wednesday</date></title>.</p>
<p>Arrived on a cold, damp morning - came to the <name type="hotel"
ref="#Hotel_Miramare">Hotel <hi rend="underline">Miramare</hi></name>
<pb n="87"/> which had been recommended to us. Until we passed <placeName
ref="#Azores_Islands">the Azores</placeName>, we had a fairly good passage -
though warm and misty - after that to <placeName ref="#Gibraltar"
>Gibraltar</placeName> bright weather, but tremendous rolling - slow but
very disconcerting. After leaving <placeName ref="#Gibraltar"
>Gibraltar</placeName> we began a hustle with <placeName
ref="#Mediterranean_Sea">the Mediterranean</placeName>, and we were cruelly
battered about - could not make the landing at <placeName ref="#Algiers"
>Algiers</placeName> and the little mail boat that came out to us, on
returning lost her Captain and one sailor. We find this hotel the most
delightful one we have ever known - large airy appartments, with every luxury -
perfect service - and food. My corner room is the most charming room I have ever
seen - and the large sitting room next it is the same - and the string of
bedrooms next the salon, made an imposing array. <persName
```

- We have begun marking up data using our Historical Markup Tool, which helps prevent errors in encoding and takes a little less time to do. It's available via the project website. Guidance for use can be found at the end of this course book, in the Appendix.

- Historical metadata - we use the list of people's names to create a database (the Emmapedia) where we capture metadata about the individual, like their date of birth/death, a brief biography, notable publications as well as any open source images we can find. This is uploaded into the Omeka content management system that generates the Emma B. Andrews website.

Tips and Tricks for Transcribers (contributed by Allison, EBA transcription intern 2013-2015)

- Names and places can be especially difficult, especially for those not as familiar with Egyptology or with the lives of Emma and Theodore. Obviously we're all learning a ton on this project, but there are some resources that can make your life far easier.

- Read the diaries and letters themselves - even just a volume or two will help you make more sense of what you're transcribing.

- Read about the period: if you've got the time, take a look at John Adams' *The Millionaire and the Mummies*. It is invaluable in learning what names are likely to show up, for Emma and Theodore's acquaintances, the places they all frequented, including Egypt. Other popular books on Ancient Egyptian history can be helpful too, and interesting. Ask Sarah for recommendations; she has copies of most material and is happy to share.

- *Who's Who in Egyptology*? Sarah scanned this and put it up on the shared drive. It's not searchable per se but the entries are all in alphabetical order by last name. Usually at least the first letter or two are legible, so use that to skim through the names until you see something that looks similar, then read their entry to see if it makes sense in context.

- Our team's historical research: we have interns working on researching people found in the diaries and it's searchable on the project website, and a spreadsheet can be found on the shared drive. *Note*: some people may be marked as 'private' on the website, so they're not visible to the public. Ask Sarah for collaborator access to access the backend and all the people.

- Utilize your teammates: flip through the other transcriptions to see what the other interns have done. Have a letter but you can't read the signature? Scroll through the sources until you see another in the same handwriting and look at the transcription of that one to see if someone else was able to figure it out.

- Language: for the most part, our sources are all in English aside from random Egyptian hieroglyphics. Remember that when trying to decipher texts. Don't just look at each letter, look at the words as a whole. If it makes no sense or looks like gibberish, use some of the other tricks here to decipher it.

- Context: consider which words are frequently used in conjunction with each other. "He it" doesn't make sense grammatically, but "He is" does. Sometimes figuring out just a word or two this way can make a whole sentence suddenly make sense.

- On challenging handwriting: look out for handwriting quirks by noting words you or others have figured out already. Emma (and Theodore to a degree) floats her 'T' crosses off to the side. Davis' 'p' looks an awful lot like 'fr', but once you realize it's a 'p', words like 'photograph' go from a jumble of scribbles to real words. This is true for spacing as well. These weren't written with ballpoints, so writing habits were different, such as more running words together and sharper turns.

- Change your perspective: look at texts both zoomed in and zoomed out. Sometimes seeing a complete line or sentence helps, since our minds try to fill in blanks logically. Other times it can be helpful to zoom in to distinguish between squiggles.
- Frustrated? Don't bang your head against a wall. Give it up for a few minutes, hours, or even days, then go back with fresh eyes. This always makes a difference.
- Proofread your work! You don't have to redo the entire transcription, but take a few moments to go back over what you've typed to ensure there aren't any glaring typos or autocorrect errors.

Textual Features to be aware of for markup

line breaks

page breaks

page numbers (match file page # or original page #)

tabs

spaces

alignment

smart quotes

smart dashes

underline

handwriting

strikethrough

hieroglyphs

questionable text (things we're unsure about)

illegible text

missing text (not filled in)

missing text (filled in)

margin notes

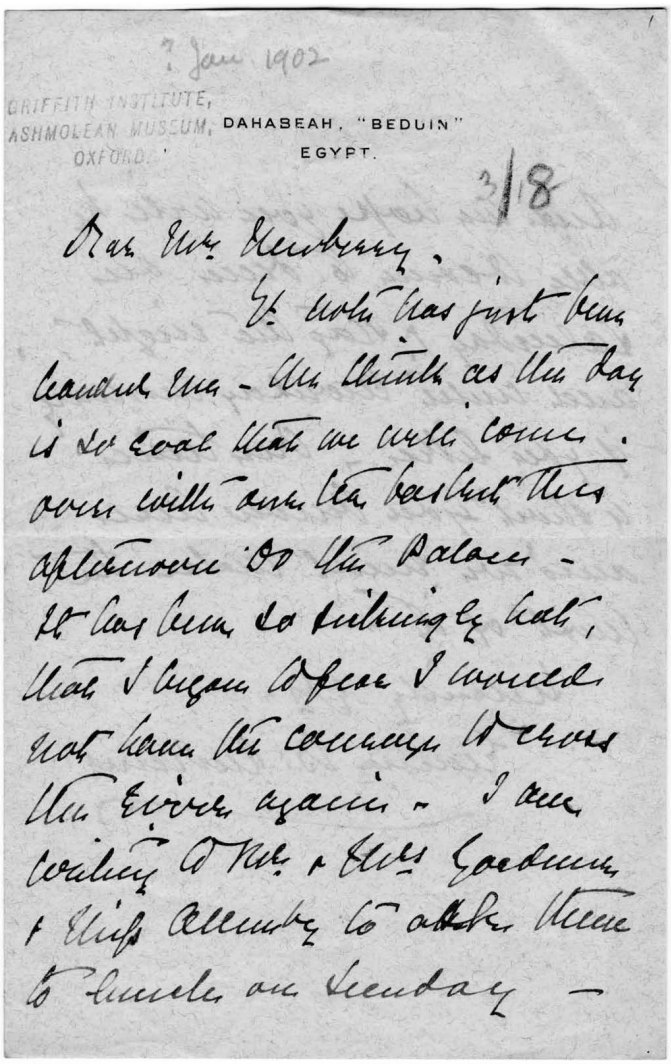
letterhead

sic (obvious typos by original transcriber or source misspelling)

notes added by a writer other than the author

Activity #1 (see Discussion Post for more info)

Take a look at the brief letter below, written in January 1902 by Emma Andrews to 'Dear Mr. Newberry'. Review the hints and tips above and see if you can make any sense of it. The original images are here should you wish to download them and zoom in/out to take a closer look at the text.



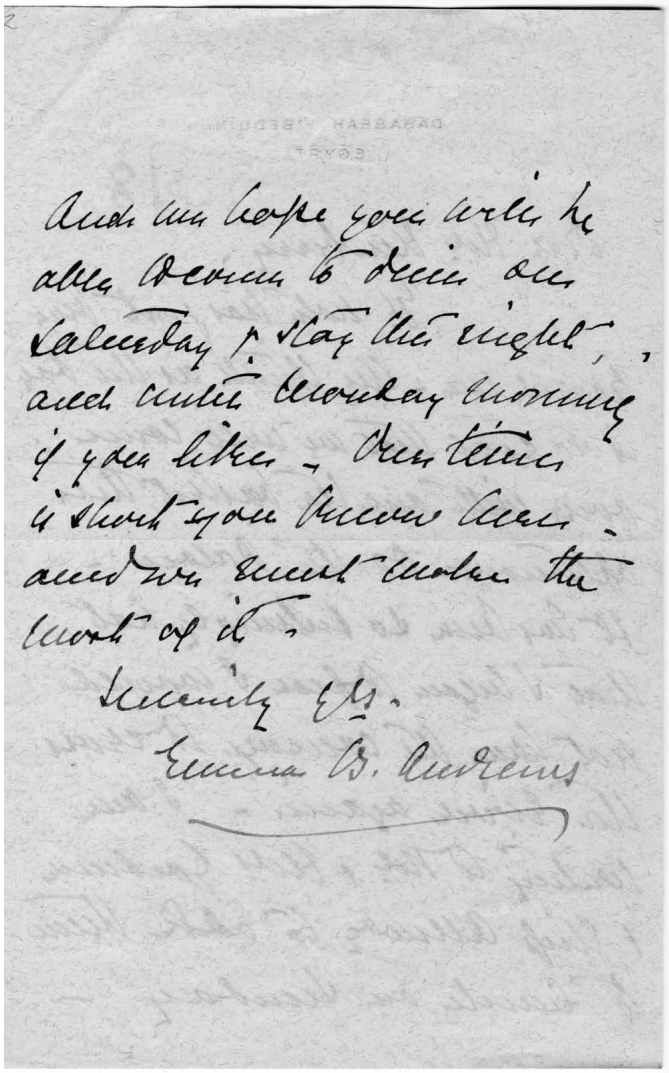
7 Jan 1902

GRIFFITH INSTITUTE, ASHMOLEAN MUSEUM, OXFORD. DAHABEAH, "BEDUIN" EGYPT.

3/18

Dear Mr Newberry,

Your note has just been handed me - the thinks as the day is so good that we will come over with our lunch this afternoon to the Palace - It has been so strikingly hot, that I began to fear I should not have the courage to cross the River again - I am writing to Mr + Mrs Goodwin + Miss Allen to ask them to lunch on Sunday -



And we hope you will be able to come to dine on Saturday & stay the night, and until Monday morning if you like - Then time is short you know even - and we must make the most of it -

Sincerely Yrs.

Emma B. Andrews

Can’t make any headway with Emma’s handwriting? You’re not the first. Here are some other samples to try if you are stuck:

Joseph Lindon Smith

Helen Winlock

Theodore Davis

1. Enter your transcription for one page of a chosen letter into the Discussion Post. Make sure you label which letter you’re transcribing with the correspondent’s name (eg ‘letter written by Mrs. Emma Andrews).
2. What challenges did you encounter? What stylistic peculiarities did you identify in the handwritten text?
3. What specific textual features do you identify in your letter, based on the list above?

Getting to grips with a dataset when you’re not familiar with the topic

Learning Objective: The purpose of this exercise is to familiarize yourself with the traditional workflow, terms and output for a **text mining project** using one of the most popular DH tools for this purpose. You’ll be working with some of the text data from week 2a. The purpose is to help you identify some themes in the data which you may wish to explore further over the course of the quarter. Much of the primary source material will be unfamiliar to you; we are leveraging a digital tool to extract a list of the most prevalent terms in your corpus of data. This will provide a starting point for your project.

Background: What is Text Analysis? Distant Reading and Using Digital Tools for Thematic Analysis

The term data mining refers to any process of analysis performed on a dataset to extract information from it. That definition is so general that it could mean something as simple as doing a string search (typing into a search box) in a library catalogue or in Google. Mining quantitative

data or statistical information is standard practice in the social sciences where software packages for doing this work have a long history and vary in sophistication and complexity.

But data mining in the digital humanities usually involves performing some kind of extraction of information from a body of texts and/or their metadata in order to ask research questions that may or may not be quantitative. Supposing you want to compare the frequency of the word “she” and “he” in newspaper accounts of political speeches in the early 20th century before and after the 19th Amendment guaranteed women the right to vote in August 1920. Suppose you wanted to collocate these words with the phrases in which they were written and sort the results based on various factors—frequency, affective value, attribution and so on. This kind of text analysis is a subset of data mining. Quite a few tools have been developed to do analyses of unstructured texts, that is, texts in conventional formats. Text analysis programs use word counts, keyword density, frequency, and other methods to extract meaningful information. The question of what constitutes meaningful information is always up for discussion, and completely silly or meaningless results can be generated as readily from text analysis tools as they can from any other.

Johanna Drucker, *Intro to Digital Humanities*, 2013

This class is primarily focused on text or data mining as a humanities research methodology. For the purposes of data analysis, we are using the plain text transcriptions of primary source material created by the Emma B. Andrews Diary Project, as well as the underlying OCR text of scanned primary source documents found in Gale Primary Sources and the Digital Scholar Lab.

Reading

- Matthew L. Jockers and Ted Underwood. “Text-Mining the Humanities.” *A New Companion to Digital Humanities*, Wiley-Blackwell, 2015, pp. 291–306. *Wiley Online Library*, doi: 10.1002/9781118680605.ch20 .

Abstract: “This chapter provides a broad overview of how text mining can be usefully employed in humanistic research. The chapter begins by addressing the question of why scholars in the humanities should care about text mining and what they might expect to gain by embracing what are deeply computational and deeply quantitative methods. We then offer a quick synopsis of the key watersheds in the history of text mining. The bulk of the chapter discusses central methodologies used in humanistic text mining. Using examples from the humanities, we unpack the differences between supervised and unsupervised learning and discuss how tools developed by researchers in other fields can be usefully employed to address humanistic questions. Drawing from personal experience, we address some of the significant challenges associated

with data quality, metadata, and copyright restrictions before moving to a discussion of a few exemplary projects and resources for further study."

- Ted Underwood, Seven ways humanists are using computers to understand text, *The Stone and the Shell,* June 4 2015 https://tedunderwood.com/2015/06/04/seven-ways-humanists-are-using-computers-to-understand-text/

Ted Underwood is one of the preeminent scholars working with text mining in the humanities. His blog, *The Stone and the Shell*, has a wealth of materials about text analysis, visualizations, and digital humanities, in general. In this blog post, Underwood describes why humanities scholars might use text analysis in their research, but he also states the pitfalls of using statistical analysis to perform corpus-based scholarship. This post does a thorough job of giving examples of things you might do with a text in order to give "a loose sense of how different activities are related to different disciplinary traditions" within the realm of text mining.

- Geoffrey Rockwell, What is Text Analysis, Really?, *Literary and Linguistic Computing*, Volume 18, Issue 2, June 2003, 209–21, https://doi.org/10.1093/llc/18.2.209

- Ted Underwood, A Genealogy of Distant Reading , *Digital Humanities Quarterly,* Volume 11, Number 2, 2011, http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html

- Marti Hearst, *What is Text Mining?*

Example Text Mining Projects using Newspaper Content

- Ryan Cordell and David Smith. *Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines* (2017). http://viraltexts.org.

"This site presents data, visualizations, interactive exhibits, and both computational and literary publications drawn from the Viral Texts project, which seeks to develop theoretical models that will help scholars better understand what qualities—both textual and thematic—helped particular news stories, short fiction, and poetry "go viral" in nineteenth-century newspapers and magazines. During this period, texts published in newspapers and magazines were not typically protected as intellectual property, and so literary texts as well as other non-fiction prose texts circulated promiscuously among newspapers as editors freely reprinted materials borrowed from other venues. In the *Viral Texts* project, we're asking: What texts were reprinted and why? How did ideas—literary, political, scientific, economic, religious—circulate in the public sphere and achieve critical force among audiences? By employing and developing

computational linguistics tools to analyze the large textual databases of nineteenth-century newspapers newly available to scholars, this project will generate new knowledge of the nineteenth-century print public sphere."

- Lincoln Mullen. *America's Public Bible: Biblical Quotations in U.S. Newspapers, website, code, and datasets* (2016). http://americaspublicbible.org .

"*America's Public Bible: Biblical Quotations in U.S. Newspapers* tracks Biblical quotations in American newspapers to see how the Bible was used for cultural, religious, social, or political purposes. Users can either enter their own Biblical references or choose from a selection of significant references on a range of topics. The project draws on both recent digital humanities work tracking the reuse of texts and a deep scholarly interest in the Bible as a cultural text in American life. The site shows how the Bible was a contested yet common text, including both printed sermons and Sunday school lessons and use of the Bible on every side of issues such as slavery, women's suffrage, and wealth and capitalism."

- Kristi Palmer, Ted Polley, and Caitlin Pollock . *Chronicling Hoosier* (2016). http://centerfordigschol.github.io/chroniclinghoosier/index.html

"This project tracks the origins of the word "Hoosier." The site's maps visually demonstrate the geographic distribution of the term "Hoosier" in the Chronicling America data set. This distribution is measured by the number of times the term appears on a newspaper page. Each point on the map shows a place of publication where a newspaper or newspapers contain the term. Another feature on the web site is the Word Clouds by Decade visualizations, which are created by looking at the word "Hoosier" in context. The text immediately surrounding each appearance of the word is extracted and from this the most frequently occurring terms are plotted."

Key Concepts

Some general definitions and important points related to text analysis, to reinforce what you've been reading about:

- Text analysis: A form of data mining, using computer-aided methods to study textual data.
- Distant reading: As compared to close reading, which finds meaning in word-by-word careful reading and analysis of a single work (or a group of works), distant reading takes large amounts of literature and understands them quantitatively via features of the text. (Conceptualized by Franco Moretti)

- Non-consumptive research: Research in which computational analysis is performed on text, but not research in which a researcher reads or displays substantial portions of the text to understand the expressive content presented within it.
- Algorithm: A process a computer follows to solve a problem, creating an output from a provided input.
- Text corpus/corpora: A “corpus” of text can refer to both a digital collection and an individual's research text dataset. Text corpora, the plural form, are bodies of textual data.
- Content Set: In the Digital Scholar Lab environment, a Content Set is a sub-collection of Gale Primary Sources content created by users.
- Volume: Generally, a digitized book, periodical, or government document.
- Optical character recognition (OCR): Mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text. The quality of the results of OCR can vary greatly, and raw, uncorrected OCR is often described as “dirty”, while corrected OCR is referred to as “clean”.

Key points

Introduction to text analysis research in the humanities and social sciences: key approaches and examples	• Text analysis: the process by which computers are used to reveal information in and about text.
	• Text analysis usually involves breaking text into smaller pieces; reducing (abstracting) text into things that a computer can crunch; counting words, phrases, parts of speech, etc.; using computational statistics to develop hypotheses.
	• Text analysis impacts research by shifting the researcher’s perspective of the text, and makes it possible to ask questions that cannot be answered by human reading alone, larger corpora for analysis, and longer periods of study.

	• Text analysis research questions often involve change over time, pattern recognition, and comparative analysis.

Finding and gathering text

- Text can be approached as data and analyzed by corpus/corpora.
- Before analyzing textual data, it is important to ensure the text is of sufficient quality (e.g., OCR-ed data is cleaned up) and fully prepared (certain unnecessary elements are discarded).

Methods for accessing and downloading textual data

- Finding text suitable for computational analysis is challenging, especially with issues of copyright and licensing restrictions, format limitations, and hard-to-navigate systems.
- Three commonly used sources to find textual data are vendor databases, digital collections, and social media. Each source has its own strengths and challenges when it comes to downloading text. For this class you will be using material from the first resource, i.e. Gale Primary Sources along with the open source material generated by the team at the Emma B. Andrews Diary Project.

Factors that affect choice of textual data:

- How much flexibility is needed for working with the data?
- What is the the technical skillset of the researcher?
- Are there funding limitations?

In Class Activity (Week 3b)

Data is available here

I. About Voyant

Voyant is a web-based reading and analysis environment for digital texts which is freely available for users:

https://voyant-tools.org

The tool is open source, and is widely used by Digital Humanists using text analysis and visualization for research. One drawback of the tool is that it is hosted on the McGill University servers, and so its ability to process very large datasets is limited. It is possible, however, to install it on a home or local server.

Key features include:

1. importing documents in various formats (plain text, HTML, XML, PDF, RTF, MS Word, ODF, etc.)
2. several tools for studying term frequencies and distributions within documents and within a collection of documents (a corpus)
3. a full-text reader that supports very large texts and includes interactive features
4. interaction between the the tools that facilitates navigation and exploration at different scales (from "close reading" to "distant reading")
5. a mechanism for bookmarking and sharing instances of Voyant Tools (specific texts and tools) through persistent URLs

Screencast and Doc Tutorials

- Screencast giving general tools overview: https://youtu.be/00V3Xbr1XA4
- Screencast describing of individual tools in Voyant can be found here: https://www.youtube.com/playlist?list=PLDCADF35691404F54
- Written overview of the tools that Voyant supports is here: https://voyant-tools.org/docs/#!/guide/tools

Hands-On Assignment

I. Gather a selection of documents to use as the basis for a thematic text analysis from the dataset above. Upload these documents to Voyant.

To upload a **group** of documents, you must first create a zip file of your corpus. Upload the zip file to Voyant: From the landing page, select the zip folder you have just created. Click ‘upload’. Voyant will do the work of expanding the archive and processing all of the documents in your dataset.

II. Understanding the Dashboard View

Familiarize yourself with the dashboard.

```
List three pieces of information about your content set that you can see at a glance from the dashboard view.
```

```
What are your overall impressions of the Voyant dashboard? Do you find it intuitive and user friendly? If not, what do you find unclear or challenging?
```

III. Voyant Suite of Tools

Voyant provides a range of tools and options for text analysis. What information can you learn from the following tools and visualizations? Record your answers as brief paragraphs.

1. Cirrus
2. Document Terms
3. Mandala
4. Contexts

IV. Explore your Content Set using Voyant

An opportunity to explore your content sets using the tools embedded in Voyant. The goal is for you to experiment with your data, to customize tool options and to create a visualization or two.

A: Most Frequent Words comparison

Open two Voyant windows:

Load your corpus of texts in one window, and load a single text in the other Voyant window. Compare the word clouds.

```
Are the most frequently used words in the single document the same ones that appear most frequently in the larger corpus? Describe any differences you observe.
```

B: Using Stopwords

In the window containing your full content set, apply the English stopwords.

In a second window, load up your full content set texts, but don't apply the stopwords.

```
Look at the two word clouds. How are they different?
```

Hover your mouse over the top right of the tool panel and use the button to generate a URL. A new window will open.

```
Copy the URL and then paste it into your assignment.
```

This enables other researchers to look at your research results.

C: Explore the Topics in your Content Set

Open the topic modeling panel by selecting the tool from the dropdown list:

```
What are the most common topics the tool identifies?
```

Week 3 Discussion

Using Voyant as a starting point, identify a theme or topic to develop for the rest of this quarter. It doesn't need to be tied to Near Eastern Archaeology; you can explore something that is of interest to you which is connected to the primary source material we have been looking at for the past couple of weeks.

Answer these questions for this week's discussion:

1. List three pieces of information about the texts that you've chosen that you can see at a glance from the dashboard view.

2. What are your overall impressions of the Voyant dashboard? Do you find it intuitive and user friendly? If not, what do you find unclear or challenging?

3. What information can you learn from the following tools and visualizations? Record your answers as brief paragraph:

    - Cirrus
    - Document Terms
    - Context

4. Most Frequent Words comparison - open two Voyant windows:
   Load your corpus of texts in one window, and load a single text (or a different corpus of texts) in the other Voyant window. Compare the word clouds.
   Are the most frequently used words in the single document or corpus A the same ones that appear most frequently in corpus B? Describe any differences you observe.

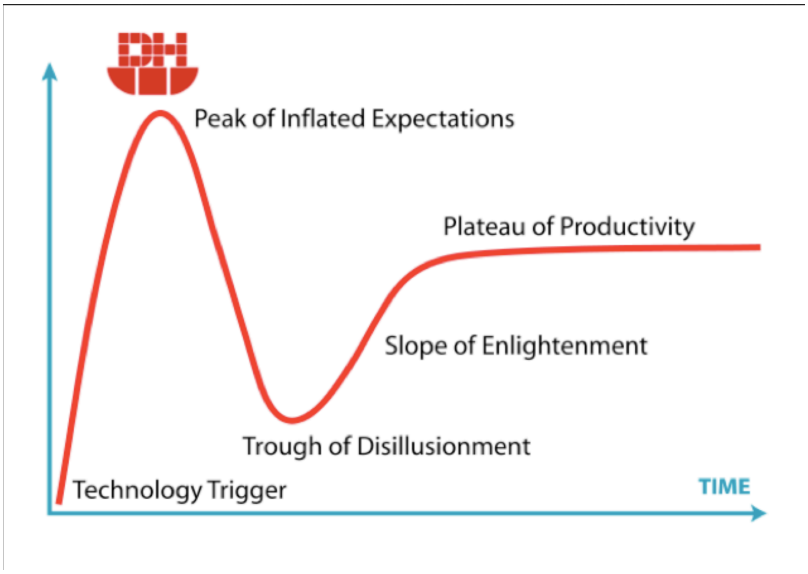
5. Using Stopwords In the window containing your full content set, apply the English stopwords.
   In a second window, load up your full content set texts, but don't apply the stopwords.
   Look at the two word clouds. How are they different?
   Hover your mouse over the top right of the tool panel and use the button to generate a URL. A new window will open.
   Copy the URL and then paste it into your Discussion Post.
   This enables other researchers (in this case, me) to look at your research results.

6. Explore the Topics in your Content Set

Open the topic modeling panel by selecting the tool from the dropdown list: What are the most common topics the tool identifies?

Replies: Due Tuesday of Week 4 11.59pm

WEEK 4a: Planning & Managing Digital Projects



Before launching any project, it's important to consider how you are going to PLAN and MANAGE your work. You will need to identify what your goals are and what the scope of the work is. We'll discuss planning an appropriate time frame, accountability issues, teamwork, funding and documentation. We'll look at some broad project management issues and work on developing a project charter and project one pager for your quarter's work.

Preliminary Reading

1. Paige Morgan, 'How to get a Digital Project off the Ground' *HASTAC* blog, June 6 2014.
2. Collaborators' Bill of Rights
3. Lynda Gratton & Tamara J. Erickson, 'Eight Ways to Build Collaborative Teams', *Harvard Business Review*, November 2007.
4. Stan Ruecker & Milena Radzikowska, 'The Iterative Design of a Project Charter for Interdisciplinary Research', *DIS '08: Proceedings of the 7th ACM conference on Designing interactive systems*, February 2008, https://doi.org/10.1145/1394445.1394476
5. Jennifer Guiliano & Simon Appleford, 'Building your First Work Plan', www.devdh.org

6. Large Project Management (DHSI Coursepak compiled by Lynne Siemens - a lot of articles here, which you can you as a reference moving forward).

Project Management from start to finish

Sarah Storti and Brooke Lestock for UVa Scholars' Lab Praxis Program, 2011-12

1. Decide whether the project should happen.

- (Does it make a worthwhile scholarly or technical intervention? Is it sustainable? What resources do we have/need?)

2. Devise a clear project workplan.

- A good workplan is a statement of deliverables with a path to completion: a "necessary fiction" that accounts for both the intellectual vision and reality.
- Putting together a workplan involves making frank assessments of the skills/resources available and getting honest estimates from team members.
- The workplan must be flexible, but only changed for good reasons.
- The workplan must lay out clear, core deliverables, and be broken down into small, manageable chunks.

3. Delegate tasks and maintain momentum.

- Let team members help define how they will be involved.
- Let the person taking on an assignment help to set the due date.
- As PM, take responsibility for making final decisions when necessary.

4. Manage communication.

- Schedule meetings and keep them on track. Meetings should include regular check-ins with the whole team as well as smaller, task-focused group meetings.
- Each meeting should address:
    - what has been done,

    - what needs to be done,
    - and what may be holding people back.
- The PM is responsible for devising a clear, easy way to track progress and work.
- Team members should report regularly: in public, to self, and to partners.

The PM's main goal is to keep all team members on task, and to deliver what was promised.

Basic Tools for Project Management

- Basecamp (free educational account)
- Trello
- Airtable
- Gantt charts (Google Sheets Add On)
- Slack
- Google Drive/Microsoft OneDrive/Dropbox
- Github
- Microsoft Project

Data Management

The organizational structure of your data can help **secondary users** find, identify, select, and obtain the data they require. As you work through the material for this section, keep these questions in mind:

How should you source your data?

Why should you organize your data?

How should you organize your data?

In which format should you organize your data?

How should you name your files?

When developing a digital humanities project, planning and building sustainably will mitigate the risk that your work becomes inaccessible, trapped in an obsolete platform within a few short years. While your work on your class project may not extend beyond the timeframe of this course, you will need to develop an awareness of best practices for project planning and managing the data that you work with and present online or in other digital formats. The resources below will give you an introductory framework for:

- developing a Data Management Plan which is a document outlining how a researcher plans to manage data during and after a research project including how it will be organized, maintained, and shared.
- assessing the risks that are associated with building a digital project.

Start with UW Libraries' Data Management Guide

Additional resources:

Write a Data Management Plan, MIT Libraries

DMPTool (log in with your UWID)

http://www.data-archive.ac.uk/create-manage

Risk Assessment

In working with digital data, here is a list of risks to consider and plan for in building a DH project:

- **Media Failure**. All storage media must be expected to degrade with time, causing irrecoverable bit errors, and to be subject to sudden catastrophic irrecoverable loss of bulk data such as disk crashes or loss of off-line media.
- **Hardware Failure**. All hardware components must be expected to suffer transient recoverable failures, such as power loss, and catastrophic irrecoverable failures, such as burnt-out power supplies.
- **Software Failure**. All software components must be expected to suffer from bugs that pose a risk to the stored data.
- **Communication Errors**. Systems cannot assume that the network transfers they use to ingest or disseminate content will either succeed or fail within a specified time period, or will actually deliver the content unaltered. A recent study "suggests that between one (data) packet in every 16 million packets and one packet in 10 billion packets will have an undetected checksum error".
- **Failure of Network Services**. Systems must anticipate that the external network services they use, including resolvers such as those for domain names and persistent URLs, will suffer both transient and irrecoverable failures both of the network services and of individual entries in them. As examples, domain names will vanish or be reassigned if the registrant fails to pay the registrar, and a persistent URL will fail to resolve if the resolver service fails to preserve its data with as much care as the digital preservation service.
- **Media & Hardware Obsolescence**. All media and hardware components will eventually fail. Before that, they may become obsolete in the sense of no longer being capable of communicating with other system components or being replaced when they do fail. This problem is particularly acute for removable media, which have a long history of remaining theoretically readable if only a suitable reader could be found.
- **Software Obsolescence**. Similarly, software components will become obsolete. This will often be manifested as format obsolescence when, although the bits in which some data was encoded remain accessible, the information can no longer be decoded from the storage format into a legible form.
- **Operator Error**. Operator actions must be expected to include both recoverable and irrecoverable errors. This applies not merely to the digital preservation application itself, but also to the operating system on which it is running, the other applications sharing the

same environment, the hardware underlying them, and the network through which they communicate.

- **Natural Disaster**. Natural disasters, such as flood, fire and earthquake must be anticipated. Other types of threats, such as media, hardware and infrastructure failures, will typically manifest then.
- **External Attack**. Paper libraries and archives are subject to malicious attack; there is no reason to expect their digital equivalents to be exempt. Worse, all systems connected to public networks are vulnerable to viruses and worms. Digital preservation systems must either defend against the inevitable attacks, or be completely isolated from external networks.
- **Internal Attack**. Much abuse of computer systems involves insiders, those who have or used to have authorized access to the system. Even if a digital preservation system is completely isolated from external networks, it must anticipate insider abuse.
- **Economic Failure**. Information in digital form is much more vulnerable to interruptions in the money supply than information on paper. There are ongoing costs for power, cooling, bandwidth, system administration, domain registration, and so on. Budgets for digital preservation must be expected to vary up and down, possibly even to zero, over time.
- **Organizational Failure**. The system view of digital preservation must include not merely the technology but the organization in which it is embedded. These organizations may die out, perhaps through bankruptcy, or their missions may change. This may deprive the digital preservation technology of the support it needs to survive. System planning must envisage the possibility of the asset represented by the preserved content being transferred to a successor organization, or otherwise being properly disposed of. For each of these types of failure, it is necessary to trade off the cost of defense against the level of system degradation under the threat that is regarded as acceptable for that cost.

adapted from Rosenthal, David S. H., et al. "Requirements for Digital Preservation Systems." *D-Lib Magazine*, vol. 11, no. 11, 2005, doi:10.1045/november2005-rosenthal.

Additional reading:

Bailey, Jefferson. *I Review 6 Digital Preservation Models So You Don't Have To*, 2014, www.jeffersonbailey.com/i-review-6-digital-preservation-models-so-you-dont-have-to/

Digital Preservation Handbook, *Digital Preservation Coalition*

Tips for choosing a File Format

Pick a file format that's less likely to become obsolete.

- Open standards (so not owned by any particular corporation), not patent-encumbered
- With at least one open-source reader/writer
- In broad use
- As high-quality (whatever that means given the medium) as practical

Key point: you may need two copies!

- One for access/use, one preservation master
- Make the preservation master first! The other is derivable.

Backing up your data

The 3 - 2 - 1 backup strategy is one which is widely recognized as being effective:

> 3 – Keep 3 copies of any important file: 1 primary and 2 backups.
>
> 2 – Keep the files on 2 different media types to protect against different types of hazards.
>
> 1 – Store 1 copy offsite (e.g., outside your home or business facility).

Example: I have a favorite photo of my dog with a file name of woody.jpg that I want to preserve. I have a copy on my laptop (primary), and I keep a copy in my Dropbox account (Cloud storage, offsite, backup #1). I also backup to an external hard drive every Sunday (backup #2, second type of media).

Paul Ruggiero and Matthew A. Heckathorn, 'Data Backup Options', *Carnegie-Mellon University for the United States Computer Emergency Readiness Team (US-CERT),* 2012

Developing a Project Charter, Project One Pager & Data Management Plan

This session, you'll have the opportunity to discuss and develop a preliminary Project Charter, and Project One Pager which you'll be using to guide your research and StoryMap development this quarter. You'll also spend a little time formulating a draft Data Management Plan for the materials you gather.

Project Charter

Once you have designated a Project Manager, they will be responsible for finalizing and overseeing the Charter, which will be included in the final project presentation.

Overview and Purpose

Even if you think everyone in your group is on the same page, it's still a really good idea to have a discussion about expectations, ways of working, and even pet peeves. Think of a charter as an excuse to have a healthy discussion.

1. Choose three words to describe the spirit in which your group will work together.
2. How will you communicate with each other (e.g., text messaging, email, Google group, Trello, etc.)?
3. Where will you store your files (e.g., Dropbox, Google Drive, server, Github etc.)?
4. When you work on a document collaboratively, how will you ensure that you don't overwrite each other's changes?
5. How often will you meet outside of class? Where will you meet? Do you need a regular meeting time? If you'll schedule meetings as necessary, what days and times are generally good for people?
6. When are people planning to be out of town or especially busy? How can you work around this?

7. Assign the following roles to project member(s). Please note that no single team member is responsible for any of these roles; rather, the specialist coordinates activity related to this work and assigns tasks to team members. If your team has more than six members, multiple people may be assigned to one role. If your group has fewer than six members, please combine two roles.

    1. **Project Manager**: Pays close attention to schedule and milestones. Alerts the team to possible roadblocks or time-crunches. Ensures that communication among team members is efficient and harmonious. Keeps track of all project documentation. Takes notes at meetings. Communicates team needs (for example, additional training on a tool) to the professor, TA and/or IT Assistant. Communicates with subject-matter expert. Submits milestones on time via Canvas -

    2. **Web Specialist**: Oversees the design and structure of the site on Omeka. Works with the CMS (or HTML files) to ensure that the site performs to the team's specifications. Installs any required updates to the CMS. Archives the project and submits the files to the professor by the end of the quarter.

    3. **Data Specialist**: Oversees the cleaning, refining, and augmenting of the group's dataset. Teaches other team members how to use OpenRefine. Ensures the data is standardized, usable, and well-formatted.

    4. **Mapping Specialis**t: Oversees the project's maps. Geolocates data. Learns how to use (and teaches teammates how to use) the appropriate tools. Fine-tunes map display. Adds maps to site.

    5. **Data Visualization Specialist**: Oversees the project's data visualizations. Ensures that data is in the right format. Learns how to use (and teaches teammates how to use) the appropriate tools. Fine-tunes data visualizations and adds them to site.

    6. **Content Specialist**: Oversees the authoring of the site's main narrative and ensures that the data visualizations and maps integrate neatly with the written content. Writes section headers and captions. Obtains necessary images and embeds them in the site. Oversees the creation of the "About" page.

8. Do all decisions need to be unanimous, or is "majority-rules" OK?

9. How will you prevent meetings from going off-track?

10. What are group members' pet peeves from previous collaborations? How will you avoid these?
11. What will happen to the project when you're done with it? Will you maintain it, or let it expire?

Here is a more formal Sample Project Charter Template for ContentDM/Omeka. *Source*: *Utah State University*

Project One-Pager

This document will be a 'work in progress' for the next couple of weeks as you get to grips with the data and the scope of the work. A copy of this document will be included in the Final Project. Look at the example to give you a sense of what information you'll be including.

1. Project Name
2. Objective Statement
3. Requirements
4. Out of Scope
5. Team
6. Schedule

It's ok to change/edit the document as the scope of the work becomes clearer over the next couple of weeks.

Developing a Data Management Plan

How will you back up your work?

Choose at least three of the risks listed in the Risk Assessment section above, which you consider relevant for the material you are working with. Discuss strategies for dealing with these risks in the planning stages of your digital project work.

How will you ensure the (hypothetical) long-term sustainability and accessibility of your project and the data you are collecting?

WEEK 4b: Digital Archives, Gale Primary Sources & the Digital Scholar Lab

> The first step of the OCR software is to analyse the structure of the newspaper page. It divides the page into elements such as blocks of texts (columns), tables, images, etc. The lines are divided into words and then into characters. Once the characters have been singled out, the program compares them with a set of pattern images stored in its database. It analyzes the stroke edge, the line of discontinuity between the text characters, and the background. Allowing for irregularities of printed ink on paper, each algorithm averages the light and dark along the side of a stroke, and advances numerous hypotheses about what this character is. Finally, the software makes a best guess decision on the character. This character is given a confidence rating…A secondary level analysis may then take place at word level (since now a word is formed). The built-in English dictionaries and possibly dictionaries of other languages are checked to see if the word matches.
>
> Rose Holley, "How Good Can It Get?" (2009)

Creating Digital Archives

This OCR section will give you an understanding of HOW the text you are working with came to be created, where it came from and what factors influence its quality. As digital humanists, it is crucial to develop this awareness in order to make informed decisions about your data and its limitations.

Curious to try working with OCR text for yourself? ABBYY offers a 30-day free trial if you want to test the process.

Transkribus is an HTR (Handwritten Text Recognition) platform which is also worth exploring.

Reading/Viewing

Begin by reading the following articles:

How does OCR document scanning work? https://www.explainthatstuff.com/how-ocr-works.html

Rose Holley, 'How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs', *DLib Magazine* 15, n. 3/4, March/April 2009 http://www.dlib.org/dlib/march09/holley/03holley.html

What about OCR accuracy? https://www.hsassocs.com/what-is-ocr-accuracy/

Strange, Carolyn, et al. "Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers ." Digital Humanities Quarterly, vol. 008, no. 1, Apr. 2014.

Abstract*: "Digital humanities research that requires the digitization of medium-scale, project-specific texts confronts a significant methodological and practical question: is labour-intensive cleaning of the Optical Character Recognition (OCR) output necessary to produce robust results through text mining analysis? This paper traces the steps taken in a collaborative research project that aimed to analyze newspaper coverage of a high-profile murder trial, which occurred in New York City in 1873. A corpus of approximately one-half million words was produced by converting original print sources and image files into digital texts, which produced a substantial rate of OCR-generated errors. We then corrected the scans and added document-level genre metadata. This allowed us to evaluate the impact of our quality upgrade procedures when we tested for possible differences in word usage across two key phases in the trial's coverage using log likelihood ratio [Dunning 1993]. The same tests were run on each dataset – the original OCR scans, a subset of OCR scans selected through the addition of genre metadata, and the metadata-enhanced scans corrected to 98% accuracy. Our results revealed that error correction is desirable but not essential. However, metadata to distinguish between different genres of trial coverage, obtained during the correction process, had a substantial impact. This was true both when investigating all words and when testing for a subset of "judgment words" we created to explore the murder's emotive elements and its moral implications. Deeper analysis of this case, and others like it, will require more sophisticated text mining techniques to disambiguate word sense and context, which may be more sensitive to OCR-induced errors."*

Lecture

https://youtu.be/TOCazHs5wEE

This lecture was recorded for the Fall 2018 'Introduction to DH' class. Ray Bankoski and Michelle Fappiano from Gale joined the class via Zoom to discuss 'Making an Archive'. In it, you'll learn about the practical and business considerations that inform the selection and creation of the archives you're using in the Digital Scholar Lab. There's also an overview of what OCR and HTR are, and the limitations of both.

In Class

We'll begin working with the Digital Scholar Lab in class, starting with a demo, then you'll have an opportunity for hands-on work as you begin to gather material into content sets to investigate for your research project. You'll have an opportunity to look at some of the OCR output in the DSL in detail.

Week 4 Discussion

Choose 1-2 articles for each of this week's topics and summarize the main points for each piece, and highlight what you feel is most relevant to the work you will be doing in class. Write 200-300 words for each.

Then consider these questions:

1. **Week 4a - Project Management -** Developing a Project One-Pager
   This document will be a 'work in progress' for the next couple of weeks as you get to grips with the data and the scope of the work. A copy of this document will be included in the Final Project. Review notes from Monday's class to help you develop your initial draft document for this discussion post.
   - Project Name
   - Objective Statement
   - Requirements (see final project rubric)
   - Out of Scope
   - Team
   - Schedule
2. It's ok to change/edit the document as the scope of the work becomes clearer over the next couple of weeks.
3. **Week 4b** - OCR

What are the main challenges of working with OCR text? What factors can influence its quality?

Provide 1-2 screenshots as you develop a content set of what you feel is 'challenging OCR'. What is the OCR Confidence Level for this material and what does this term mean, in your own words?

What questions do you have about text analysis or OCR?

Replies: Due Tuesday of Week 5 11.59pm

Consider if you need to revise your Project One Pager to reflect the current status of your work.

WEEK 5a & 5b: Corpus Building/Preparing Texts for Analysis

Text Cleaning

Much of the text cleaning work you do in this class will be experimental - you may toggle cleaning options on and off to determine what effect they have on your content set. *You will not end up with perfectly cleaned documents!* In order to do this, it would be necessary to intervene with manual cleanup which is not an option currently in the DSL. It also takes a very long time!

The following reading will provide a framework for understanding WHY we clean data:

Katie Rawson, Trevor Muñoz, "Against Cleaning", http://curatingmenus.org/articles/against-cleaning/, July 6 2016

Miriam Posner, “Humanities Data: A Necessary Contradiction”, https://miriamposner.com/blog/humanities-data-a-necessary-contradiction/, June 25 2015

Julia Flanders, Trevor Muñoz, “An Introduction to Humanities Data Curation”, https://guide.dhcuration.org/contents/intro/

Words to numbers: text preprocessing choices

The following excerpt lists seven of the textual features that you may want to consider removing from your text in order to conduct a meaningful analysis. Essentially, we are working with a 'bag of words' model which is often mined quantitatively to extract patterns and frequencies. Many of these choices can be made by checking the appropriate box on the DSL cleaning page, but this list will give you some understanding of why you might want select one or more cleaning options.

Punctuation: The first choice a researcher must make when deciding how to preprocess a corpus is what classes of characters and markup to consider as valid text. The most inclusive approach is simply to choose to preprocess all text, including numbers, any markup (html) or tags, punctuation, special characters ($, %, &, etc), and extra white-space characters. These non-letter characters and markup may be important in some analyses (e.g. hashtags that occur in Twitter data), but are considered uninformative in many applications. It is therefore standard practice to remove them. The most common of these character classes to remove is punctuation. The

decision of whether to include or remove punctuation is the first preprocessing choice we consider.

Numbers: While punctuation is often considered uninformative, there are certain domains where numbers may carry important information. For example, references to particular sections in the U.S. Code ("Section 423", etc.) in a corpus of Congressional bills may be substantively meaningful regarding the content legislation. However, there are other applications where the inclusion of numbers may be less informative.

Lowercasing: Another preprocessing step taken in most applications is the lowercasing of all letters in all words. The rationale for doing so is that that whether or not the first letter of a word is uppercase (such as when that words starts a sentence) most often does not affect its meaning. For example, "Elephant" and "elephant" both refer to the same creature, so it would seem odd to count them as two separate word types for the sake of corpus analysis. However, there are some instances where a word with the same spelling may have two different meanings that are distinguished via capitalization, such as "rose" (the flower), and "Rose" the proper name.

Stemming: The next choice a researcher is faced with in a standard text preprocessing pipeline is whether or not to stem words. Stemming refers to the process of reducing a word to its most basic form. For example the words "party", "partying", and "parties" all share a common stem "parti". Stemming is often employed as a vocabulary reduction technique, as it combines different forms of a word together. However, stemming can sometimes combine together words with substantively different meanings ("college students partying", and "political parties"), which might be misleading in practice.

Stopword Removal: After tokenizing the text, the researcher is left with a vector of mostly meaningful tokens representing each document. However, some words, often referred to as "stop words", are unlikely to convey much information. These consist of function words such as "the", "it", "and", and "she", and may also include some domain-specific examples such as "congress" in a corpus of U.S. legislative texts. There is no single gold-standard list of English stopwords, but most lists range between 100 and 1,000 terms. Most text analysis software packages make use of a default stopword list which the software authors have attempted to construct to provide "good performance" in most cases. There are an infinite number of potential stopword lists, and for this class we are using the Glasgow list.

n-gram Inclusion: While it is most common to treat individual words as the unit of analysis, some words have a highly ambiguous meaning when taken out of context. For example the word "national" has substantially different interpretations when used in the multi-word expressions:

“national defense”, and “national debt”. This has lead to a common practice of including n-grams from documents where an n-gram is a contiguous sequence of tokens of length n. For example, the multi-word expression “a common practice” from the previous sentence would be referred to as a 3-gram or tri-gram (assuming stopwords were not removed). Previous research has tended to use 1,2,and 3-grams combined, because this combination offers a reasonable compromise between catching longer multi-word expressions and keeping the vocabulary relatively smaller. After extracting all n-grams from a document, a number of approaches have been proposed to filter the resulting n-grams, but here we choose to focus only on the most basic case of considering all 1,2, and 3-grams together without any filtering. So, the decision of whether include 2 and 3-grams (along with unigrams, which are always included) is the sixth preprocessing choice we consider.

Infrequently Used Terms: In addition to removing common stopwords, researchers often remove terms that appear very infrequently as part of corpus preprocessing. The rationale for this choice is often two-fold; (1) theoretically, if the researcher is interested in patterns of term usage across documents, very infrequently used terms will not contribute much information about document similarity. And (2) practically, this choice to discard infrequently used terms may greatly reduce the size of the vocabulary, which can dramatically speed up many corpus analysis tasks. A commonly used rule of thumb is to discard terms that appear in less than 0.5-1% of documents, however, there has been no systematic study of the effects this preprocessing choice has on downstream analyses. The decision of whether include or remove terms that appear in less than 1% of documents is the seventh and final preprocessing choice we consider.

List excerpted and adapted from Denny, Matthew and Spirling, Arthur, Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It (September 27, 2017), p6-9. Available at SSRN: https://ssrn.com/abstract=2849145 or http://dx.doi.org/10.2139/ssrn.2849145

Text Cleaning in the Digital Scholar Lab

Here's the 'Cleaning' DSL video: https://youtu.be/nDNyM6KPxcA

- In the Digital Scholar Lab, continue to search and build Content Set(s) related to your chosen research topic.
- In the Doc Explorer view, compare the side-by-side original image with its OCR text output carefully. Can you identify any recurrent errors in the text?

- Begin the process of creating a custom Cleaning Configuration to apply to your content set.
- Test your cleaning configuration by following the 'test configuration' process, downloading 10 cleaned and 10 uncleaned documents NOTE: THIS IS SLOW, ITERATIVE WORK!
- Repeat the process of tweaking your configuration, and re-testing.

From the DSL 'Help' Pages

Cleaning & Configuration

One of the most important elements of text analysis is making sure that your texts are formatted in a way that suits the kind of analysis you want to carry out. The Clean feature of the Digital Scholar Lab lets you edit all the Documents within a Content Set. It's designed to fit into existing Analysis Tools, as well as the download process for a Content Set. While we can edit or alter Documents on the fly for particular tools, cleaning is a critical part of the preparation for any text analysis. The Digital Scholar Lab breaks it out as a separate feature, so you can ensure that the Documents in different Content Sets are prepared in precisely the same way, and that you, as a scholar, can decide how they're altered, and make adjustments according to your research needs. It will be a one-stop shop for tinkering with the texts you want to Analyze before sending them off to an Analysis Tool Job.

Configurations & Replication / Method

The Clean feature is designed as part of the broad commitment to transparency and method within the Digital Scholar Lab. We want you to be able to replicate or reproduce your results with the same Content Set, or compare similar methods across different Content Sets. Clean lets you build a Configuration you can reuse, alleviating the need to remember 'what you did', as much as allowing you to return to your analysis easily after being away from the Digital Scholar Lab for a period of time. In short, a Configuration creates a kind of standardized method for preparing documents that you can send for Analysis, and lets you use that standardized preparation - like a cookie cutter - for any of your Content Sets, combined with any Analysis Tool.

Default & Custom Configurations

Clean allows you to create Configurations which you can reuse or associate with a specific Analysis Job for an Analysis Tool. Default Configurations can be used directly, or they can act as a template or starting point for the creation of new Custom Configurations. As new Tools come online we'll provide you with a series of Default Configurations best suited to specific Tools,

which you can tweak and alter as you see fit. You can save the Configurations at any time - just provide a new name, and description, which will help you pick the correct Configuration from a dropdown list in a Tool configuration area.

Select the text cleaning configurations

Each Configuration consists of a series of correct, removal and replacement or substitute options, alongside a possible stop word list. In theory, you can have a Configuration that's empty - it won't do anything, but the algorithm would still use it before running an Analysis Job.

Corrections

- The only correction currently available is case correction, altering all text to lower case. This is useful in contexts where an Algorithm or Tool might be case sensitive and have no internal options to alter cases.

Removal

- Remove all number characters
- Remove all special characters. Users can set specific special characters to remove, such as currency symbols, slashes, underscores etc. Remove all punctuation. Users can set specific punctuation to remove

Replacement

- Reduce multiple spaces to one space (ex: "hello there" becomes "hello there")
- Replace _____ with _____ allows users to define what kinds of replacements they'd like to make on the fly. This function is useful for controlling orthography or spelling variants, e.g. all instances of 'colour' can be altered to 'color', to make sure that the Analysis Tools treat them as the same token, not distinct words.

Stop words

You can also decide to use a stop word list as part of your configuration. The default stop word list contains English words; however, you can edit this list as you see fit. Each stop word should be listed on a separate line. If you'd like, you can cut and paste an entirely new stop word list here.

Configurations and Tools

You can use any configuration on any Tool in the DS Lab environment. That said, different selections will impact certain tools in specific kinds of ways. For instance, MALLET - the software powering the Topic Modeling Tool - is case sensitive. If you decide to make everything lower case, it won't distinguish between Smith (perhaps someone's last name), and smith (an occupation, like a blacksmith). MALLET also fails handles possessive apostrophes in slight awkward manner, turning them into their own words - you can add 's to the Stop Word list to prevent this from happening. In some Sentiment Analysis Tools, punctuation, specifically periods, matter because they carve up Documents by Sentence. In the current version of the Sentiment Analysis Tool, this doesn't matter, but later versions will allow you to calculate Sentiment Scores in different ways for a Document. Removing all special characters will also remove currency symbols like $, £, ¥, or €, which might be important for tracking amounts. The Ngram tool, and many others, Tokenize, or cut up, Documents using whitespaces. It's prudent to replace all tabs or other characters (that might have slipped by our OCR processing) with single spaces to make sure that Documents are Tokenized appropriately. Because of this, we've pre-selected the whitespace options in our Default Configuration; you can change these, if you'd like.

In the end, there are no 'incorrect' Configuration options. But it is important to note how certain choices will affect or shape the results of an Analysis Job. Often the best method is a combination of reading up on what kinds of analyses you'd like to carry out, understanding how certain preparations can affect them, and testing them out. If you find something isn't quite right, the great thing is that you can change your Configuration, and run the Job again!

Name and save the cleaning configurations

When you first open up the Clean feature you'll see the 'Default Configuration', which contains an English stop word list, and a couple of configurations checked. You can alter this Default and save it as something different, simply provide a new Name, and if you think it's useful, a description. Each time you view and alter a Configuration, you can either save it, rewriting over the existing Configuration, or you can save it as something new, using save as. This allows you to create any number of Configurations for your work! Remember, though, to name them so you can remember the settings you've chosen! Using a kind of shorthand in the name helps, e.g. No Punctuation - Lower case - English for a Configuration which removes all punctuation, transforms all text to lower case, and uses the English stop word list.

Review the output of the cleaned content set

Most researchers will want to see what a Cleaning Configuration does to their texts before using it on a large Analysis Job. This is understandable - often it helps to look at edits to appreciate how they might affect a larger computational task even though we have an expectation of what they should do, it's nice to know what the output might look like.

Sample

To check out your Configuration, just click on Test Configuration in the top Control Bar, and select the Content Set you'd like to test it on. This will submit the Cleaning Job and return a sample of 10 documents (original and cleaned texts; 20 documents total) from the Content Set for you to review. If you like what you see, you're good to go! If you don't like the results, you can then make changes accordingly, and rerun the test.

Content set up to 5,000 documents

At the moment Downloads are limited to 5,000 Documents per Content Sets. For Content Sets over 5,000 Documents, a Download will contain a randomized 5,000 Documents from your Content Set. You may apply a cleaning configuration at point of download to ensure you receive a cleaned version of your selected texts up to 5,000 documents.

Pass it through an analysis tool to generate a visualization

The last step for Clean is putting your Configuration into action. When you select an Analysis Tool, click on Configure, and you'll find your Configurations listed in the drop-down box at the top of the Configuration panel. Select the one you'd like to use for this tool, and then Run your Analysis Job as usual.

Cleaning Practice outside the Digital Scholar Lab

Option 1: Researchers have the option of downloading up to 5000 documents of OCR text from the DSL to clean or analyze outside the platform. While this allows for a little more flexibility in cleaning individual documents at a granular level, at the moment it's not possible to re-upload your cleaned documents in the DSL. So if you download your .txt files, chances are that you plan to analyze them outside of the DSL. However, for the purposes of this Module, it will be a useful exercise to explore a different cleaning options outside the DSL which are commonly used by researchers in digital humanities. There are three options listed below, and you're welcome to explore each of them.

Start by downloading your uncleaned content set (assuming it's less than 5000 documents) to your computer. You can do this from the 'My Content Sets' page in the Digital Scholar Lab.

Option 2: Use the data/primary source diaries and letters in the Data folder.

Lexos

Lexos is an open source, web-based tool which enables the researcher to upload, clean and analyze text material. It was developed and is maintained by Wheaton College.

From the Wheaton College Lexomics Department Website:

"Lexos is a web-based tool to help you explore your favorite corpus of digitized texts. Our primary motivation is to help you find the explorer spirit as you apply computational and statistical probes to your favorite collection of texts. Lexos provides a workflow of effective practices so you are mindful of the many decisions made in your experimental methods."

Full details are available here, including options to install on local machine. For the purposes of this class, I recommend their server instance:

http://lexos.wheatoncollege.edu/upload

The video tutorials are comprehensive and well worth watching. The full playlist is available here.

Summary: In this activity we will explore the impact of text cleaning on our data sets using the Lexos tool.

Instructions:

1. Open up your selected primary source documents from the Data folder and/or use OCR text downloaded from the Digital Scholar Lab.
2. Review the documents and consider any issues with the text. Consider how you would address those issues to get the best result.
3. Open up the Lexos tool using this link: http://lexos.wheatoncollege.edu/upload

4. Upload your data set by clicking the “Browse” button and selecting the files in your folder. You can also drag and drop the individual files into the “Drag files here” container. Successful uploads will appear in the “Upload List.”

5. Once your files have been uploaded, click the “Prepare” link in the upper right hand corner and click the “Scrub” option in the drop down list.

6. Take a moment to experiment with the options on the page.

    a. Use the Previews section and click the “Preview” button to review your changes and see the impact on the text.

    b. Here are steps you can follow to use the Stop/Keep Words option

        i. Download the English_StopWord_List.txt.file from your Group folder

        ii. Paste the words in the English_StopWord_List.txt file in the text box

        iii. Click the “Stop” option

    c. Navigate to the “Previews” section on the page and click “Preview” button to see the changes to your text files.

    d. Tweak the English_Stop_Words.txt file and repeat the steps above until you’re happy with your results.

    e. Once you feel the text is in a good state, click the “Apply” button.

7. Now navigate to “Visualize” and select the Word Cloud option.

    a. Examine the terms in the word cloud

8. Navigate to the Scrub page again and download your cleaned data set by clicking the “Download” button in the upper right hand corner of the “Preview” section

The next two methods - Regular Expressions and Open Refine are considered 'medium difficulty'. Don't let this put you off! The Programming Historian tutorials will walk you through the process in each case and give you some exposure to methodologies you have perhaps not tried before. Make sure you take notes as you work, both for your assignment and as general good practice so that if you make a mistake, you can backtrack and pick up at the point where things went wrong.

Regular Expressions

The OCR texts you download from the DSL will be .txt files, or plain text. Plain text documents contain no hidden extra code. Word processor documents (.doc files, for example) can only be opened with word processors. Plain text documents can be opened with all and any text editors. Recommended editors which are free, even if you do get occasional popup boxes asking if you want to purchase include Sublime, Atom.io , BBEdit and Notepad++

Tutorial: Laura Turner O'Hara, "Cleaning OCR'd text with Regular Expressions ," The Programming Historian 2 (2013).

Another tutorial to consider: https://www.regular-expressions.info/quickstart.html

Additional Resources:

Beth Seltzer, "Text Scrubbing Hacks: Cleaning Your OCRed Text", https://sites.temple.edu/tudsc/2014/08/12/text-scrubbing-hacks-cleaning-your-ocred-text/

'Understanding Regular Expressions' https://github.com/OpenRefine/OpenRefine/wiki/Understanding-Regular-Expressions

Open Refine

Find Open Refine here, along with a selection of tutorials. Per the developers, it is: a *free, open source, powerful tool for working with messy data*

While you may not have time to appreciate all its features this course, the following tutorial will give you a taster.

Tutorial: The Programming Historian 'Text Cleaning with Open Refine ' (Seth van Hooland, Ruben Verborgh, and Max De Wilde, 2013)

Week 5 Discussion

This week's discussion post has a couple of elements which you should have completed already, focusing particularly on your record of hands-on work and research question development.

1. Include a summary of reading you completed this week related to text cleaning. What challenges did you encounter as you experimented with text cleaning in the DSL and/or with Lexos? (digits bug notwithstanding).

2. Provide a complete record of your work this week. Include:

- details about the development of your content set(s).
- a discussion about how the development of your content set(s) affected the scope of your research question.
- the search terms you have been using to build your content set. This can be a screenshot from the 'My Content Sets' area of the DSL, or a typed string of text.
- a summary of the challenges have you encountered. How did you address them?

Replies: Due Tuesday of Week 6 11.59pm

Remember to revisit your Project One-Pager regularly to update the objective summary as you continue to refine it.

WEEK 6a Microhistory & Text Encoding

Microhistory as a theoretical & methodological approach to writing history

First developed by Italian historians in the 1970s as an experiment, microhistory swiftly became one of the most innovative ways of researching and writing history. The first microhistorians were 'born' through their dissatisfaction with predominant social history methods that focused on broad subjects over very long periods of time.

Like all good histories, a microhistory begins with a research question or a set of questions. It's the second step that distinguishes microanalysis: the **reduction of the scale of analysis**, sometimes drastically.

The Methods of Microhistory

In addition to zooming in on an individual, a community or a unique event, a historian might use other microhistorical practices to illuminate the past, perhaps switching to narration not just to "tell a story" but also as a method of shedding bright light on hidden aspects of a historical person or group of people: for example, Karen McCarthy Brown's *Mama Lola,* or Alessandro Portelli's *The Order Has Been Carried Out*.

Some commonly used methods and interpretive microhistorical practices include:

- Privileging first-hand accounts ('ego documents') to explore historical actors' experiences.
- Tracking clues through multiple sources to **discover hidden connections**, like a sleuth following every lead to its smallest detail to see where those details unexpectedly collide.
- Reconstructing webs of social networks.
- **Scaling an analysis down or up** to highlight specific historical contexts and perspectives.

Microhistory draws on the fields of cultural anthropology, ethnography, and literary and philological studies, among other disciplines.

Microhistory reinfuses the past with its own vibrant energy because finely crafted microhistories capture the drama of everyday life. They let readers understand people as agents of change for the worlds they live in, often in the face of overwhelming difficulties. For example, the return of a soldier to his village becomes a riveting tale about identity and imposterhood in Natalie Davis's *The Return of Martin Guerre*. An exchange of gifts among women tumbles a land into panic, revealing the frightening and hidden dynamics of witchcraft driven into the life of one woman at the center of the storm, as told in Thomas Robisheaux's *The Last Witch of Langenburg*. Rumors of a slave uprising that spread through a city allow Jill Lepore in *New York Burning* to reveal the dynamics of race at the street level in New York of the 1740s. The way women create whole worlds in a society is sometimes best explored through microhistory, like Laurel Ulrich's *A Midwife's Tale* (see also resources below), or Jon Sensbach's account of a slave woman missionary who helped inspire the rise of black Christianity in the Atlantic world, in *Rebecca's Revival*.

Microhistories can bring to light the experiences of everyday people in big, well-known historical events, sometimes in ways that challenge the common wisdom, surprise and even shock us. For instance, the legacy of the German Occupation of Rome in World War II looks entirely new when told through the stories of everyday Romans collected, re-told and interpreted by Alessandro Portelli in *The Order Has Been Carried Out*. James Goodman's *Stories of Scottsboro* shows in acute brushstrokes how everyday racial oppression in the Jim Crow South was even more violent and harrowing than many histories reveal.

Today microhistories often serve as correctives to grand historical narratives, big theories, and Big Data studies. Well crafted microhistories discover microworlds of experience barely glimpsed at larger scales of historical study, illuminating the dynamics of human history in rich colors and textures.

Microanalytical Methods

- Reduce the scale of analysis and use different scales of analysis. Finding just the right focus is the mark of well-designed microanalysis.
- Creatively use narrative to craft a story or construct an analysis.
- Identify and interpret "ego documents" – objects or parts of documents that reveal historical figures' own perspectives on their experiences.
- Design a project around the "exceptional normal," an event that seems unusual, striking or sensational, but which can become a window onto important everyday patterns, values or ideas.

- Identify and track clues in documents to discover hidden connections. This method involves learning the unique features of different genres of sources, the ways they were created, the language and concepts that shape them, and in the process heightening your sensitivity to be able to identify something striking, unusual or important that you might have otherwise missed.
- Deploy historical contexts for meaning and interpretation of evidence or sources. Every source points to multiple historical contexts. Which ones should you be alert to?
- Do social network analysis. Most people live life in a web of relationships: family, friends, neighbors, and others. Learn to track them, identify the significant ones for your study, and interpret them to reveal hidden or subtle social dynamics.
- **Digitally visualize historical networks** to reveal connections that might otherwise go unnoticed.
- Traditionally biography is one way to understand the individual in her or his historical world. Microhistory allows other ways to foreground one small moment or aspect of an individual's life.
- Discern the difference between a case study and microhistory. Frequently confused with the case study, microhistorical methods set up an investigation around singular, unique objects, not patterns or "cases." What can the singular and the unique teach us?
- Use trial records and other tricky sources. Microhistories have pioneered new ways of using records that have in the past been dismissed as too biased.

A microworld **involves something singular, specific, unique** — the world of experience of an individual or group of connected individuals at a particular moment in time. It's a slice of intimately lived history, a personal story that reveals hidden details of the past — conflicts or beliefs or values — that broad histories can often glide over or smother.

You can build a microworld from a person or group of people, but also from a single object or event, making that object or event the center of your microhistorical analysis.

Credit: microhistory summary adapted from Tom Robisheaux, Duke University

Reading and Resources

"Microhistory ." Encyclopedia of European Social History . . Encyclopedia.com. (February 1, 2020). https://www.encyclopedia.com/international/encyclopedias-almanacs-transcripts-and-maps/microhistory

Giovanni Levi, “On Microhistory,” in Peter Burke, ed., *New Perspectives on Historical Writing* (1991)

Sigurdur Gylfi Magnusson, ’What Is Microhistory?’, *History News Network* https://historynewsnetwork.org/article/23720

http://www.microhistory.org/ *is run by the Center for Microhistorical Research at the Reykjavik Academy in Iceland. This is a website for the various projects about microhistory, including bibliography on published microhistorical works, studies on memory and postmodernism.* (note: the website doesn’t seem to be maintained any more)

Examples of Microhistory Works/Projects

- Natalie Zemon Davis, The Return of Martin Guerre (1984) describes a soldier’s return to his village. See also this Wikipedia entry describing the 1982 movie. Interestingly, Zemon Davis’s book stems from her work as the historical consultant on the movie. Zemon Davis’s summary biography is relevant. See also Robert Finlay, ‘The Refashioning of Martin Guerre’, The American Historical Review, Vol. 93, No. 3 (Jun., 1988), pp. 553-571.
- John Brewer, *Sentimental Murder: Love and Madness in the 18th Century* (2004) describes the Martha Ray case of 1779: bloody murder on the steps of a theater.
- Laurel Thatcher Ulrich, *A Midwife's Tale*, by Laurel Thatcher Ulrich describes Martha Ballard’s diary. More information here, although the site is a little dated. Cameron Blevins also carried out a topic modeling project on the same material.
- Thomas V Cohen’s *Love and Death in Renaissance Italy* is a record of a dying woman’s conversations.
- Storytelling and the Global Past is a discussion between three microhistorians including a discussion about the secret life of a globe-trotting religious convert,

- Donwon Shin, Applying the Methodology and Practice of Microhistory: The Diary of a Confucian Doctor, Yi Mun-gon (1495-1567), *Korean Journal of Medical History* 2015;24(2): 389-422.
- Farah Griffin talks about her book *Harlem Nocturne: Women Artists & Progressive Politics During World War II* which focuses on three African-American artists.
- Peter Marshall, *Mother Leakey and the Bishop: A Ghost Story* (2007) describes sightings of a ghost on Halloween.

Introducing the Text Encoding Initiative (TEI)

Overview and Purpose

We'll take a broad look at XML and the TEI guidelines, and practice analyzing a selection of documents for structure and content.

Resources

TEI Consortium's website at http://www.tei-c.org/index.xml

TEI guidelines at http://www.tei-c.org/Guidelines/P5/

TEI by Example at http://www.teibyexample.org/TBE.htm

TEI mailing list <TEI-L@LISTSERV.BROWN.EDU>

TEI wiki at http://wiki.tei-c.org/

Further resources provided by the TEI council and Oxford computing centre:

ROMA at https://roma2.tei-c.org/ = customizing TEI schemas for XML validation

OxGarage : https://github.com/TEIC/oxgarage

http://oxgarage.tei-c.org/

online resource for conversion between common file formats, using TEI P5 as pivot format. Can be used to produce TEI P5 XML from a .docx file.

DHOxSS at http://digital.humanities.ox.ac.uk/dhoxss/ = providing the material (including slides and exercises) for years of summer schools.

TEI Boilerplate at http://teiboilerplate.org/ = a light-weight solution for publishing styled TEI P5 content directly in modern web browsers (the implementation uses XSL which is processed by the browser).

Sample TEI Projects:

TAPAS Project - http://tapasproject.org/

Shelley Godwin Archive

Digital Mitford

Women Writers Project

In Class

Take some time to work through the TEI Guidelines, looking specifically at the groups of element sets.

Take a look at these texts: https://drive.google.com/drive/folders/1-hAkjB1bc8ximHaDveIGXMpSg3dJnUI5?usp=sharing

What do you think it is important to capture in your markup for each document?

Referring to the TEI Guidelines, which TEI tags would you consider using?

WEEK 6b Qualitative or Quantitative? Considering which tool to choose through the lens of sample projects

---

Sample Project #1

The Rise of Electricity in the late 19th & Early 20th Centuries

Synopsis

The harnessing and production of electricity is one of the defining watersheds of the late 19th century. This source of power, which acts and looks unlike anything other than the supernatural, opened numerous scientific and economic doors, while terrifying and amazing peoples and societies with its potential. What electricity might do was unclear - it could power the new machines and technologies of the industrial revolution, but could it also reshape and repower, or transform, the human body and self, too? As a substance electricity shared much with light and spirit, which were two crucial paradigms for how human beings understood themselves and their place in the natural and spiritual worlds they inhabited. That electricity could be coursing through human bodies was a truly astounding idea. And so electricity itself is a topic that bridges science and spirituality, industry and invention, as well as fantasy and reality. After all, electricity played a role in Frankenstein as much as it did the lightbulb.

This project compares how electricity features in two serial publications between the end of the Civil War in 1865 and the end of the First World War in 1918.

- Banner of Light one of many spiritualist journals / serials - runs from 1857 to 1907
- Scientific American, a leading science journal runs from 1845 to the present

Core Research Question:

- How do two serial publications treat the topic of Electricity in the late 19th & early 20th century?
- How can the same topic be compared between two serial publications, or for that matter, two distinct content sets?

More Precise Questions:

1.What themes are evident in the Banner of Light and Scientific American when discussing the topic of ‘electricity’?

2.Do the Banner of Light and Scientific American share any topics or similar points of view in their treatment of electricity in the late 19th century?

3.What sentiments appear in discussions of electricity?

4.Are there any other distinguishing features surrounding electricity in these journals that may reflect on contemporary views of industrialization, invention, and their effects on men and women in the late 19th and early 20th centuries?

Thinking about Methodology & Specific Tools

- Topic Modeling - we can use this tool to see if there are any themes or topics which cut across a collection of texts
- nGram - we can use this tool to track different kinds of phrases or terms which might occur together, and the number of times a phrase appears
- Sentiment Analysis - we can use this tool to examine whether the contents of the document were overall positive or negative according to the AFINN dictionary.

Building the Content Set

Searching

The initial content sets were constructed with the same variables in mind, but distinct serial publications. The archive used for both was the American Historical Periodicals. The first serial content set is derived from ‘Banner of Light’ as the publication title, and the second ‘Scientific American’

Limits & Parameters

Content Type ("Article" or “Essay”)

Archive - American Historical Periodicals

Publication Date (1865-1918)

Keywords (in individual rows)

Entire Document: Electricity

Statistics & Info

Banner of Light

Content Set Name: 1865-1918 Banner of Light - Essays and Articles

Content Set ID: 1580766975821

Number of Documents: 1296

Scientific American

Content Set Name: 1865-1918 Scientific American - Essays and Articles

Content Set ID: 1580767055855

Number of Documents: 3356

Specific Tools

None of the tools required specific content sets. It was easier to create cleaning configurations that removed all punctuation, set all characters to lowercase, and also removed extended ASCII characters. Numbers were also removed.

Specific Questions

Question 3 suggests that the sentiment analysis tool might be an option. But the sentiment analysis tool provides a metric using the AFINN dictionary - it doesn't necessarily tell researchers what specific sentiments may be present within a text. In this respect, Topic Modeling and the nGram tool allow us to explore the kinds of phrases and words which could support and help contextualize and explain the results of the sentiment analysis tool. This is an example of understanding how the results of one tool can buttress or reinforce, or even help explain, the results of another.

Cleaning the Content Set

Specific research questions:

The content sets were both cleaned for punctuation, as well as numbers, and special characters as above. However, each required the creation of their own specific stop word lists. The Banner of Light was published in Boston, Mass. for most of its existence, and referenced other spiritualist centers in the United States. Since our research questions are focused on electricity, rather than places, it made sense to include state abbreviations as well as common cities and place names such as Boston, Mass., Washington, Philadelphia, etc. Also, since each content set is derived from a periodical with advertisements, columns, and sections, it made sense to remove common words associated with prices, publication sections, like pages, etc. These were far too common when running Topic Modeling and nGrams; putting them on the stop word lists helped to clean up the 'noise' in these results.

Running Tools

Selecting particular views for each tool was extremely straightforward. We selected an approach that reflected the size of our content sets: we looked for more things and raised the bar for what made the cut for the results. Both were for very simple reasons: Topic Modeling as a tool statistically discerns what words are more likely to appear near to one another. More topics, and more words lowers the threshold of what is 'significant', meaning we get a finer grained picture of what the statistical analysis could suggest. In very similar documents, like those appearing in serial publications like Banner of Light and Scientific American, the chance is that there will be similar words related to advertisements, questions posed by readers, comments and notices, etc. Selecting more words and more topics is a good way of sifting through some of these 'known' similarities, and can work in tandem with stop word lists to help 'drill down' into a large content set. For nGrams, we took a similar approach to thinking about potential 'noise' - we want to see what turns up. But the highest count in a result doesn't always translate into the most meaningful or interesting. There's a balance between number and noise.

Topic Modeling

It seemed best to cast a wider net in part to see what kinds of words appeared in the models created by the MALLET software that powers the tool. Requesting more words than the default,

and double the topics produces finer grained topics, in reflection of the size of the content set. We opted for 15 word topics, and 20 topics.

Sentiment Analysis

This tool has no settings other than selection of the cleaning configuration.

nGrams

Like Topic Modeling, it seemed worthwhile to go beyond the default settings given the size of our content set. We raised the threshold for the number of times an nGram had to appear to be considered useful, and set it at 4. Equally, we wanted to find collocates rather than just single words, so we set the minimum nGram size to 2 (biGram), and the maximum size to 5. These settings translate into a search for "nGrams of between 2 to 5 words that appear in documents at least 4 or more times".

Understanding Results

This project involved numerous iterations of cleaning configurations to obtain initial clear results.

Topic Modeling

There was some clear overlap, as expected, between the two periodicals when it came to Topic Modeling, but also distinct topics and concerns. Banner of Light's topics focused more on the self and the body, and how electricity fit into human existence - not surprisingly given the spiritualist nature of the publication. Scientific American shared some of these concerns, but was also very interested in the use of electricity as an invention or means of bettering society. Where the two seemed to overlap is around the idea of betterment, often seen somewhat with words revolving around health, medicine, and discovery.

Banner of Light

- life,spirit,man,spiritual,human,mind,nature,power,thought,soul
- powders,positive,diseases,cure,office,cured,negative,sent,disease,healing
- medium,table,room,hand,said,hands,came,spirits,spirit,saw

- cloth,light,spiritual,paper,book,banner,free,place,rich,white
- force,matter,form,motion,light,forms,atoms,heat,substance,forces
- electricity,electric,life,brain,blood,current,water,body,electrical,air
- spiritualism,phenomena,science,facts,scientific,spirits,subject,fact,truth,spiritual
- medical,healing,medicine,disease,health,practice,physicians,law,treatment,physician
- spirit,spirits,spiritual,earth,medium,life,body,power,form,conditions

Scientific American

- water,steam,inch,boiler,power,engine,pressure,iron,pipe,use
- apparatus,prof,valuable,steam,contained,description,iron,method,electric,interesting
- lightning,electricity,earth,animal,plants,rain,death,ground,electric,animals
- electric,power,light,motor,electricity,car,horse,lamps,engine,lighting
- science,professor,scientific,prof,society,electrical,electricity,american,discovery,year
- light,force,motion,heat,matter,electricity,energy,sun,theory,rays
- current,wire,electricity,electric,battery,magnet,iron,machine,wires,placed
- telegraph,telephone,bell,instrument,wire,line,patent,cable,apparatus,wires
- Battery,wire,use,power,cells,writes,motor,current,coil,used

For this project, we ‘named’ our Topics in the results. This is purely optional, and it’s crucial to note that researchers have to come up with their own interpretations of what the lists above might represent as a coherent ‘topic’, rather than as a statistically created list of words. But once this is done, we have a better sense of what the Topic Modeling comparison view can do to help us understand specific metrics and the topics created by the tool.

Before moving to the topic comparison view, it’s worth looking at a single topic. Here we see on the left a summary of the topic measures, and how many documents where the tool has

identified the topic. On the right, the four columns each contain information on the number of

Body and Electricity

IDENTIFIED IN

500 DOCUMENTS

TOPIC MEASURES

Tokens	21164
Document Entropy	5.7517
Average Word Length	6.2
Coherence	-64.3552
Uniform Distance	4.0033
Corpus Distance	2.9952

TERMS	COUNT	PROBABILITY	DOCS
electricity	411	0.0194	248
electric	375	0.0177	204
life	212	0.01	122
brain	205	0.0097	94
blood	196	0.0093	93
current	192	0.0091	104
water	191	0.009	109
body	191	0.009	97
electrical	160	0.0076	91

words you've selected for your topics, and the number of documents in which the word appears.

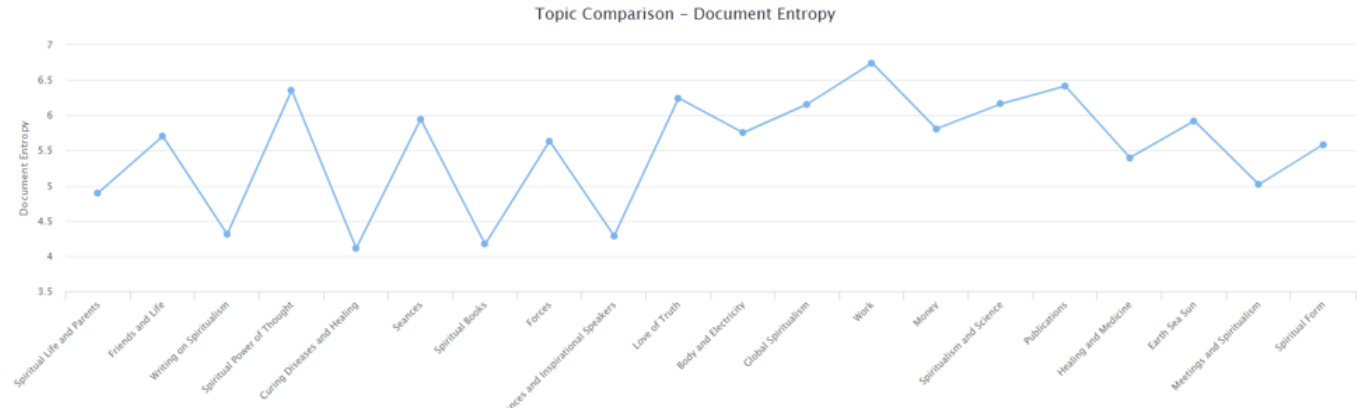
Examining the Topic Modeling comparison view we find specific metrics which can help us work through what kinds of subjects our Content Sets contain. The most important thing - unsurprisingly - is that the documents also discuss topics that have nothing to do with Electricity. That said, electricity does appear alongside other related words such as 'force', 'energy', 'light','spirit', etc.

Banner of Light

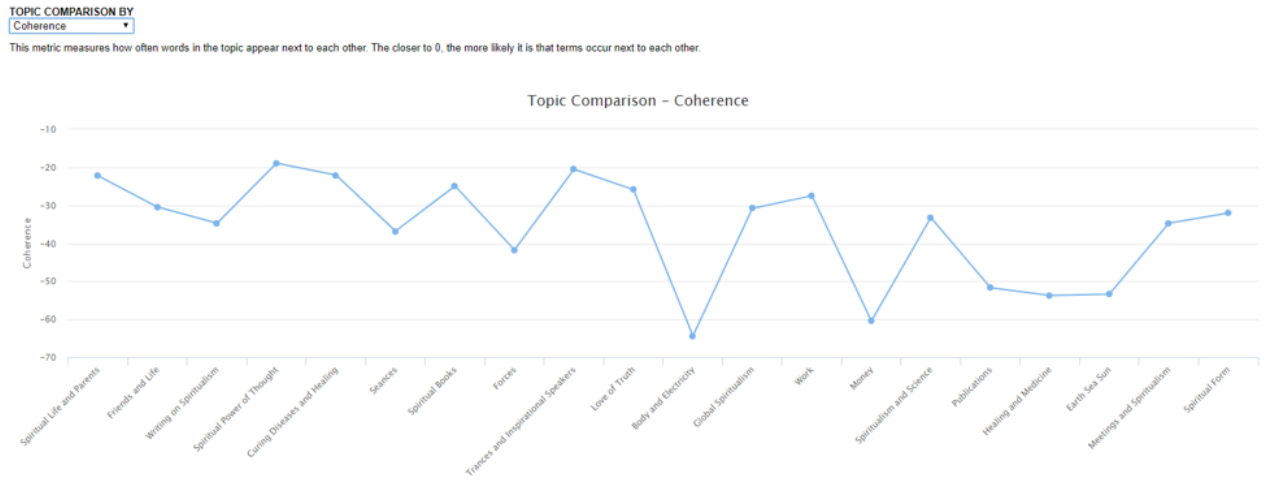
TOPIC COMPARISON BY

Document Entropy

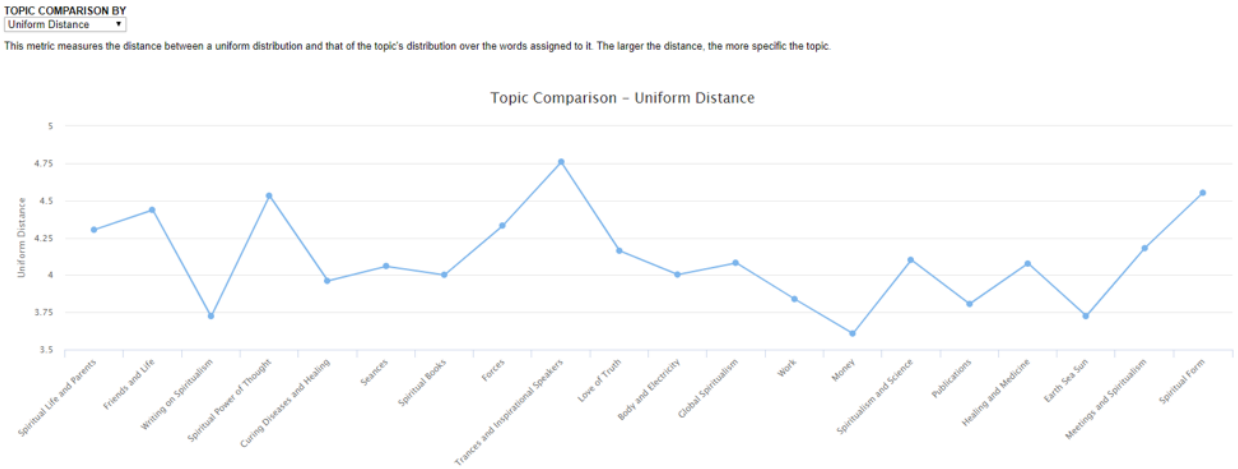
This metric measures the probability any given document will be in the topic. Low entropy topics will come from a small set of documents while higher entropy topics will come from a wider set of documents.



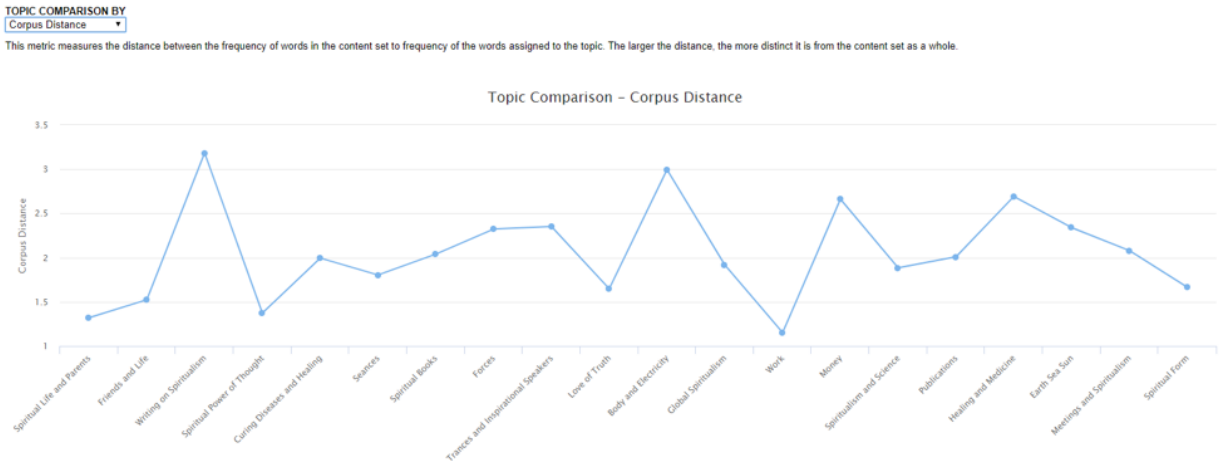
Document Entropy offers a way of thinking about the prevalence of topics within a content set as indicates how extensive a topic is across all the entire set. In the example above, the topic we've named as 'work', is the most prevalent. The next is 'publications', and the third 'spiritual power of thought'. The results suggest that electricity really doesn't appear as a distinct topic across the entire content set. Despite its presence in various topics, it's not pervasive.



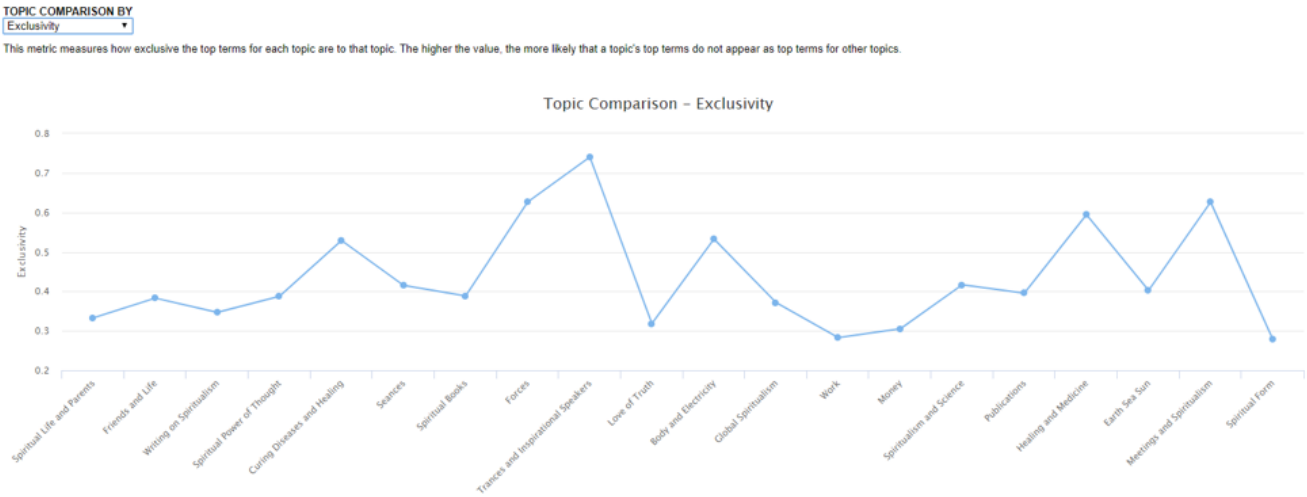
Coherence provides a metric that suggests how closely knit the words in the topic are within the texts. Since a topic is composed of words that have a greater statistical chance of appear near each other, this shows how great that proximity actually is for a topic. Notably out of all the topics, ‘body and electricity’ has the lowest coherence, indicating that the words that the tool suggest are a topic, are in fact spread out further from one another in contrast to those which make up the topics ‘spiritual power of thought’ or ‘trances and inspirational speakers’.



Uniform distance essentially compares a topic to the ways in which words are distributed within texts. Conceptually, it’s related to the coherence metric, but instead of comparing the distance of words within a topic, it compares those to how all words are distributed, and in relation to the topic words themselves. This metric helps discern how specific a topic might be - in this case ‘trances and inspirational speakers’ is the highest, with ‘spiritual form’ coming in second.



Corpus distance offers a metric for ascertaining how unique the words that make up a topic might be within a content set. The frequency measure shows how distinct words within a topic are from the entire content set. In this case 'body and electricity' comes in a very pronounced second, suggesting that this topic really is quite distinct from the rest of the language within a content set.

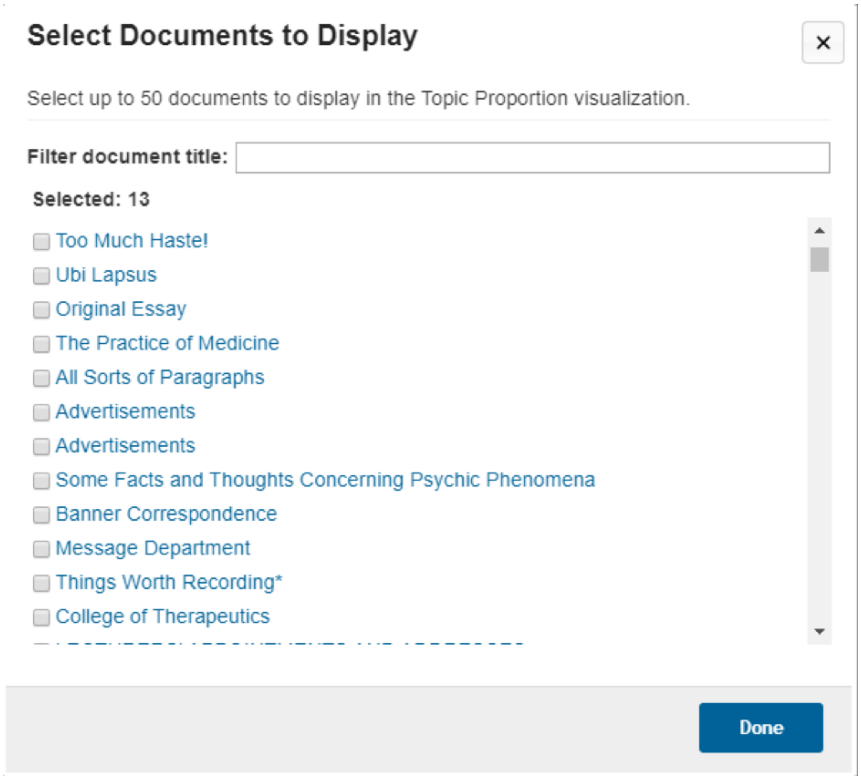


Exclusivity shows us another way of thinking about uniqueness or distinctiveness of a topic in relation to the broader content set. Since each topic is made up of a group of statistically proximate words, there's a good chance those words overlap with one another. The degree to which they don't - ie that they are distinct to a specific topic - suggests that the vocabulary that composes a topic is itself limited to that topic, making it more 'exclusive' rather than woven into and possibly appearing within other topics. Here we see 'trances' yet again, but also 'forces',

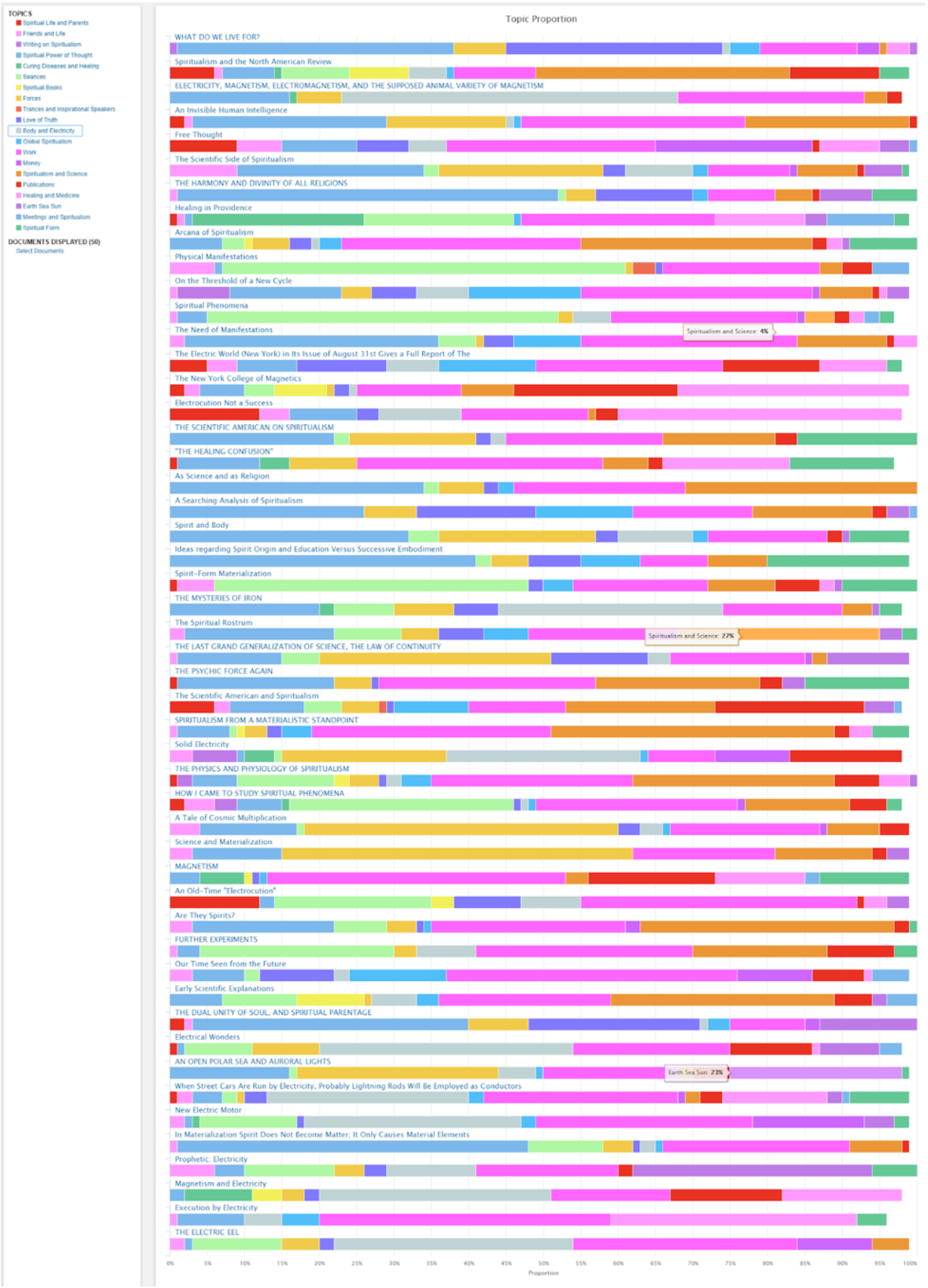
'healing and medicine', and 'meetings and spiritualism'. The words that make up 'body and electricity' are 5th in terms of exclusivity.

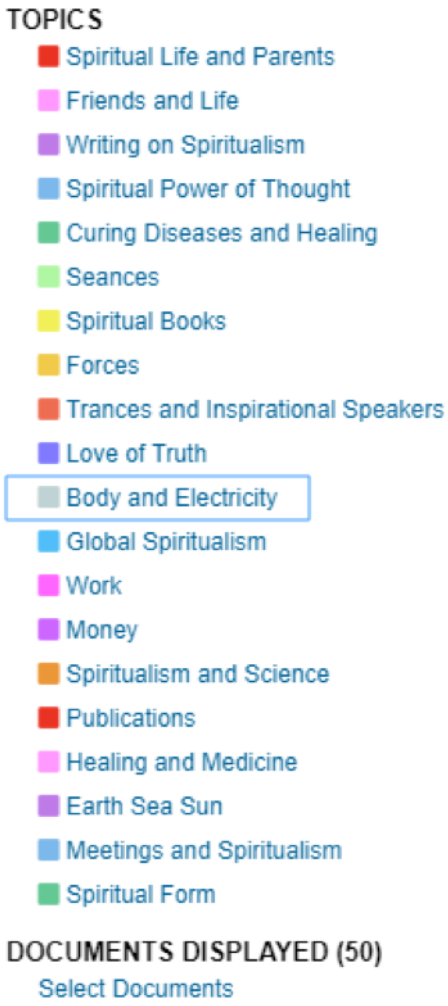
Configuring the Topic Proportion

The Topic Proportion view of the Topic Modeling tool, permits us to compare how much of the texts within a subset of a content set are allocated to each topic. This visualization provides a quick means of seeing how prevalent topics are against one another, rather than always relying on numbers. The visualization is limited to 50 documents, which are initially randomly selected from the content set. But you can select your own and refresh the visualization.



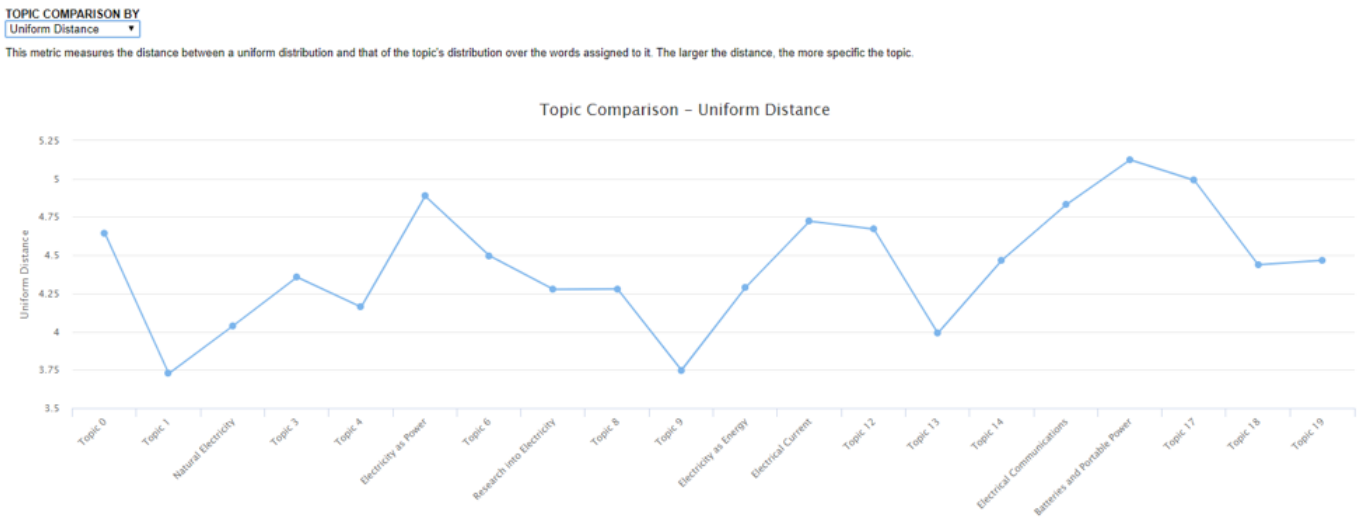
Along with the Topic Comparisons, it helps to create names for your topics when using this view. If you click on a specific topic, the visualization will shift to display only the percentage of each text where the topic appears.


TOPICS
Spiritual Life and Parents
Friends and Life
Writing on Spiritualism
Spiritual Power of Thought
Curing Diseases and Healing
Seances
Spiritual Books
Forces
Trances and Inspirational Speakers
Love of Truth
Body and Electricity
Global Spiritualism
Work
Money
Spiritualism and Science
Publications
Healing and Medicine
Earth Sea Sun
Meetings and Spiritualism
Spiritual Form
DOCUMENTS DISPLAYED (50)
Select Documents
Topic Proportion
WHAT DO WE LIVE FOR?
Spiritualism and the North American Review
ELECTRICITY, MAGNETISM, ELECTROMAGNETISM, AND THE SUPPOSED ANIMAL VARIETY OF MAGNETISM
An Invisible Human Intelligence
Free Thought
The Scientific Side of Spiritualism
THE HARMONY AND DIVINITY OF ALL RELIGIONS
Healing in Providence
Arcana of Spiritualism
Physical Manifestations
On the Threshold of a New Cycle
Spiritual Phenomena
The Need of Manifestations
Spiritualism and Science: 4%
The Electric World (New York) in Its Issue of August 31st Gives a Full Report of The
The New York College of Magnetics
Electrocution Not a Success
THE SCIENTIFIC AMERICAN ON SPIRITUALISM
"THE HEALING CONFUSION"
As Science and as Religion
A Searching Analysis of Spiritualism
Spirit and Body
Ideas regarding Spirit Origin and Education Versus Successive Embodiment
Spirit-Form Materialization
THE MYSTERIES OF IRON
The Spiritual Rostrum
Spiritualism and Science: 27%
THE LAST GRAND GENERALIZATION OF SCIENCE, THE LAW OF CONTINUITY
THE PSYCHIC FORCE AGAIN
The Scientific American and Spiritualism
SPIRITUALISM FROM A MATERIALISTIC STANDPOINT
Solid Electricity
THE PHYSICS AND PHYSIOLOGY OF SPIRITUALISM
HOW I CAME TO STUDY SPIRITUAL PHENOMENA
A Tale of Cosmic Multiplication
Science and Materialization
MAGNETISM
An Old-Time "Electrocution"
Are They Spirits?
FURTHER EXPERIMENTS
Our Time Seen from the Future
Early Scientific Explanations
THE DUAL UNITY OF SOUL, AND SPIRITUAL PARENTAGE
Electrical Wonders
AN OPEN POLAR SEA AND AURORAL LIGHTS
Earth Sea Sun: 23%
When Street Cars Are Run by Electricity, Probably Lightning Rods Will Be Employed as Conductors
New Electric Motor
In Materialization Spirit Does Not Become Matter; It Only Causes Material Elements
Prophetic: Electricity
Magnetism and Electricity
Execution by Electricity
THE ELECTRIC EEL
0% 5% 10% 15% 20% 25% 30% 35% 40% 45% 50% 55% 60% 65% 70% 75% 80% 85% 90% 95% 100%
Proportion

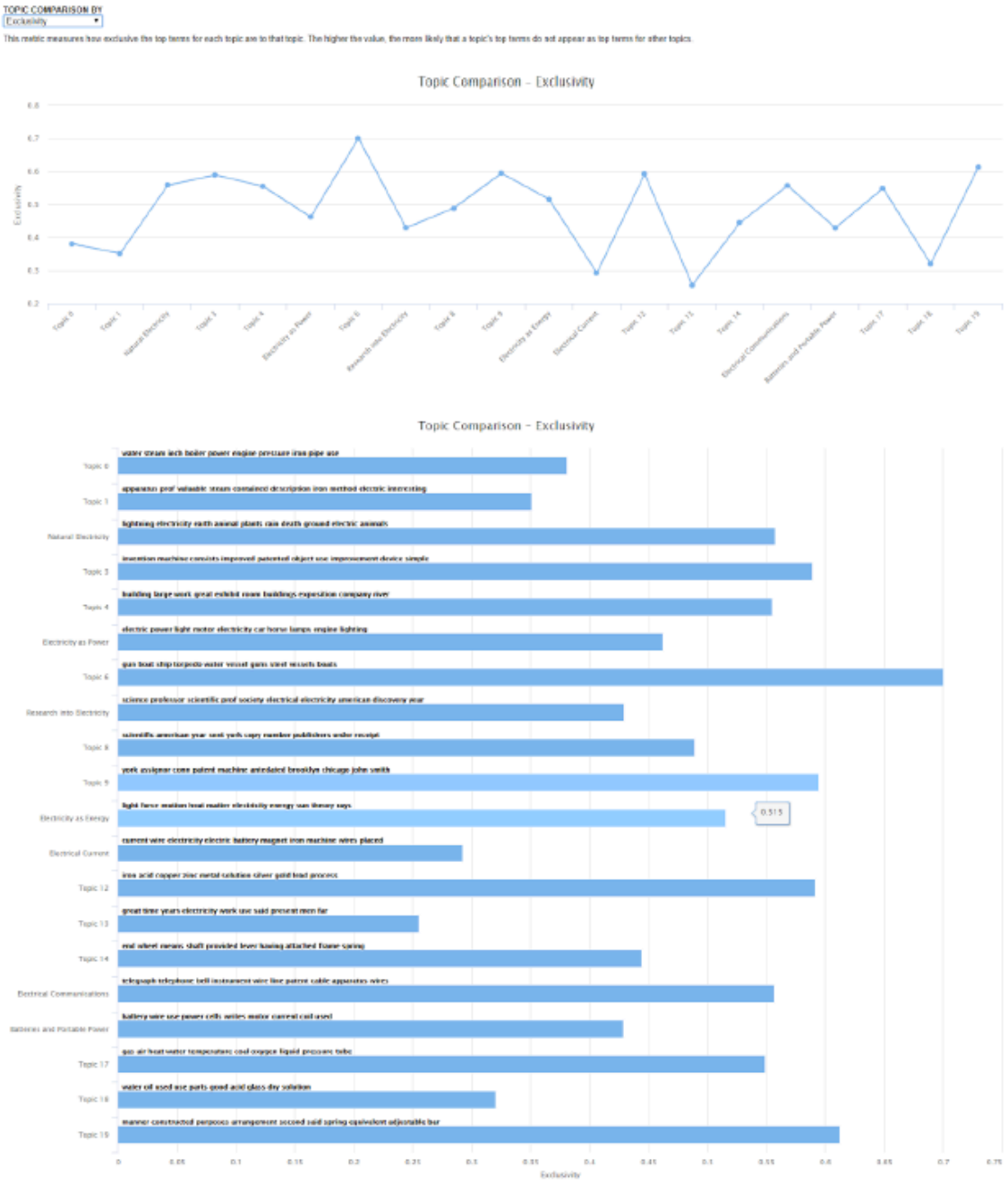


Scientific American

With this overview of Topic Modeling, we'll just provide some highlights here from the Scientific American results to compare against the Banner of Light results. There aren't any obvious overlaps.



The uniform distance measure does indicate that the most specific topics with electricity are ‘Electrical Communications’, ‘Batteries and Portable Power’, ‘Electrical Current’, and ‘Electricity as Power’.

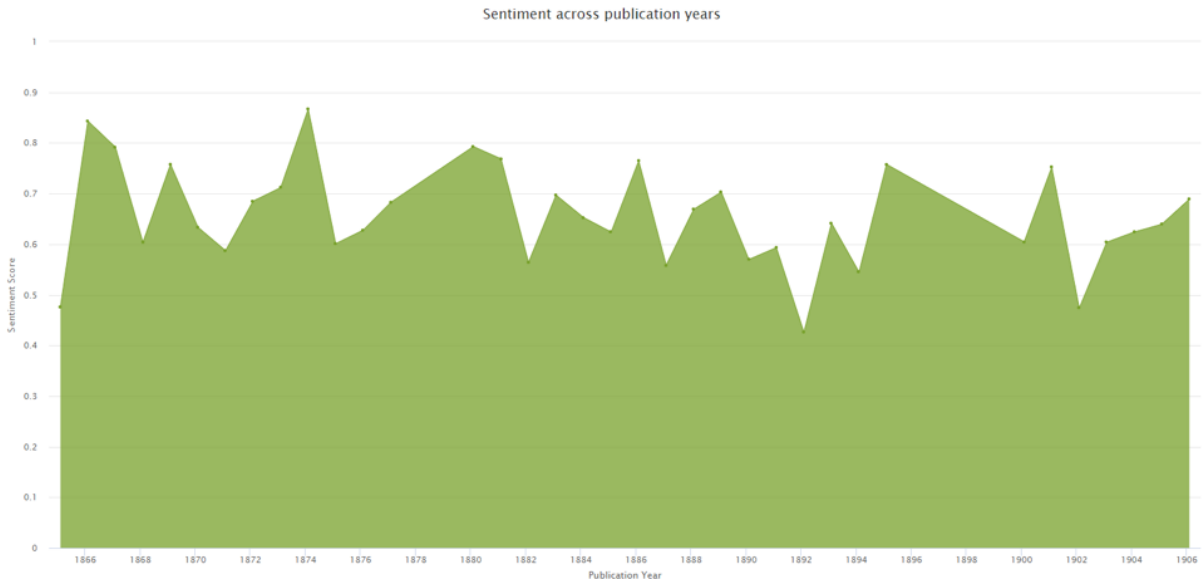


Exclusivity seems somewhat inconclusive as well. In the end the measures for these topics indicate a wide array of topics, and though there are some clear exclusive topics such as Topic 6 - involving guns and boats, there's also a healthy mix of topics in this content set.

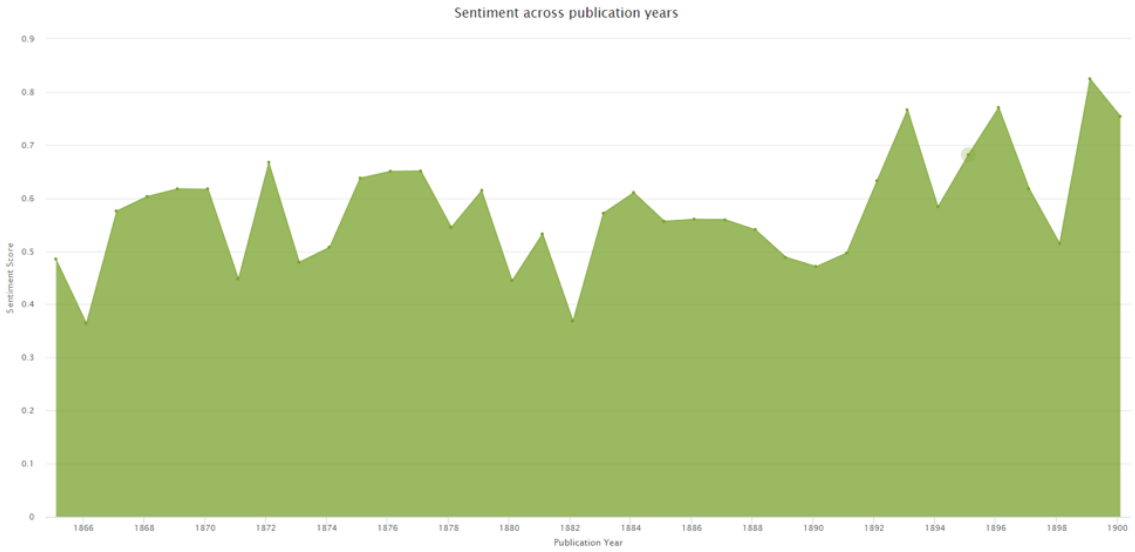
Sentiment Analysis

Perhaps unsurprisingly, the results for sentiment analysis for both content sets is consistently positive for both periodicals. This could suggest that electricity was an overwhelmingly positive topic, but we must remember that this tool provides a measure across the entire document - and that in serial publications with highly varied and mixed content, there's no clear means of ascertaining if such positive sentiment is in fact due to electricity, or perhaps rather, something else. They could also be a matter of the topics of both journals - spiritual well-being and growth, and scientific advancement and progress. Both of these broad areas of interest tend to focus on either personal or social / scientific / economic development, which tend (usually) to have more positive vocabularies than negative. The idea of betterment and progress, which was such a part of the culture of invention and discovery showcased by Scientific American, might explain why the highest values are seen in the 1880s and 1890s in the Scientific American content set. Consequently, though the results might look like something overwhelmingly positive, more precise content sets are likely required to drill down into these results.

Banner of Light



Scientific American

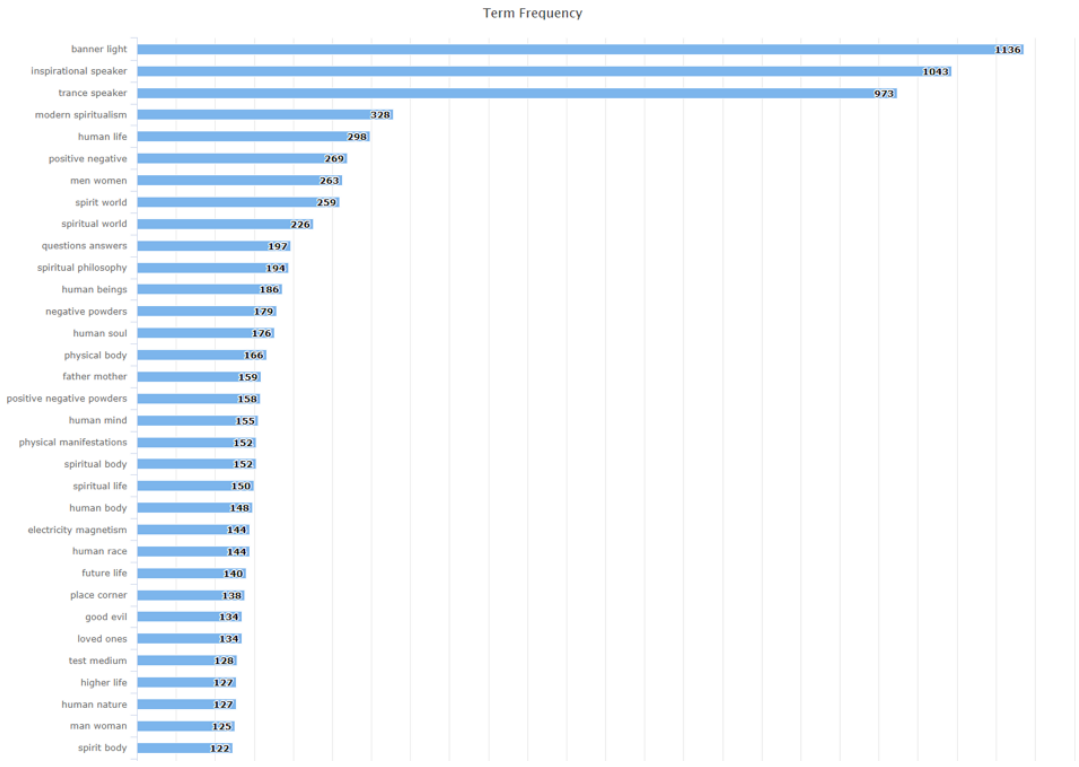


nGrams

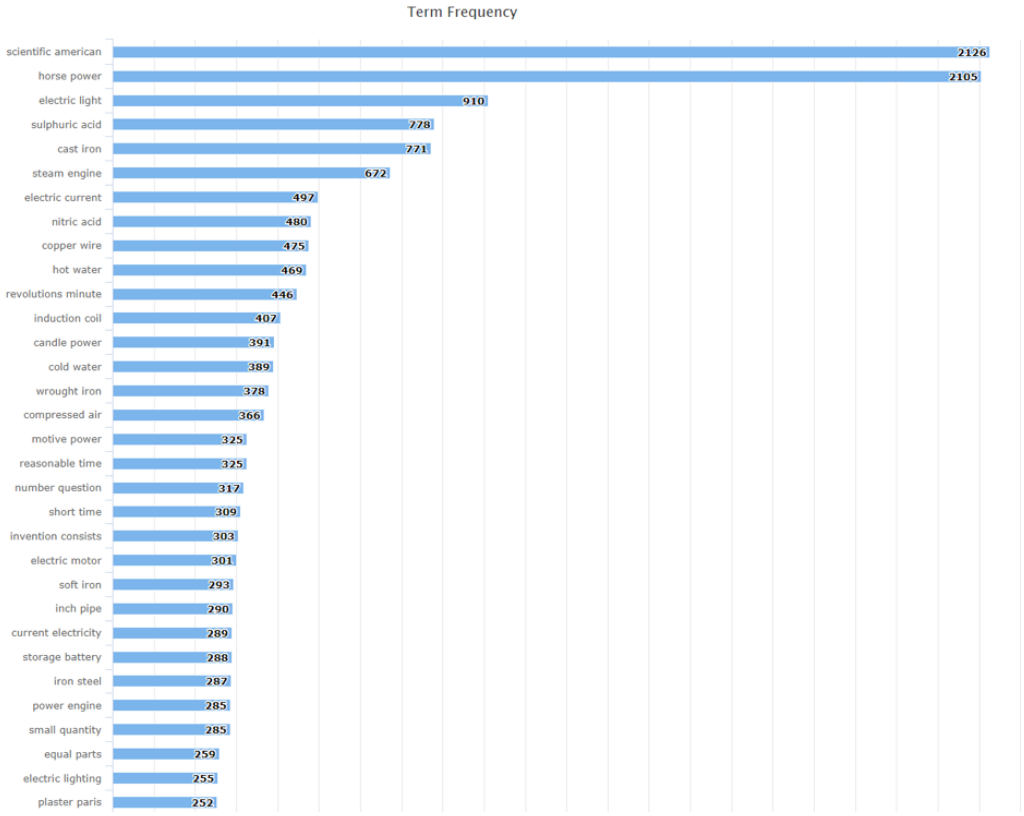
Configuration: Min 2 Max 5, Threshold 5

Cleaning Configuration: Electricity - Banner or Scientific American, no Punctuation, No Numbers, No Special Characters

Banner of Light



Scientific American



The most interesting thing about the nGram results is not merely that electricity isn't very prevalent, but that in the Banner of Light results, the word which appears with electricity is magnetism, not something more apparently to do with spirituality as we might expect. Equally, in Scientific American, the word light appears before current - 'light', as you can see above, is an important word in the Banner of Light topics. But it's also one of the most important scientific inventions involving electricity - the light bulb. With the Banner of Light nGram results, we find some hints, finally of some overlap between the two periodicals, and a new research question: what's the link between electricity and magnetism across the two publications?

Research Outcomes

The results are inconclusive, but we do find some commonalities. The overwhelming positivity of the Sentiment Analysis results, and the low appearance of electricity in the nGram results seem to indicate that more refinement of the Content sets are needed.

Original Questions

1. What themes are evident in the Banner of Light and Scientific American when discussing the topic of ‘electricity’?

    A. It’s clear that the two publications have fairly distinct interests. Where the Banner of Light mentions or treats topics related to electricity, it has to do with the body, as well as spiritual ‘force’. In contrast Scientific American is very much concerned with practical application of discoveries to new inventions.

2. Do the Banner of Light and Scientific American share any topics or similar points of view in their treatment of electricity in the late 19th century?

    A. At first glance, it doesn’t appear so. However, familiarity with the period, and the history of science and religion, there are some possible similarities. This is where results of text analysis really require deeper knowledge of a particular field of research. The idea of spiritual ‘force’, and the nGram result in Banner of Light for ‘electricity magnetism’, relate directly to the connection between current and electricity as a force that was also a way of describing the soul and the spirit. We can see this in the topic found in the Banner of Light results:

    - force,matter,form,motion,light,forms,atoms,heat,substance,forces
    - electricity,electric,life,brain,blood,current,water,body,electrical,air

3. What sentiments appear in discussions of electricity?

    A. They’re overwhelmingly positive - as we discussed above, this requires some reflection as there might be other factors at play here.

4. Are there any other distinguishing features surrounding electricity in these journals that may reflect on contemporary views of industrialization, invention, and their effects on men and women in the late 19th and early 20th centuries?

    A. Not that we can see at the moment.

New Questions

Question 2 above presents a possible new line of inquiry - Electricity and Spirit, perhaps. Or health, or body, or force. This will require more precise Content Sets to explore.

Perhaps the most fascinating discovery was a piece in the Banner of Light discussing Scientific American - this document could act as the foundation for an entire study on how the two periodicals reflect overlapping and maybe competing concerns of the era. Electricity clearly appears in the discussion - as the document is part of our content set.

Reflections on Method

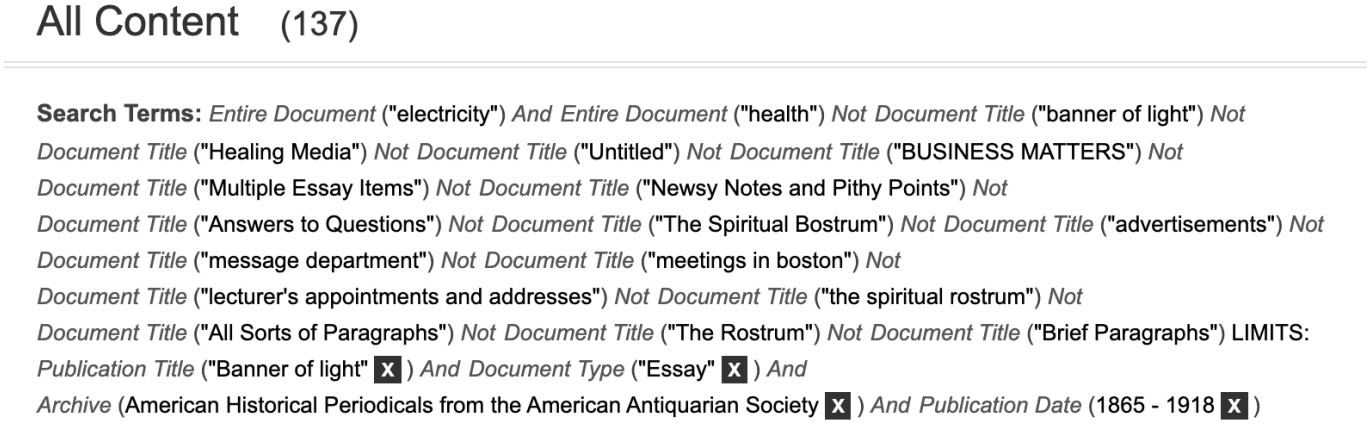
Content Set Building

The content set building for this comparative project consisted of finding appropriate temporal boundaries for two serial publications - we opted for 1865 as this is a significant year in American history - the end of the Civil War. 1918, as the end of the First World War or World War I, is also an important cultural moment. In between these two conflicts, the place of electricity within Americans society moved from a fairly limited notion through industrial and scientific development, into practical applications. At the same time, what it meant, and what it was, was a topic of intense cultural fascination and discussion. We can see both of these concerns in the serial publications selected for this project. Building the content sets, as a result, only varied in the selection of the periodicals themselves. Everything else remained the same - the temporal boundaries, and the word used: ‘electricity’.

Iteration

As easy as it was to build the initial content sets, each required creation of distinct cleaning configurations as the nGram and Topic Modeling tools revealed new words that obscured meaningful results. Both content sets need their own stop word lists, in order to remove the ‘noise’ - words arising from serial publications and advertisements, as well as unuseful information, like placenames for Banner of Light, and scientific experimentation words for Scientific American.

Both content sets, however, were also muddied by the presence of documents that often appear in serial publications - advertisements, notes and queries, letters from readers, set or repeating editorial sections etc. The easiest method of scrolling through the titles of the documents, to find repeating titles (and thus standard sections of the publications), was to browse the documents in Topic Proportions, and make a list. Then the original search parameters were revised (see Search History), and the titles added as rows to the advanced fields with the ‘not’ selection. Here’s an example search for Banner of Light with these repeated document titles excluded.



The new content sets, and their results, are linked below. It's worth comparing them against the original sets - can you find any further clarification to the original research questions?

Banner of Light - Essays and Articles (with titles excluded)

https://go.gale.com/ps/textAnalysisTools?method=updateTools&userGroupName=gdc_all&prodId=DSLAB&authType=Google&contentSetName=1581035076845

Scientific American - Essays and Articles (with titles excluded)

https://go.gale.com/ps/textAnalysisTools?method=updateTools&userGroupName=gdc_all&prodId=DSLAB&authType=Google&contentSetName=1581036274641

Understanding Outcomes

Revising Questions

As discussed in the guide, it's not only normal to revise your research questions after running analysis tools on a Content Set, it's an integral part of the research process. Often, analysis will turn up new questions which could lay beyond the scope of your current project. This is how researchers develop new projects and lines of scholarship - by following clues and new questions that come up while pursuing other research.

Limitations

- This project did not build content sets using actual cases, nor were they built following a close reading of the documents included in the sets. A more precise content set could be built by determining, following examination of each document, whether or not it was appropriate to include in a Content Set focused on the specific parameters of the project.
- Iteration is not restricted to cleaning - it's a key part of content set building as well. It became clear in this project that the recurring sections of serial publications or periodicals can add considerable noise to a content set. Excluding documents with titles that repeat can substantially change outcomes.

Beyond the Lab

Presentations

All of the tool outputs can be downloaded as images to use in Powerpoints, or embedded in webpages or other ways to present your work.

New Visualizations

It's also possible to download the data which power the visualizations as comma delimited (CSV) or javascript object notation (JSON) files, allowing you to create and format your own visualizations. If you have the skills, it's possible to collate or create new visualizations that may combine outputs from similar visualizations into one, allowing you to compare and contrast in new ways that the DSL tool does not. The Topic Modeling tool downloads are especially rich with possibilities for new visualization. The Topic view download is large and contains results for each document and measure for the Tool - much more data than the Topic Model visualizations can currently display. If you're a programmer, this is the ideal place to start to explore the data created by the DSL using other tools and visualization designs.

The development of electricity, and the comparative nature of this project, could be greatly extended beyond the lab by plotting downloadable data along timelines. The first might be breaking out the scores for topics by publication date, and plotting them along the time scale, as in the Topic Modeling Martha Ballard's Diary project (http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/).

And it could be done with a mind to moments in the development of electricity itself - using resources like the Electricity Timeline (http://resources.schoolscience.co.uk/britishenergy/14-16/index.html).

Refining the Content Sets

Understanding the limitations of the DSL allows us to consider what can be done to both to build content sets, and to use the results produced by its tools. As precise as the revised content sets might be, the many words surrounding electricity - not only permutations of it like electric, electrical - but those like current, force, spark, power, motion, spirit - suggest that perhaps using the Topic Modeling tool might offer a rich method of building more precise content sets when closer reading isn't possible.

Downloading the Content Sets

As powerful as the DSL's tools are, they offer fairly standardized configurations, and are not customizable at the moment. Downloading your content sets not only allows use of other tools, but also permits custom editing and cleaning of the documents. The DSL's cleaning configurations are not as powerful as more extensive, iterative techniques that make several passes over documents to refine and clear up problems arising from OCR digitization. Downloading also allows you, as a researcher, to find problem words and tokens that can complicate or mess up your results in the DSL. Download your content set and experiment, and use what you find to help refine your DSL projects.

Reading

Helle Porsdam, 'Digital Humanities: On Finding the Proper Balance between Qualitative and Quantitative Ways of Doing Research in the Humanities', Digital Humanities Quarterly 7.3 (2013)

Bernhard Rieder and Theo Röhle, 'Digital Methods: Five Challenges', in Berry, David M.(ed), *Understanding Digital Humanities* (2012).

Week 6 Discussion

This week's discussion post continues to build on your project research, reading and thoughts about this week's work on microhistories and qualitative/quantitative analysis.

1. Include a summary of reading you completed this week related to Monday and Wednesday's topics. Aim to discuss 2-3 works for each of the topics (Monday/Wednesday). You can discuss the Electricity sample project as one reading if you wish.

2. Provide a complete record of your work this week. Include:

- a discussion of what you have done this week.
- insights into the choices you have made as you clean your content sets, and your observations about what you feel has worked and what has not.

Replies: Due Tuesday of Week 7 11.59pm

WEEK 7a-b Text Processing & Ngrams

Processing Order of Text Cleaning Choices

Stage I:

- Select desired text segments from document(s).

Stage II:

- Apply basic latin filter, if enabled.
- Apply character drops, if any.

Stage III:

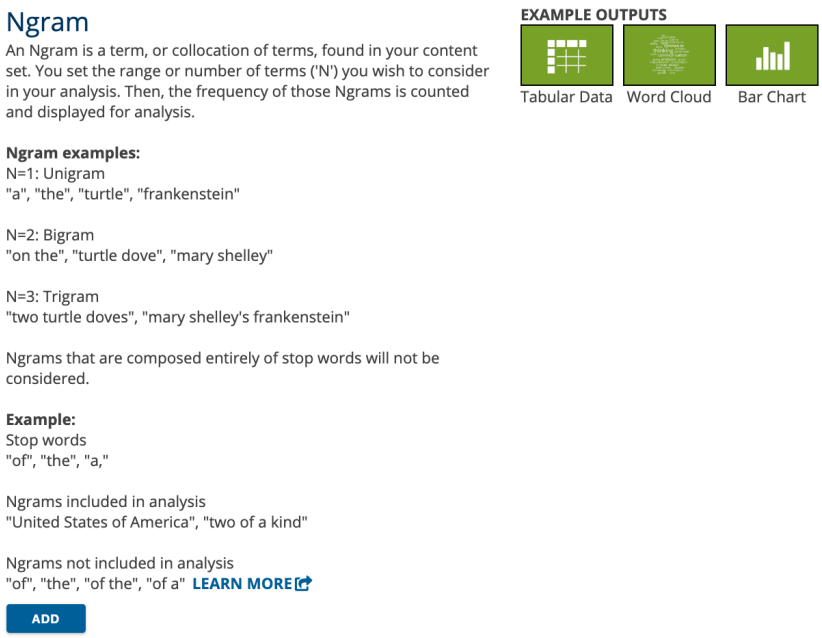
- Apply ReplaceWiths, if any. This is applied in the text sequence.

Stage IV:

- Apply stopword filter, if any.
- Apply multi-space squash if enabled.
- Apply lower-case, if enabled.

What comes out of Stage IV in the cleaning sequence is what is sent to any and all analysis jobs. None of the analysis tools have access to original document text. So, in the instance where we apply a clean configuration (with a stop word list containing "the" and "of") to an Ngrams job, is it correct to assume that "the" and "of" will be removed from "The United States of America"? Assuming "the" and "of" are included in the users stop words list and are filtered with correct (lower/upper) case, Then, yes, they will be removed. It may not be otherwise.

Ngram Tool Overview



Ngrams build collocations of words from tokens within Documents of a Content Set. A Ngram is nothing more than a sequence of words, where N represents the number of words. A unigram has 1 word, bigram, 2, trigram, 3, and so on. Ngrams are often used to compile search terms or words that are often associated with one another. This tool can be used to help determine if a Content Set contains specific terminology, or phrase, which otherwise can be difficult to trace without intensive reading of each Document.

Digital Scholar Lab Implementation

The DSL leverages the Apache Lucene tokenizer to identify strings of words based on the configuration options you choose.

- The Ngrams Tool tokenizes on whitespaces following a basic analysis.
- Importantly, stop word lists are applied after the composition of Ngrams.
- Only Ngrams where all tokens appear in the stop word list is removed from the results: Ngrams containing one token, not in the stop word list are retained and output.
- The Tool permits users to prioritize occurrence by setting a minimum threshold for the number of times an Ngram must appear in order to be added to the resulting list.
- The Tool is case sensitive; e.g. 'Apple' and 'apple' are treated as distinct tokens for the purposes of Ngram analysis.
- There are two distinct types of visualizations associated with this Tool:
    - Word Cloud which employs Highcharts.com's Word Cloud visualization.
    - Bar Chart which employs Highcharts.com's Bar Chart visualization.

Readings

Jean-Baptiste Michel et al, " Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331, 2011

Douglas Duhaime, 'Textual reuse in the Eighteenth Century: Mining Eliza Haywood's Quotations." *DHQ*, 10:1, 2016 http://digitalhumanities.org:8081/dhq/vol/10/1/000229/000229.html

Maarten van den Bos, Hermione Giffard, "Mining Public Discourse for Emerging Dutch Nationalism." *DHQ*, 10:3, 2016 http://digitalhumanities.org:8081/dhq/vol/10/3/000263/000263.html

Frederick W. Gibbs and Daniel J. Cohen. "A Conversation with Data: Prospecting Victorian Words and Ideas." *Victorian Studies*, vol. 54 no. 1, 2011, p. 69-77. *Project MUSE* muse.jhu.edu/article/468193. (I downloaded the article and it's here)

Claude S. Fischer, "Digital Humanities, Big Data, and Ngrams." *Boston Review* June 20, 2013. https://bostonreview.net/blog/digital-humanities-big-data-and-ngrams

M. Egnal, "Evolution of the Novel in the United States: The Statistical Evidence." *Social Science History*, 37(2), 2013, p.231-254. doi:10.1017/S0145553200010646 (available here).

Example projects using ngrams

Google Ngram Viewer: https://books.google.com/ngrams

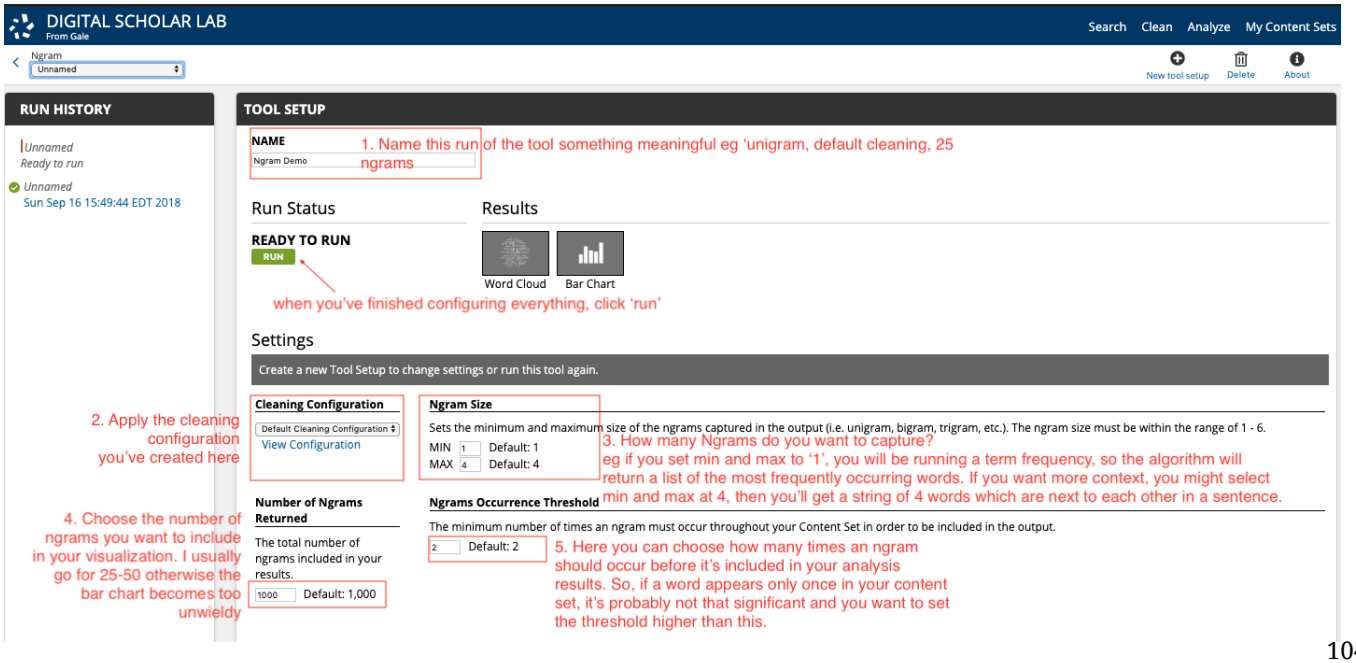
Voyant tools https://voyant-tools.org

Ben Schmidt, 'Poor man's sentiment analysis', *Sapping Attention*, Feb. 2 2012 https://sappingattention.blogspot.com/2012/02/poor-mans-sentiment-analysis.html

And projects described in the readings, above.

Configuration options in the DSL

Here are the options, also called out on the image, below.



1. Name this run of the tool something meaningful, eg 'unigram, default cleaning, 25 ngrams'

2. Apply cleaning configuration, if using.

3. How many ngrams do you want to capture? eg if you set min and max to '1', you will be running a term frequency, so the algorithm will return a list of the most frequently occurring words. If you want more context, you might select min and max at 4, then you'll get a string of 4 words which are next to each other in a sentence. The maximum n-gram size is 6.

4. Choose the number of ngrams you want to include in your visualization. I usually go for 25-50 otherwise the bar chart becomes too unwieldy.

5. Here you can choose how many times an ngram should occur before it's included in your analysis results. So, if a word appears only once in your content set, it's probably not that significant and you want to set the threshold higher than this.

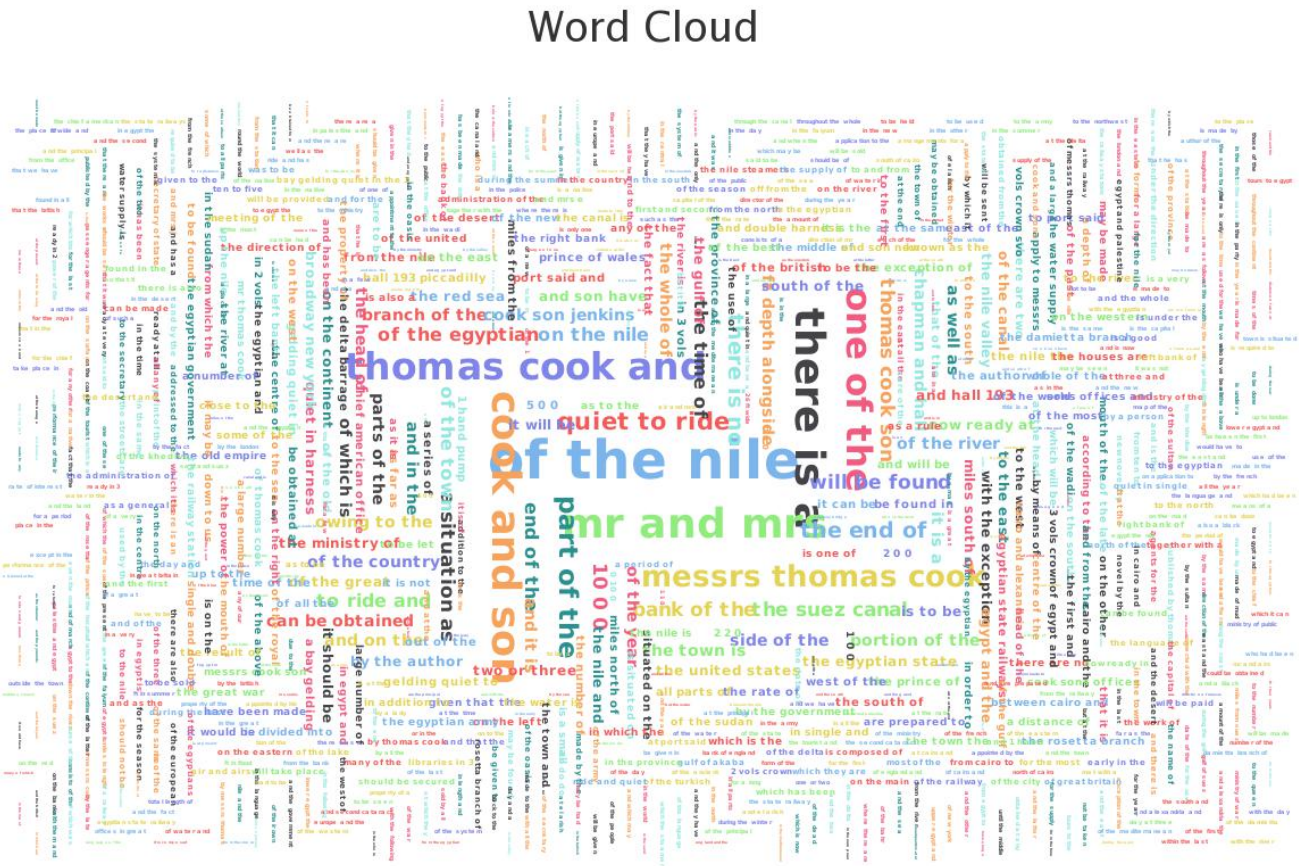
You have the option to view the results as either a bar chart or a word cloud. Both visualizations are downloadable.

- Word Cloud employs a visualization that presents Ngrams in a cluster, with those having the highest frequency or count in the center and those with lower counts towards the periphery of the visualization.
- Ngrams with higher counts are displayed with larger fonts, quickly drawing the viewer's attention to the most prominent feature as the most statistically relevant element of the visualization.
- The Bar Chart visualization represents the results as a standard Bar Chart, providing the viewer with counts of the occurrences of specific Ngrams in a Content Set.

Note: if you choose to look at collocates, a Word Cloud may not necessarily be the best visualization for these results! As you can see, it becomes almost unintelligible:

In this case, a downloaded CSV or bar graph will be more valuable. This underscores the importance of choosing the right sort of visualization for your results. For more commentary on Word Clouds, see

Jacob Harris, 'Word Clouds considered harmful', *Nieman Journalism Lab*, October 13 2011.

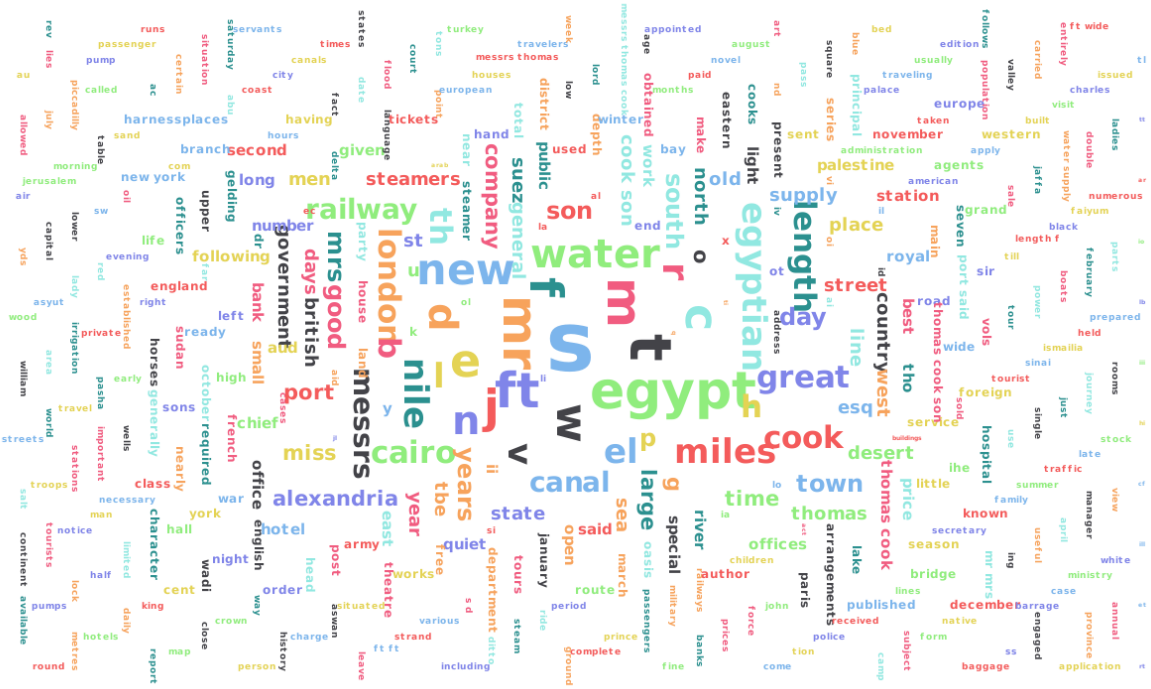


His use of the label ‘mullets of the internet’ to describe word cloud visualizations is more than enough reason to read the article in its entirety.

Using Ngrams as a tool for cleaning

You can include a simple term frequency analysis as part of your text cleaning workflow. Running the ngram tool with max and min thresholds set to ‘1’, you will return a list of the most common words in your content set. Download the output as a CSV and open the file. You can cast your eye down the list, and copy any words that you don’t want included in the final analysis. These may include ‘nonsense words’ created through poor OCR, word fragments, or words that you are simply not interested in analyzing for whatever reason.

Word Cloud



ngram	count
s	742
t	516
egypt	463
mr	438
ft	423
m	420
e	411
f	407
j	403
c	402
new	375
w	373
water	350
d	349
n	339
r	326
l	293
nile	289
miles	288
egyptian	284
el	273
v	271
cairo	261
london	259
great	257
b	246
h	245
length	242
cook	241

Once you have copied your selected words, you can paste them into the stop words list on the ‘Clean’ page. Should you identify words which are frequently mis-recognized by the OCR engine, but you want to keep the correct spelling in your analyses, you can paste the misspelling into the ‘replace this’ box, and the correct spelling into the ‘with this’ box.

DH Researcher Interviews

Dr Wendy Perla-Kurtz ‘Mediating Memory: Text Mining a Dictatorship’ (MP4 video file)

Ruth Trego ‘Topic Modeling and Tropicality in American Fiction” (MP3 audio file)

Week 7 Discussion

This week's discussion post will describe the work you have done as we begin to move forward from collecting, curating and cleaning your content sets into analysis and visualization. As we've discussed, this is an iterative process rather than a straight line moving only forwards. The Ngram tool in particular is useful for identifying recurrent word fragments and other textual oddities, so it bridges the gap between 'clean' and 'analyze'. When used to identify collocates (eg trigrams etc), it can give contextual information about prevalent words and their position in a document.

Please note that the DSL has limited flexibility within the Ngram tool should you wish to look at a particular word in context in a sentence. You are able to do this in Voyant, however.

1. Include a summary of reading you completed this week related to Ngrams. Aim to discuss 2-3 articles or projects.

2. Provide a complete record of your work this week, including processes, triumphs, research insights and questions.

3. Consider if you need to revise your Project One Pager to reflect the current status of your work.

Replies are due by Tuesday of Week 8 11.59pm.

WEEK 8a Named Entity Recognition

Named Entity Recognition is often used to identify key people, places, and things within a Content Set. This tool can be useful when collecting data around place names for mapping, which can often be challenging to aggregate without the close reading of each document.

Digital Scholar Lab Implementation

The Named Entity Recognition tool is based on the open source spaCy model. When you run the tool, it will parse (work through) your content set, and identify words classified as 'entities' by the model, which as been trained using the OntoNotes 5 corpus. For your purposes, here is the list of entities which will be recognized, along with their abbreviation:

TYPE	DESCRIPTION
PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
EVENT	Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
LANGUAGE	Any named language.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including ”%“.
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	“first”, “second”, etc.
CARDINAL	Numerals that do not fall under another type.

Reading

Suvro Banerjee, 'Introduction to Named Entity Recognition', *Medium*, 2018 https://medium.com/explore-artificial-intelligence/introduction-to-named-entity-recognition-eda8c97c2db1

Kimmo Kettunen et al, 'Old Content and Modern Tools - Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771-1910', *DHQ* 11:3, 2017 http://digitalhumanities.org/dhq/vol/11/3/000333/000333.html

Rainer Simon, Leif Isaksen, Elton Barker, and Pau de Soto Canamares, The Pleiades Gazetteer and the Pelagios Project' in Ruth Mostern, Humphrey Southall, Lex Berman, & Peter Bol, Placing Names: Enriching and Integrating Gazetteers, 2016. Project MUSE., https://muse.jhu.edu/.

Example Projects using Named Entity Recognition

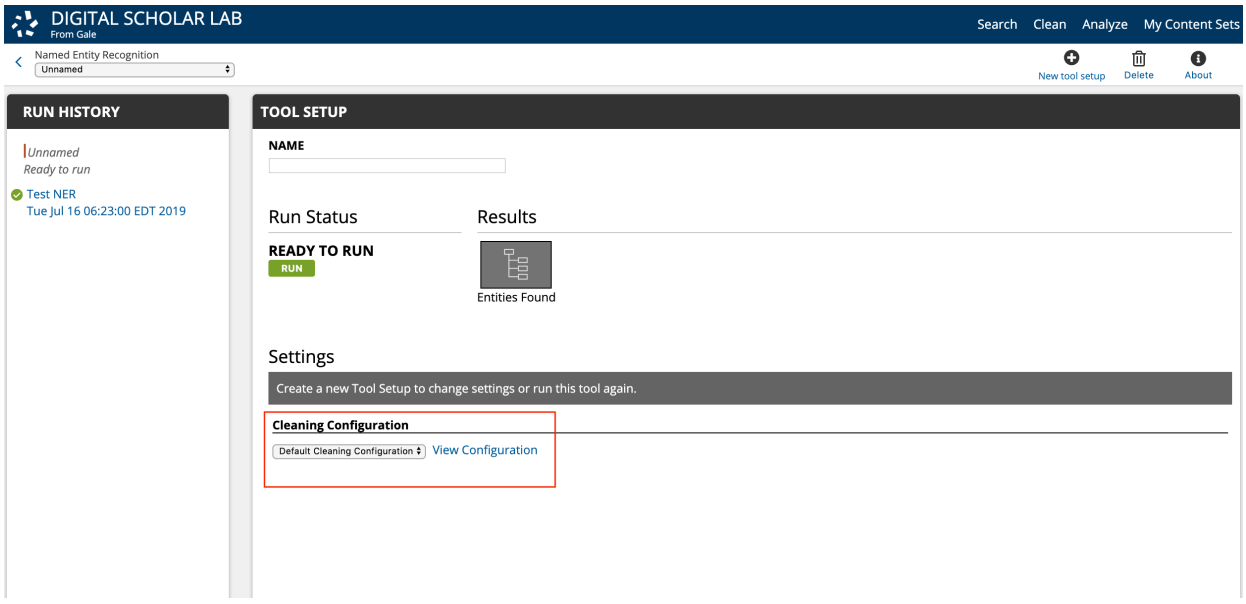
Gazetteers are one example of Named Entity Recognition in practice, in this case targeting place names in text. Often, this process is used as the starting point for building maps or historical narratives on the theme of 'place'.

Pelagios is one such project, which is well-funded and is very active. You can read more about it in the article, above. Check out the markup tool here: https://recogito.pelagios.org/

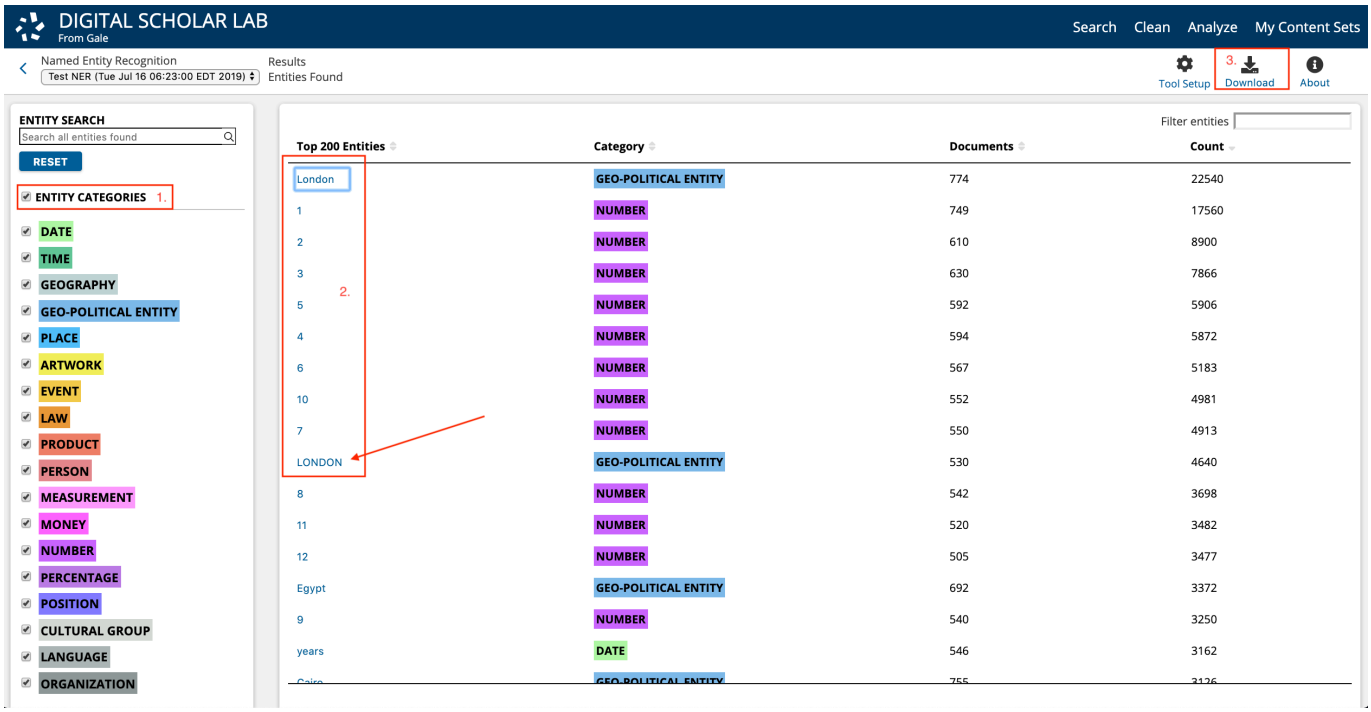
My own research is based on recognizing people and place names in Nile travel journals, letters and other ephemera from the late nineteenth century. My goal is to map social and travel networks in Egypt during this period which is known as the 'Golden Age of Egyptology'. This past year, I have been working with a Masters student in Computational Linguistics who has built a historical markup tool for our project, and others who want to use it. It essentially takes plain text (.txt) input, and then outputs text that has been encoded in a machine-readable format, in this case XML-TEI. The tool also identifies and marks up named entities in the text. You can try it out at our project website: http://www.emmabandrews.org/project/historical-markup-tool

Configuration Options in the Digital Scholar Lab

The only configuration option currently available for named entity recognition is the cleaning configuration you choose to apply to the content set.



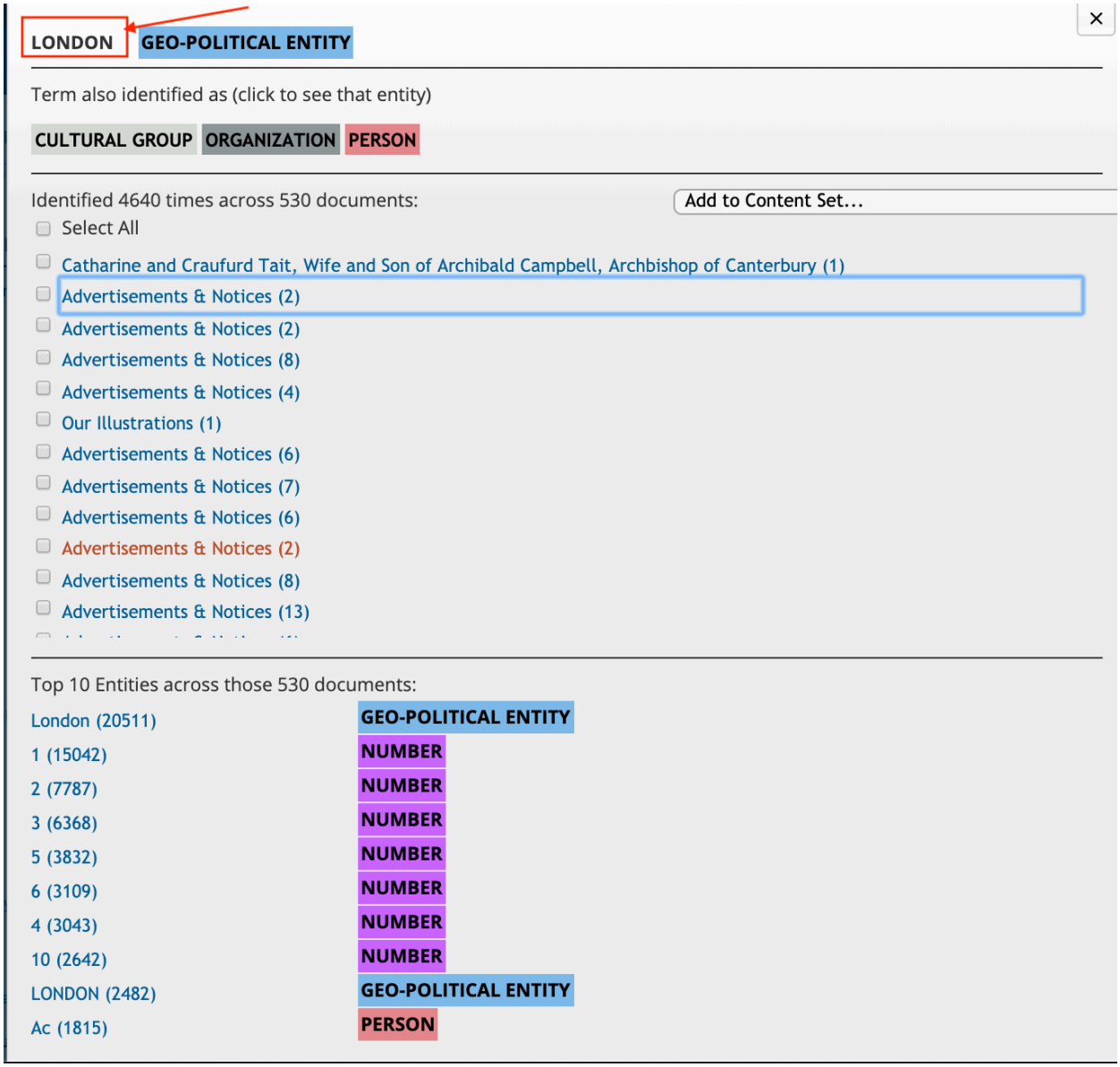
Output and Visualizations



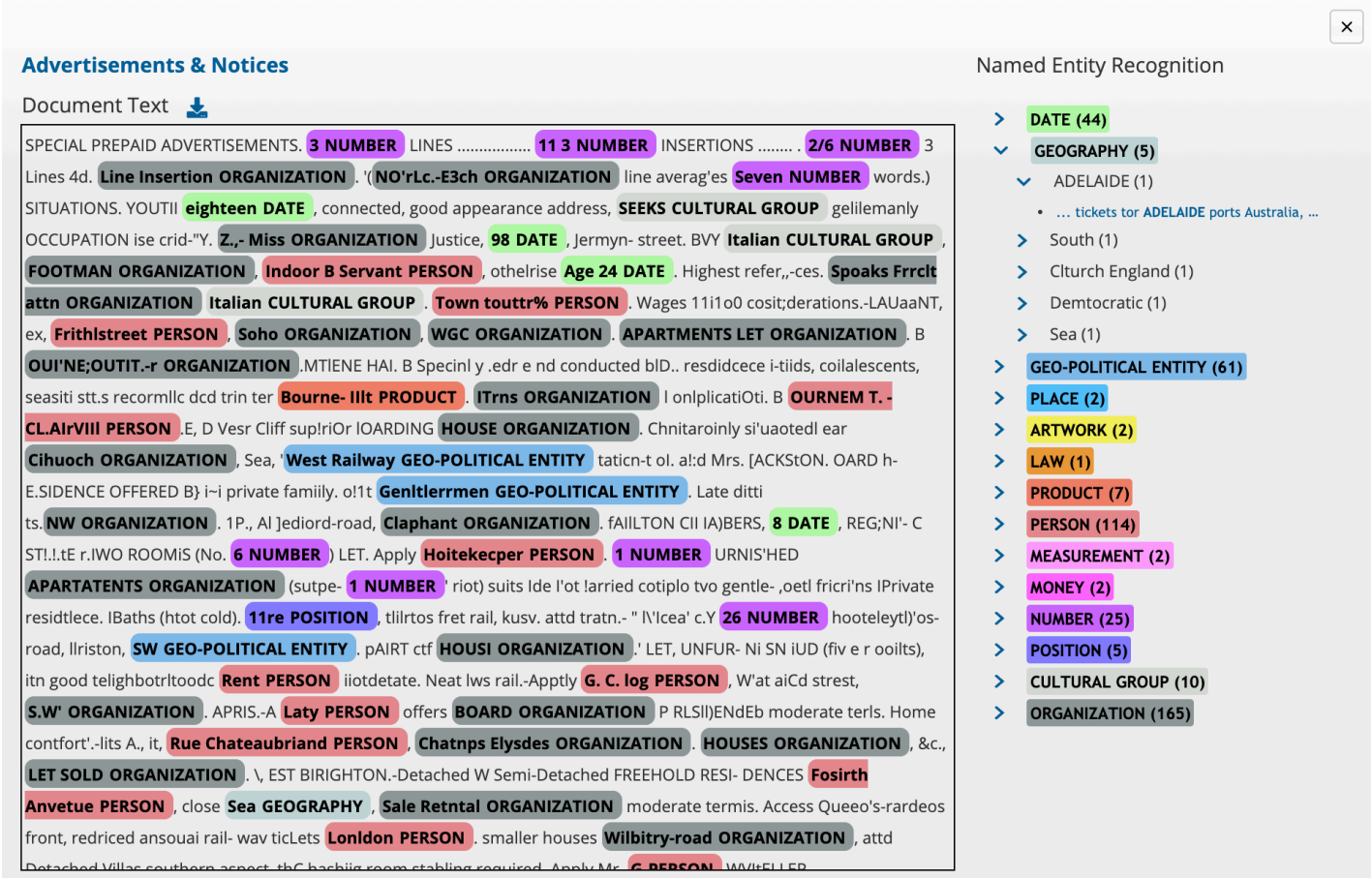
Each red box (above) is numbered:

1. In the left column, you'll see a list of Entity Categories identified by the tool in your content set, which are color coded. You can toggle each Entity Category on and off by checking or unchecking the boxes.

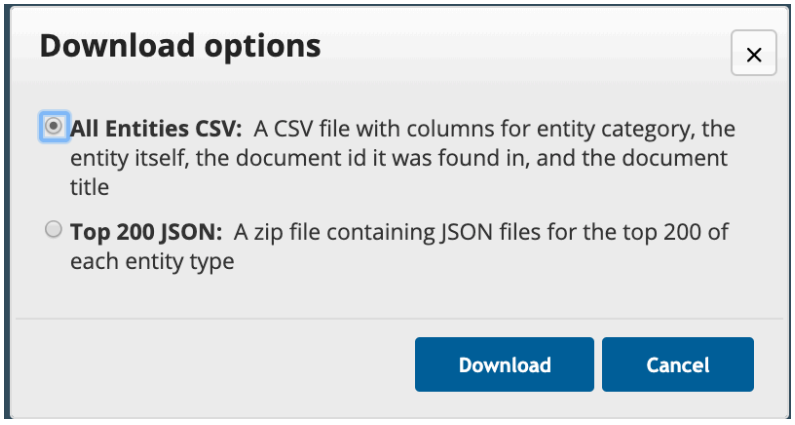
2. Individual entities are listed, along with the color-coded Entity Category, the number of documents the entity appears in, and the number of times it appears in total. The entities are clickable. Here, I've clicked into 'London':


LONDON
GEO-POLITICAL ENTITY
Term also identified as (click to see that entity)
CULTURAL GROUP
ORGANIZATION
PERSON
Identified 4640 times across 530 documents:
Add to Content Set...
Select All
Catharine and Craufurd Tait, Wife and Son of Archibald Campbell, Archbishop of Canterbury (1)
Advertisements & Notices (2)
Advertisements & Notices (2)
Advertisements & Notices (8)
Advertisements & Notices (4)
Our Illustrations (1)
Advertisements & Notices (6)
Advertisements & Notices (7)
Advertisements & Notices (6)
Advertisements & Notices (2)
Advertisements & Notices (8)
Advertisements & Notices (13)
Top 10 Entities across those 530 documents:
London (20511)
GEO-POLITICAL ENTITY
1 (15042)
NUMBER
2 (7787)
NUMBER
3 (6368)
NUMBER
5 (3832)
NUMBER
6 (3109)
NUMBER
4 (3043)
NUMBER
10 (2642)
NUMBER
LONDON (2482)
GEO-POLITICAL ENTITY
Ac (1815)
PERSON

You can then click into individual documents in the list to look closely at what has been captured by the tool.



3. You can download the full list in CSV or JSON format by clicking on the button.



Note: You will probably notice that some entities are misclassified by the tool. It's currently not possible to reclassify them in the DSL. The only way to do this is to download the CSV and then edit it.

Future updates of the NER tool will include the ability to edit and reclassified wrongly identified words.

Mapping

A natural companion to Named Entity Recognition is the mapping of place names, since they are readily extracted from text by running the analysis tool. Once the run is completed, a spreadsheet of Named Entities can be downloaded in the form of a CSV file, opened in Excel or OpenRefine and cleaned up to remove extraneous data. It's also possible to extract locations outside the DSL by exporting the OCR text and using the Clavin method, described below.

Once location data has been extracted, a number of tools exist for visualizing this data on a map. A few options are suggested below.

- Neatline - a plugin for Omeka. Tutorial here.
- StorymapJS - a free tool for creating narratives on a map base layer.
- Geolocator - activated in Omeka. Tutorial here
- Carto
- ArcGIS online
- ArcGIS Storymaps
- Edinburgh Geoparser

Geoparsing Text Data

Working with plain text files, an option to extract location names is to use the online version of the CLAVIN geoparser (http://clavin.berico.us clavin-web/) to extract location data from select pages of a dataset.

CLAVIN (Cartographic Location And Vicinity INdexer) is an open source software tool for document geo-tagging and geo-parsing. It automatically extracts location names from structured and unstructured text and resolves them against a gazetteer to produce data-rich geographic entities. It has 75% accuracy for geospatial entity resolution, can resolve 100 locations per second per CPU and process 1 million documents in under an hour on a 9-node Hadoop cluster. It can scale to billions of records.

Full books may take a little while to generate geoparsed data - be patient. Caveat: you can only extract the top 20 locations (although on testing this, I note that I was able to extract 60!).

The goal is to submit text to the online geoparser and structure that information into an online spreadsheet, so that the results can be visualized using a mapping application.

The process

Step 1: Get the data

- Download your content set from the DSL. You have the ability to clean within the platform, which I recommend doing.

- You can either concatenate your individual .txt files into a single file, or upload several files one-by-one. Python file for concatenation is available here.

Step 2: Open the geoparser and the shared spreadsheet in web browser tabs

- Navigate your browser to the CLAVIN Web interface (http://clavin.berico.us/clavin-web/) *Note that depending on screen size, you may need to 'zoom out' your browser (Use the toolbar or Ctrl+scroll down), or resize the screen in order to move the map from covering the data results list.

- On a second tab, open the geolocation spreadsheet. **MAKE A COPY** in your own Drive.

*Option 1 Repeat steps 3-4 for each individual text file

*Option 2 Run your concatenated file through the geoparser, although this may take a while and will only return the top 20 results.

Step 3: Use the geoparser to extract location data

- Open the text file with a text editor
- Select all text (Ctrl+a on PC/Linux, for example) and copy it (Ctrl + c)
- Paste (Ctrl+v) the text into the text box on the online CLAVIN geoparser and click the 'Resolve Locations' button

Step 4: Copy and paste the results to your copy of the shared spreadsheet

- Highlight and copy the locations resolved by the geoparser
  - Do not copy the column headers ("ID|Name|Lat,Lon|Country Code|#")
- Navigate to YOUR COPY of the shared spreadsheet above and paste your results into open rows
  - Ensure that your data lines up with the headings

Download the spreadsheet to a working location on your computer.

Useful Tools for Mapping

Mapwarper: Find maps and other imagery, upload, and rectify against a real map.

Reed College Geocoding Application: https://rich.shinyapps.io/geocoder/

David Rumsey Historical Map Collection: https://amica.davidrumsey.com/home

NYPL Open Source Maps: http://www.openculture.com/2014/03/new-york-public-library-puts-20000-hi-res-maps-online.html

Chris Gist, 'Projection Lessons in Maps', *Scholar's Lab Blog*, December 1, 2011 https://scholarslab.lib.virginia.edu/blog/projection-lessons-in-maps/

Living Maps Review: http://livingmaps.review/journal/index.php/LMR/index

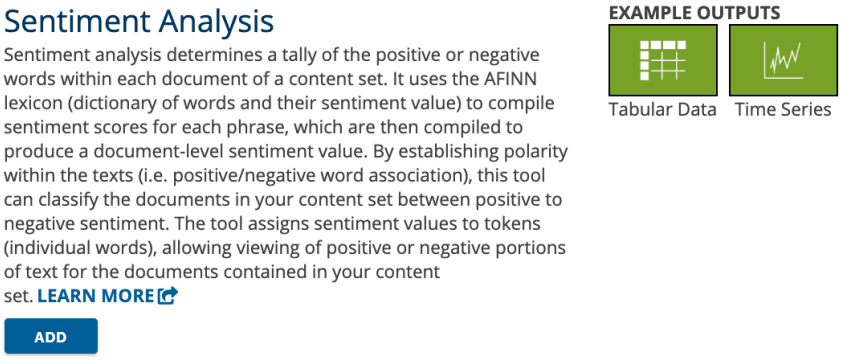
British Library Maps: https://www.bl.uk/maps/

Pelagios Network: https://pelagios.org/

The Pelagios Network is a long-running initiative that links information online through common references to places. To create and maintain these connections, Pelagios has developed:

- *a method for creating semantic annotations, based on the W3C Web Annotation standard;*
- *tools and specifications for creating and making use of these annotations, most notably Recogito, an open-source platform for geo-annotating texts, images and databases;*
- *a community of individuals and organizations working with geographic data in humanities disciplines (history, language and literary studies, archaeology, etc.), and cultural heritage (galleries, libraries, archives and museums).*

Sentiment Analysis Tool Overview



Digital Scholar Lab Implementation

The DSL measures sentiment across time, based on the timeframe covered by your content set. It measures positive and negative sentiment based on the AFINN lexicon of positive and negative

vocabulary. Future releases of the DSL will include the ability to leverage other sentiment lexicons, as well as including your own weightings.

Reading

- Jockers, Matthew L. “A Novel Method for Detecting Plot.” June 5, 2014. http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/

This short blog post by Matt Jockers is one in a series where he delves into specific examples of how to use text analysis and visualizations for literary research. By tracing sentiment in the 19th century novel, Jockers (accidentally) “discovered that the sentiment I was detecting and measuring in the fiction could be used as a highly accurate proxy for plot movement.” In the post he describes how he uses sentiment analysis for detecting plot in four disparate novels: James Joyce’s *The Portrait of the Artist as a Young Man*, *Picture of Dorian Gray* by Oscar Wilde, *The Da Vinci Code* by Dan Brown, and Cormac McCarthy’s *Blood Meridian*. In a follow-up post to the one linked above, he describes how he normalizes “the plot shapes in 40,000 novels in order to compare the shapes and discover what appear to be six archetypal plots!”

- https://programminghistorian.org/en/lessons/sentiment-analysis

- Finn Årup Nielsen, AFINN Sentiment Lexicon, 2011 http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010
- Parul Pandey, ‘Simplifying Sentiment Analysis using VADER in Python (on Social Media Text)’, https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f
- Sentiment Analysis with Python NLTK Text Classification (Try it out!)
- Acerbi A, Lampos V, Garnett P, Bentley RA, ‘The Expression of Emotions in 20th Century Books’. PLoS ONE 8(3): e59030, (2013) https://doi.org/10.1371/journal.pone.0059030

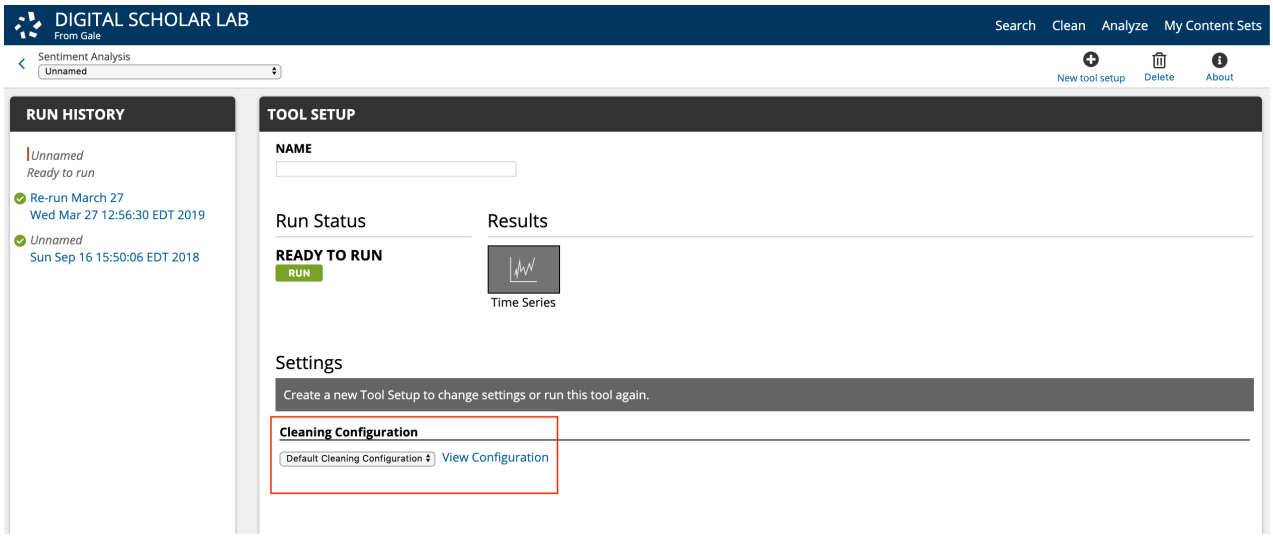
Example Projects Using Sentiment Analysis

Alexander Spangher, 'How does this article make you feel?', *New York Times* October 31, 2018 https://open.nytimes.com/how-does-this-article-make-you-feel-4684e5e9c47

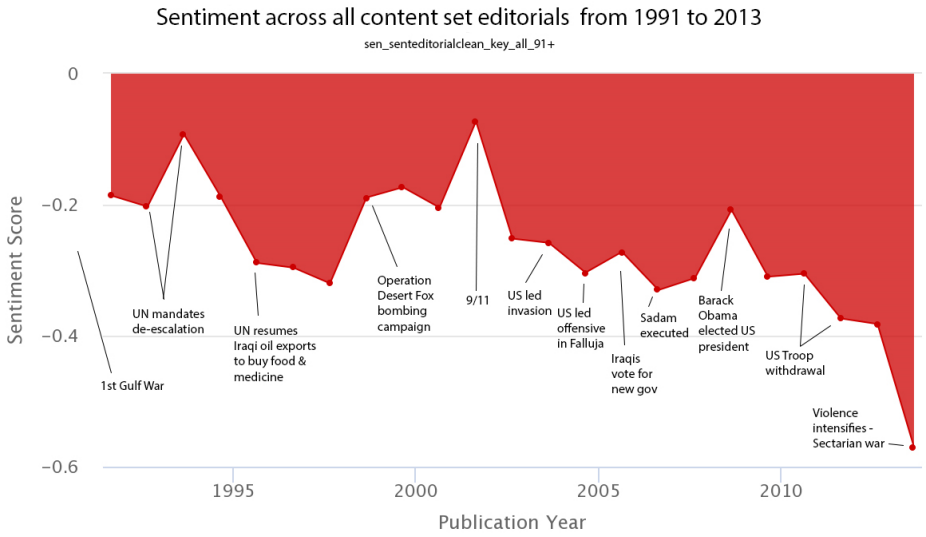
'Sentiment analysis is opinion turned into code', *Open Objects* 2015 http://www.openobjects.org.uk/2015/04/sentiment-analysis-coming-to-a-newspaper-near-you/

Configuration Options in the DSL

The only configuration option currently available for sentiment analysis is the cleaning configuration you choose to apply to the content set.



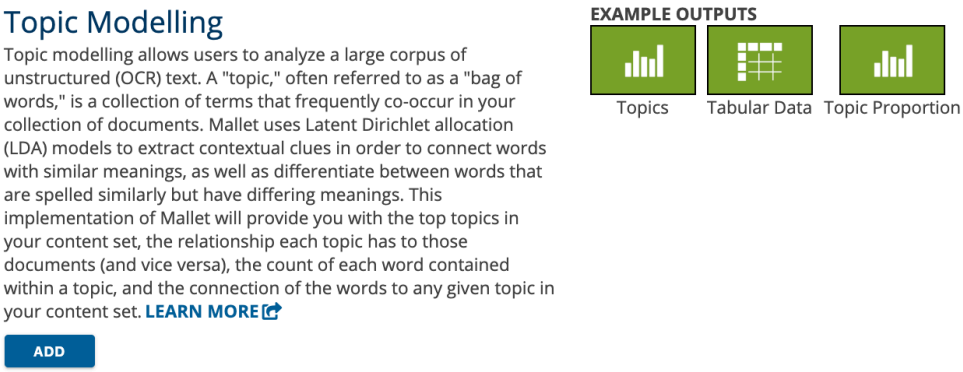
When you run your analysis and generate the visualization, you'll be able to click into each point to take a closer look at the document to determine why it was considered positive or negative.



Sometimes there are outliers which perhaps don't belong in your content set at all, and you can remove them at this stage of analysis.

WEEK 8b Topic Modeling & Clustering

Topic Modeling Tool Overview



When you run topic modeling, the algorithm will sift through your content set and group together themes or topics it considers related in some way. Posner (below) notes that: 'it should be abundantly clear that no part of this process is “scientific”; it’s just one way of getting your head around a large body of text. So there’s no right or wrong topic name, just schemas that do and don’t help you find interesting features of the text you’re looking at.’

Sample project conversation with Margaret Waligora, ‘Topic Modeling the Watergate Scandal’

Digital Scholar Lab Implementation

The topic modeling tool in the DSL is built on MALLET, which initially was accessible only via the command line but now also has a GUI implementation. It's fully integrated in the DSL making it pretty streamlined to use.

Reading

Megan R. Brett, Topic Modeling: A Basic Introduction, *JDH* 2:1, 2012 http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/

Ted Underwood, 'Topic modeling made just simple enough', *The Stone and the Shell*, 2012. https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Ted Underwood, 'Visualizing Topic Models' *The Stone and the Shell*, 2012. https://tedunderwood.com/2012/11/11/visualizing-topic-models/

Shawn Graham, Scott Weingart, and Ian Milligan, Getting Started with Topic Modeling and MALLET https://programminghistorian.org/en/lessons/topic-modeling-and-mallet

Miriam Posner, 'Very Basic Strategies for Interpreting Results from the Topic Modeling Tool', http://miriamposner.com/blog/very-basic-strategies-for-interpreting-results-from-the-topic-modeling-tool/

Example Projects using Topic Modeling

Mining the Dispatch, https://dsl.richmond.edu/dispatch/pages/home

Cameron Blevins, 'Topic Modeling Martha Ballard's Diary', 2010 http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/

Quintus Van Galen & Bob Nicholson, 'In Search of America', *Digital Journalism*, 2018. DOI: 10.1080/21670811.2018.1512879

Configuration Options in the DSL

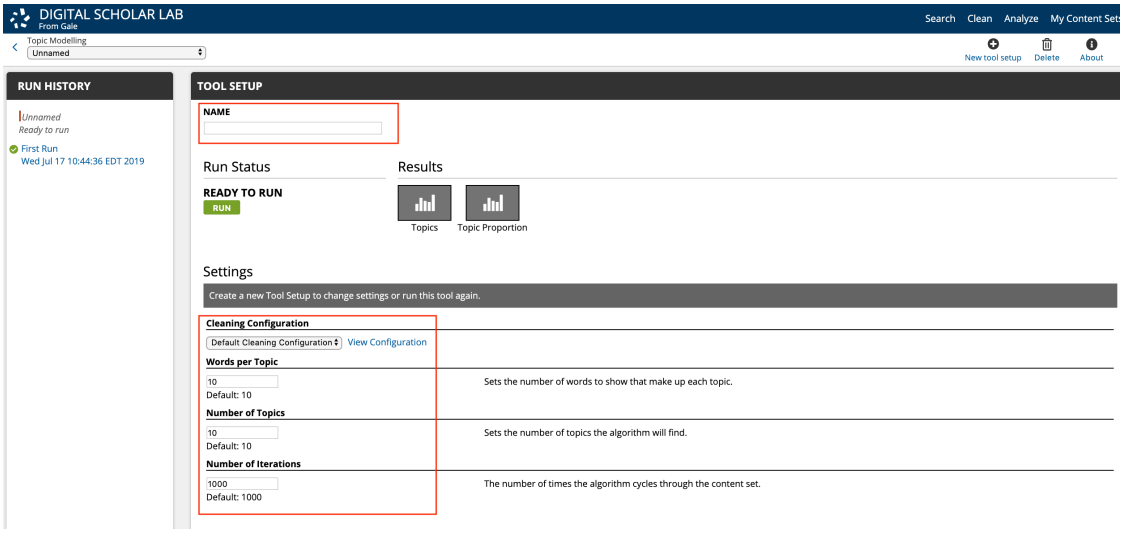
1. Name the run of your tool.

2. Apply cleaning configuration, as appropriate. Things to note: MALLET - the software powering the Topic Modeling Tool - is case sensitive. If you decide to make everything lower case, it won't distinguish between Smith (perhaps someone's last name), and smith (an occupation, like a blacksmith). MALLET also handles possessive apostrophes in a slightly awkward manner, turning them into their own words - you can add 's to the Stop Word list to prevent this from happening.

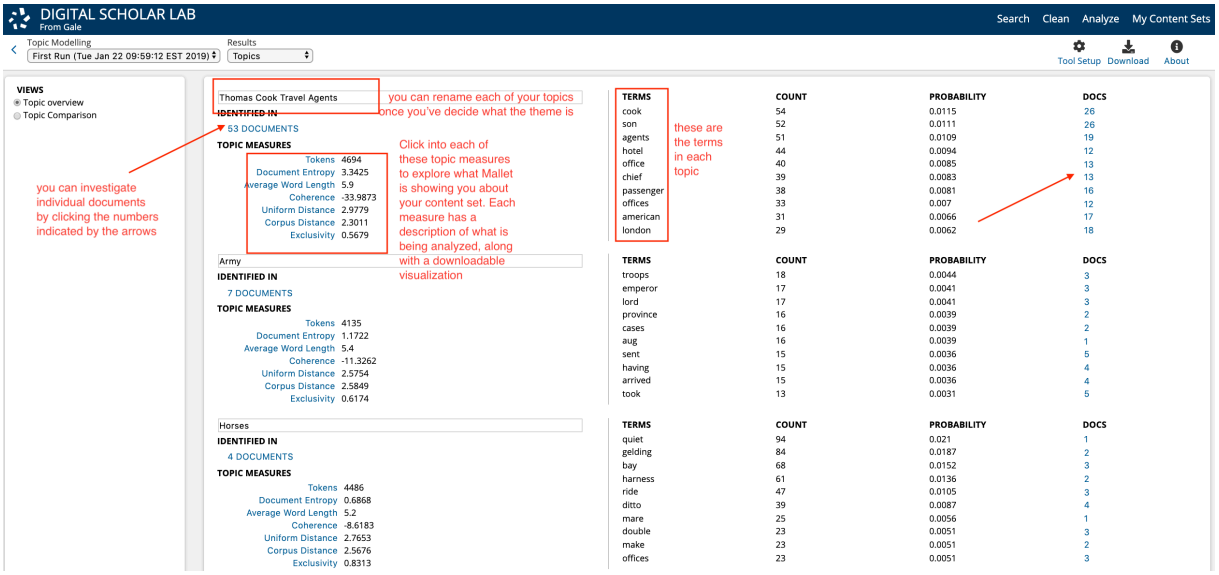
3. Choose the number of words you'd like to appear in each topic. 10 is a reasonable default.

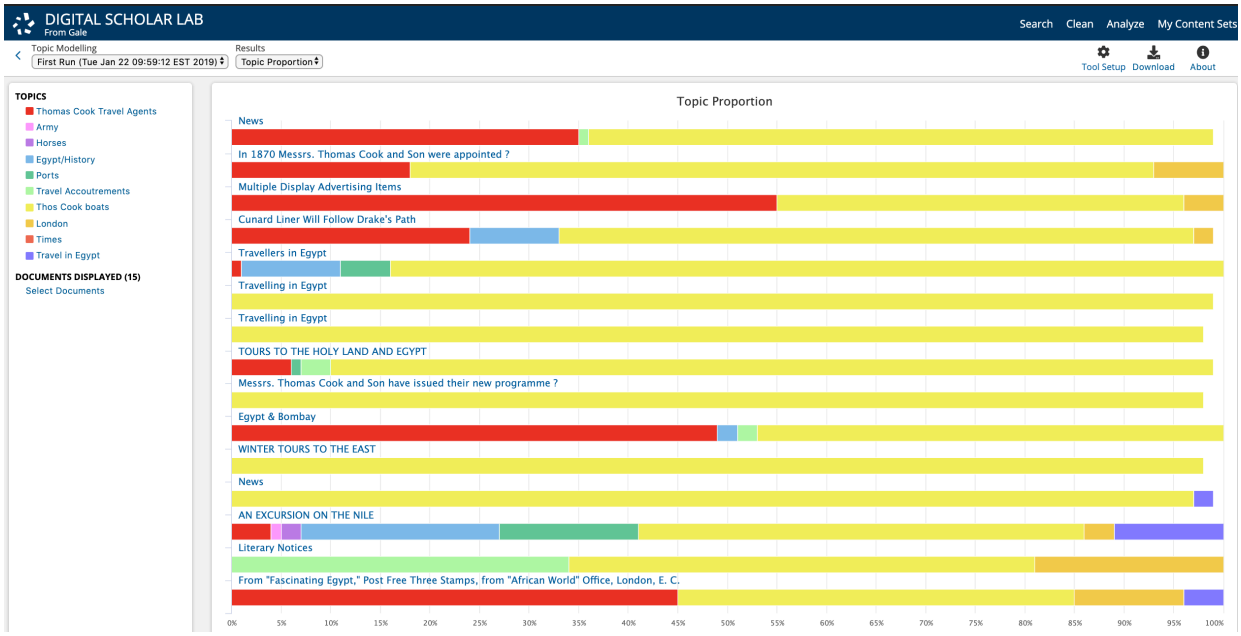
4. Choose the number of topics - you could start with 10, then increase the number to see what else you can discover.

5. The number of times the algorithm will iterate through the content set before returning a result is set at 1000 as a default. It's fine to leave this as it is.



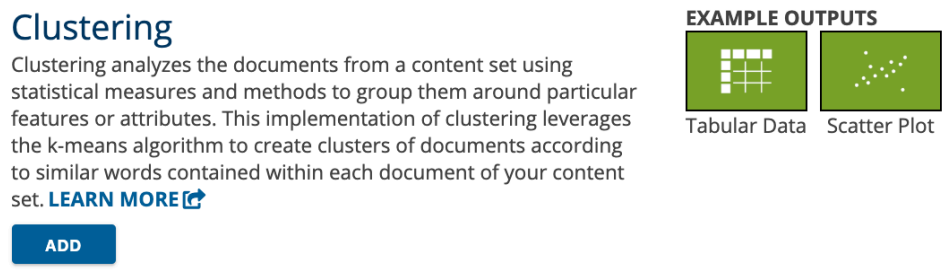
Here's an overview of what is shown on the 'Topics' page. You can also download a CSV of this data in it raw form, in the same format that a user of MALLET might expect to see.





This is the topic proportion by document, displaying the tabular data in a graphic format. The visualization is interactive, so click through to explore each aspect of the page. You can also download this visualization.

Clustering Tool Overview



K-means clustering is arguably the most challenging analysis tool to understand in the DSL.

- **Q:** What variables are being plotted on the scatter plot that is produced by the tool? The axes don't have explicit labels, and we haven't been able to find information on what features of the text are actually being plotted here.
- **A:** The x/y axis do not have labels because the scatter plot represents flattened multi-dimensional vectors. The points within the scatter plot represent documents within a Content Set. Their similarity to one another is based on their distance within these multi-dimensional vectors which are then flattened in 2-D space.

Digital Scholar Lab Implementation

The clustering tool in the DSL is built leveraging open source software, in this case scikit-learn's K-means Clustering algorithm, which is described in detail here.

Reading/Viewing

Ben Schmidt, 'Machine Learning at Sea', *Sapping Attention*, 2012 https://sappingattention.blogspot.com/2012/11/machine-learning-on-high-seas.html

Descriptive lectures/videos:

- https://www.coursera.org/lecture/machine-learning/k-means-algorithm-93VPG Watch the first 3mins 41 seconds of the Coursera video, unless you want to dig deep into the K-Means algorithm in which case keep watching!
- Lexos overview of K-means clustering: https://youtu.be/B7_cJBeofn4. There is a good explanation of what K-means is, starting around 30 seconds. Keep watching to see the visualization in 3D, vs. the flattened 2D representation currently available in the DSL. This may give you a better sense of how to interpret the clustering visualization.

Example projects using Clustering

See Ben Schmidt's Whaling project, above.

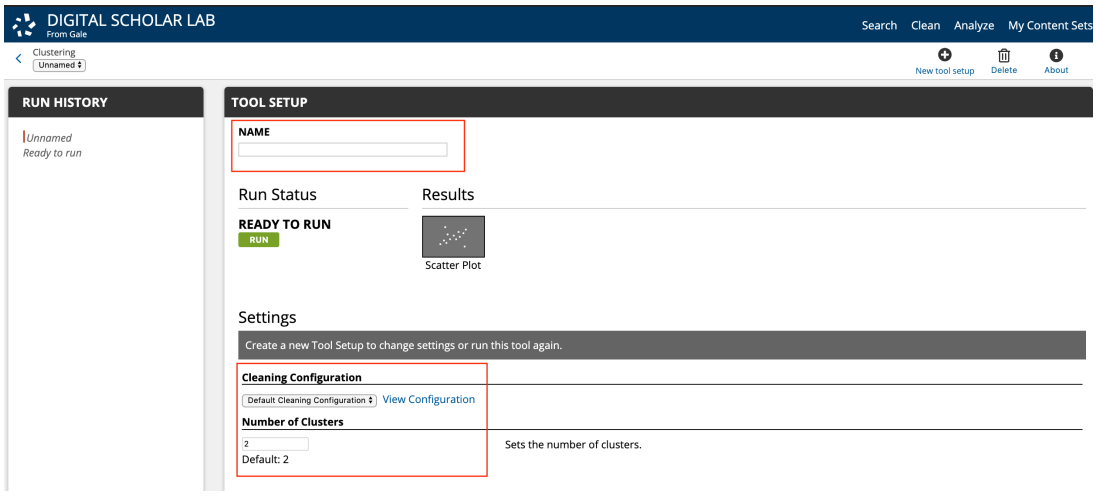
Configuration Options in the DSL

You have two configuration options for clustering.

1. Choose the cleaning configuration you want to apply.

2. Choose how many clusters you want the algorithm to group your content in.

Once you have done this, name your tool setup and click 'Run'.



Sample Project #2 Black America & The Law in the mid-20th Century

Synopsis

The mid-twentieth century in the United States was a time of immense transformation for people of color, particularly African Americans. Often referred to as the Civil Rights era, the 1960s and early 1970s saw protests, riots, violence, and eventually legislative responses to racialized injustice and discrimination. Numerous men and women fought to change how American society treated and understood the place of people of color, whether on buses, in streets, at home, in schools, or at work. Segregation was one of the main points of contention, and the focus of considerable legal effort. At the same time, cases involving African American men and women proceeded through the legal system. How that legal system, and those involved in it responded, implicitly or explicitly, to the pressures of the era in its trials and cases, can be seen by examining the US Supreme Court Records and Briefs.

Core Research Question:

- How does the language surrounding Black Americans shift in legal documents and records between 1950 and 1980?

More Precise Questions:

1. Are the Supreme Court Records and Briefs mentioning negro, black, or african americans more positive or negative in sentiment?

    1. Does this change over the course of the Civil Rights Era?

2. What are the most common phrases or collocates used in documents mentioning negro, black, or african americans?

3. What topics appear most frequently in documents mentioning negro, black, or african americans?

    1. Do these reflect themes that dominated Civil Rights Era conflicts?

4. What are the main concerns of legal records mentioning negro, black, or african americans during this period?

5. Are there any specific states, statutes, or other entities which stand out amidst the analyses?

6. Are there any differences between Document Types in the Archive (US Supreme Court Records and Briefs)?

Thinking about Methodology & Specific Tools

- Topic Modeling - we can use this tool to see if there are any themes or topics which cut across a collection of texts

- nGram - we can use this tool to track different kinds of phrases or terms which might occur together, and the number of times a phrase appears. In some cases names with several words - like United States or North Carolina - might appear

- Sentiment Analysis - we can use this tool to examine whether the contents of the documents were overall positive or negative according to the AFINN dictionary

Building the Content Set

Searching

The search for this content set was limited to one Archive, and a set Publication Date. Also, there

Limits & Parameters

Content Type ("Monograph")

Archive (U.S. Supreme Court Records and Briefs, 1832-1978)

Publication Date (1950 - 1980)

Keywords (in individual rows)

Entire Document: black, black american, black man, black woman, negro, african american, african americans, black americans

Statistics & Info

Content Set Name: 1950-1980 Black America & The Law - no subject terms

Content Set ID: 1579633475592

Number of Documents: 4289

Specific Tools

None of the tools required specific content sets.

Specific Questions

Question 6 could not be answered using the main content set, and so it required creating additional sub-content sets divided by Document Types: Briefs & Petitions; Statements, Memoranda, Appendices, etc.

Cleaning the Content Set

It took several attempts, or iterations [Link on Iteration], to get the cleaning right for each of these analyses. In the end, it requires several different cleaning configurations, as there were different stop words needed for different tools, as well as different approaches to punctuation. The main stop word list required adding single letters (for each letter, in case they appeared as abbreviations), as well as additional stop words. Also, some replacements were obvious from the test cleaning configurations.

Specific research questions:

1. No additional cleaning configurations required for Sentiment Analysis.

2. This required considerable iteration, in order to remove abbreviations and single letters, and additional stop words.

3. This required additional stop words that were unique to Topic Modeling, and were not the same as those used in nGrams - ‘state’ ‘court’, for instance needed to be retained for nGrams (for ‘United States’), but removed for Topic Modeling which treats individual tokens (ie. ‘United States’ can never occur in a Topic Model because it includes two words. The software doesn’t operate on phrases).

Running Tools

Selecting particular views for each tool was extremely straightforward. We selected an approach that reflected the size of our content set: we looked for more things and raised the bar for what made the cut for the results. Both were for very simple reasons: Topic Modeling as a tool statistically discerns what words are more likely to appear near to one another. More topics, and more words lowers the threshold of what is ‘significant’, meaning we get a finer grained picture of what the statistical analysis could suggest. In very similar documents, like court records, the chance is that there will be similar phases as questions and answers are posed, and rulings and arguments recorded. Selecting more words and more topics is a good way of sifting through some of these ‘known’ similarities, and can work in tandem with stop word lists to help ‘drill down’ into a large content set. For nGrams, we took a similar approach to thinking about potential ‘noise’ - we want to see what turns up. But the highest count in a result doesn’t always translate into the most meaningful or interesting. There’s a balance between number and noise.

Topic Modeling

It seemed best to cast a wider net in part to see what kinds of words appeared in the models created by the MALLET software that powers the tool. Requesting more words than the default, and double the topics produces finer grained topics, in reflection of the size of the content set. We opted for 15 word topics, and 20 topics.

Sentiment Analysis

This tool has no settings other than selection of the cleaning configuration.

nGrams

Like Topic Modeling, it seemed worthwhile to go beyond the default settings given the size of our content set. We raised the threshold for the number of times an nGram had to appear to be

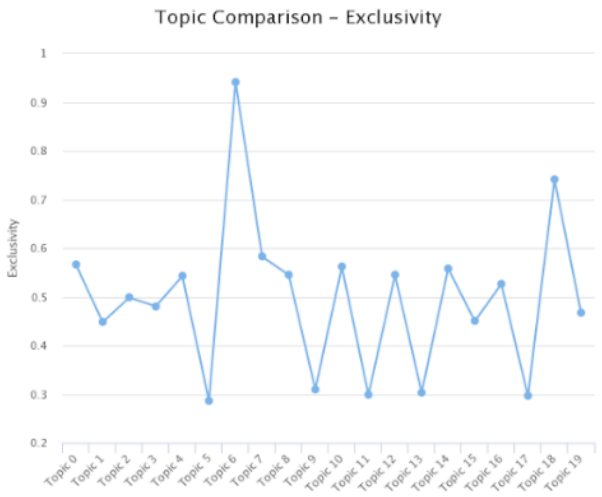
considered useful, and set it at 4. Equally, we wanted to find collocates rather than just single words, so we set the minimum nGram size to 2 (biGram), and the maximum size to 5. These settings translate into a search for “nGrams of between 2 to 5 words that appear in documents at least 4 or more times”.

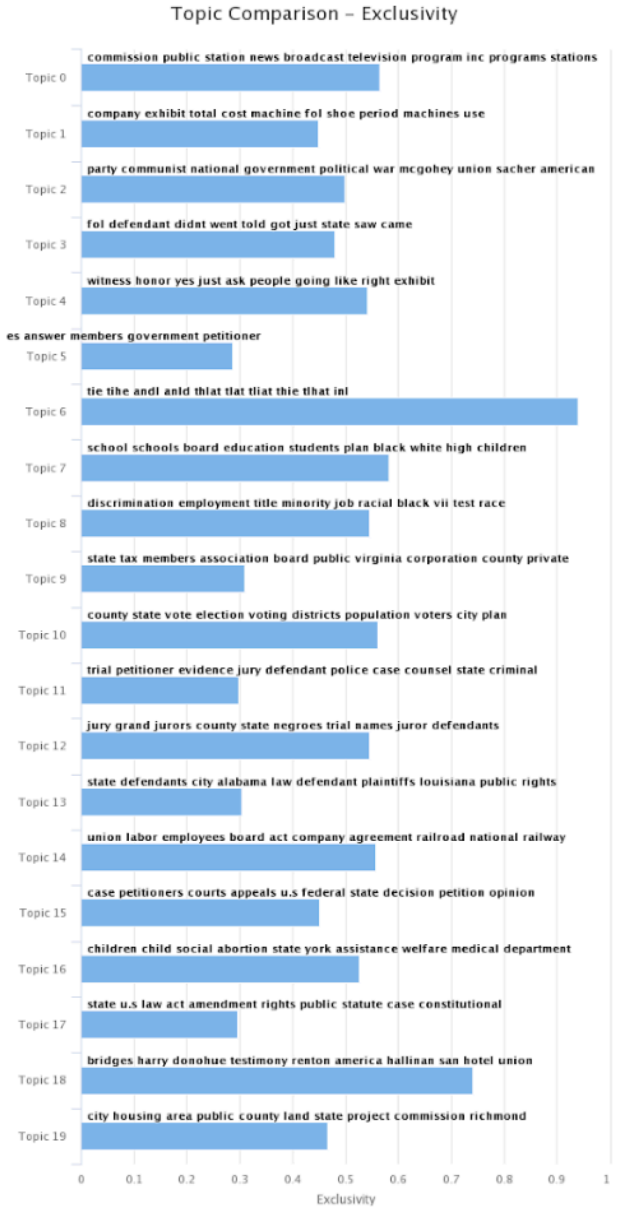
Understanding Results

Determining meaningfulness or significance is a critical set up in scholarly inquiry and how we pursue research questions. An essential element of this in a computational text analysis environment like the DSL is understanding that raw counts or ‘more hits’ doesn’t always mean something important - it could be noise, meaning its simply there because the content set wasn’t cleaned enough, or we didn’t use the right stop words, or perhaps it’s something expected and known. In the end, understanding results really requires understanding what we’ve asked of the tools in our configurations and settings, and how the results relate to the variables we’ve selected, and the algorithms which power the analysis.

Topic Modeling

Refining the stop word list for topic modeling took some time, as the normal stop word list didn’t include OCR error words. This is an example of the kinds of problems that can occur - notice Topic 6:


Topic Comparison – Exclusivity
Exclusivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Topic 0
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
Topic 6
Topic 7
Topic 8
Topic 9
Topic 10
Topic 11
Topic 12
Topic 13
Topic 14
Topic 15
Topic 16
Topic 17
Topic 18
Topic 19


Topic Comparison – Exclusivity
Topic 0
commission public station news broadcast television program inc programs stations
Topic 1
company exhibit total cost machine fol shoe period machines use
Topic 2
party communist national government political war mcgohey union sacher american
Topic 3
fol defendant didnt went told got just state saw came
Topic 4
witness honor yes just ask people going like right exhibit
Topic 5
es answer members government petitioner
Topic 6
tie tihe andl anld thlat tlat tliat thie tlhat inl
Topic 7
school schools board education students plan black white high children
Topic 8
discrimination employment title minority job racial black vii test race
Topic 9
state tax members association board public virginia corporation county private
Topic 10
county state vote election voting districts population voters city plan
Topic 11
trial petitioner evidence jury defendant police case counsel state criminal
Topic 12
jury grand jurors county state negroes trial names juror defendants
Topic 13
state defendants city alabama law defendant plaintiffs louisiana public rights
Topic 14
union labor employees board act company agreement railroad national railway
Topic 15
case petitioners courts appeals u.s federal state decision petition opinion
Topic 16
children child social abortion state york assistance welfare medical department
Topic 17
state u.s law act amendment rights public statute case constitutional
Topic 18
bridges harry donohue testimony renton america hallinan san hotel union
Topic 19
city housing area public county land state project commission richmond
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Exclusivity

Revision of the Cleaning Configuration to add ‘tihe’, ‘andl’, ‘anld’, ‘that’, ‘tlat’, ‘tliat’, ‘thie’, ‘tlhat’, ‘inl’, removed this topic upon re-run; yet it produced more Topics with less than useful words.

What we can see from the Topic Modeling are a series of possible themes within the US Supreme Court Records:

- city, state, public, police, white, petitioners, peace, alabama, law, people
- party, communist, committee, government, bridges, testimony, member, union, activities, members
- county, vote, election, voting, state, districts, city, population, voters, political
- state, plaintiffs, defendants, action, plaintiff, defendant, motion, complaint, county, law
- school, schools, board, education, plan, students, black, white, high, racial
- jury, grand, jurors, county, negroes, defendants, state, juror, names, trial
- state, u.s, act, rights, law, amendment, case, federal, public, statute

How we decide what is the most meaningful or significant measure in these results can be tricky - it depends on what our question might be, of course. The results from Topic Modeling could be meaningful simply by being unexpected or new, suggesting something that we didn’t already know, or perhaps were unaware of. At the same time, they might confirm something we already know, and can act as a touchstone to confirm that we’re on the right course with reading or analysis, or both. We see in the topics returned above, that some are clearly relevant to the Civil Rights Era. Some may not be. There are other ways of understanding these results in relation to the content set, however.

The software powering the Topic Modeling tool - MALLET - is particularly well known and refined. It offers very rich results, which can be investigated in a number of ways by looking at the Topic results section in the tool view, and by selecting ‘Topic Comparison’. Here we see a list of measures describing how the topics relate to the content set, and the analysis.

Tokens: This metric measures the number of words from the content set assigned to this topic.

Document Entropy: This metric measures the probability any given document will be in the topic. Low entropy topics will come from a small set of documents while higher entropy topics will come from a wider set of documents.

Average Word Length: This metric measures the average number of characters in the top terms. Because longer words are assumed to be more meaningful, higher word lengths indicate more specific topics.

Coherence: This metric measures how often words in the topic appear next to each other. The closer to 0, the more likely it is that terms occur next to each other.

Uniform Distance: This metric measures the distance between a uniform distribution and that of the topic's distribution over the words assigned to it. The larger the distance, the more specific the topic.

Corpus Distance: This metric measures the distance between the frequency of words in the content set to frequency of the words assigned to the topic. The larger the distance, the more distinct it is from the content set as a whole.

Exclusivity: This metric measures how exclusive the top terms for each topic are to that topic. The higher the value, the more likely that a topic's top terms do not appear as top terms for other topics.

These measures allow us to explore how the topics created by the software relate to each other, and to the content set from which they're drawn. We can look at raw counts, but also the possible kinds of interrelation the words have with each other and with the content set as a whole.

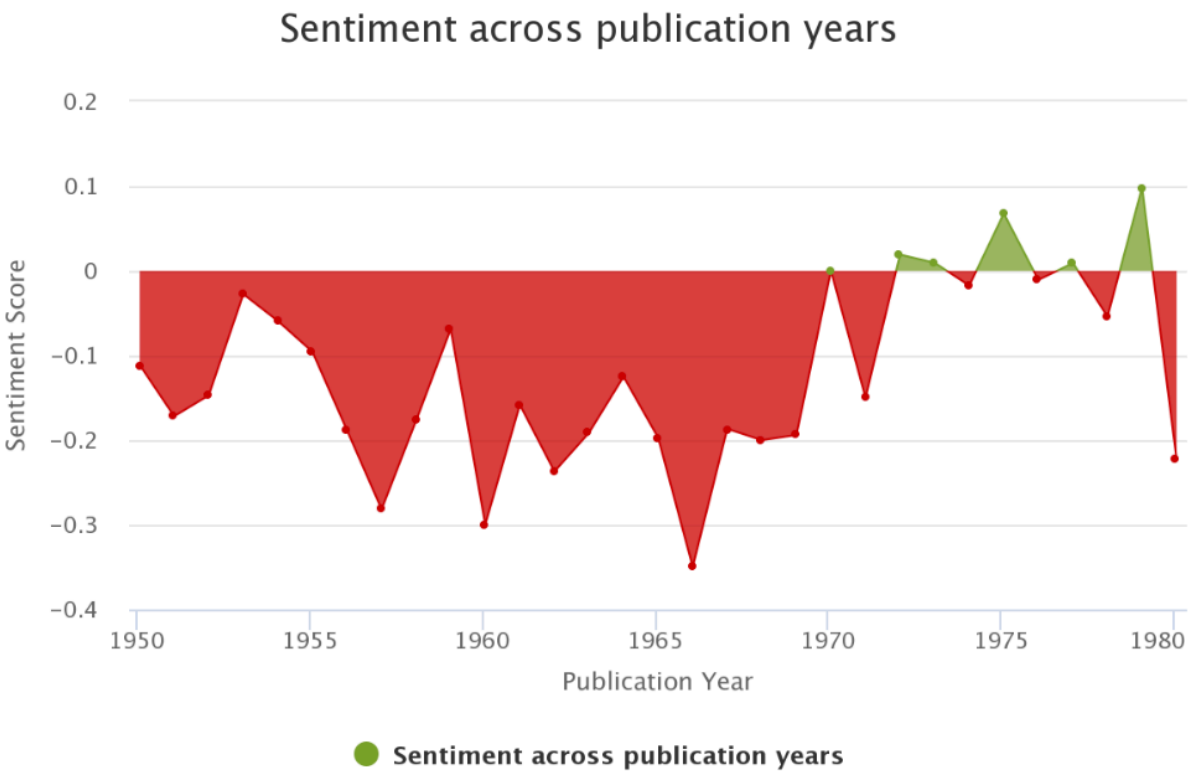
- Coherence overlaps conceptually with nGrams - can we compare the two tools in any meaningful way?
- Several measures point towards specificity, or to put it another way, the precision or clarity of a topic within the content set and documents: Document Entropy and Uniform Distance offer ways of examining how specific topics fit into the the content set, as well as within documents.
- Another question is - how unique might a topic be? Corpus Distance offers a means of thinking about exceptionality of a topic.

We can provide names for the topics created by the tool, so they can be referenced later on. These names will appear in the Topic Proportion view, replacing "Topic [number]", making it easier to navigate what the results are.

The Topic Modeling tool allows us to move through these measures and the topics themselves through the Topic Proportion view. We can select specific documents by title, and compare which Topics show up. This is particularly useful as a means of drilling into the content set itself.

In our results we see that the topics which have the greatest presence in the Proportion view in each case are those concerned with procedural or legal terms. This isn't surprising, but it also isn't particularly useful.

Sentiment Analysis



This seems unsurprising, given the fraught nature of the Civil Right Era, with its protests, violence, and focus on systemic racial discrimination. It raises several NEW research questions, however:

1. Can we associate or link the years with the lowest sentiment scores with particular events or moments in the Civil Rights Era?

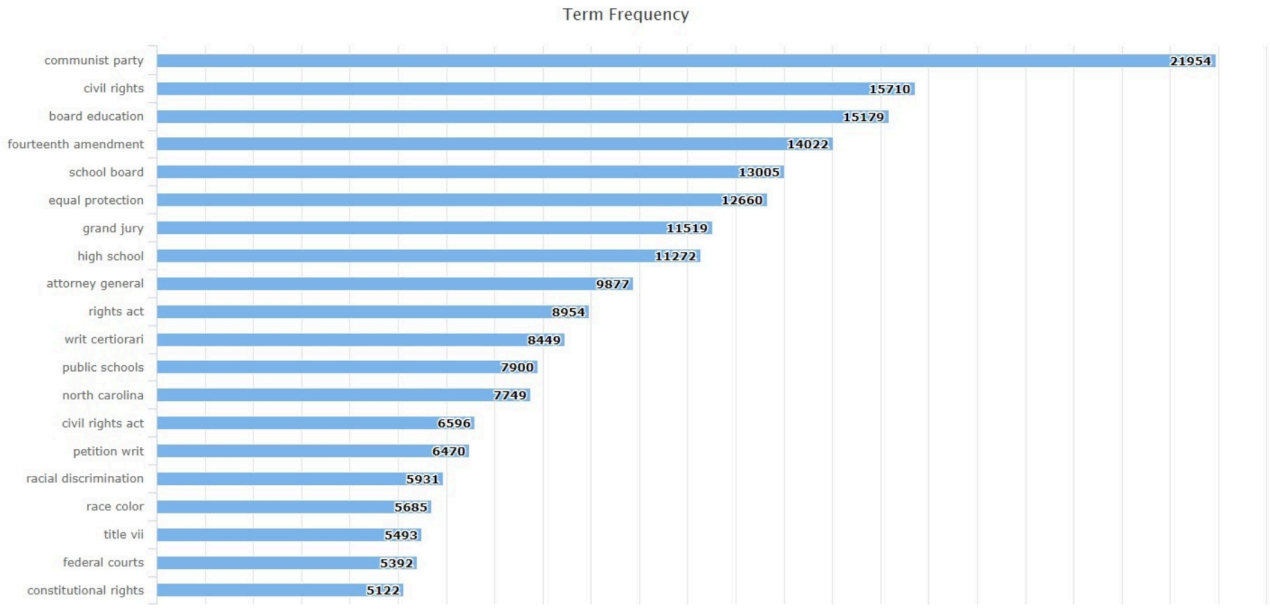
2. Why might there be years with positive scores throughout the 1970s?

3. To what degree can we attribute these scores of sentiment to the issues that shaped the Civil Rights Era versus the general nature of legal cases as contestations or conflicts, or matters of friction? IE, will most of the US Supreme Court Records have a negative sentiment because they deal with legal matters, or is there something specific about the records dealing with Black America that makes them different or stand out?

nGrams

Configuration: Min 2 Max 5, Threshold 5

Cleaning: US Legal No Punctuation No Numbers



Here we have the top nGrams, after rerunning the tool several times with different cleaning configurations. Strangely enough, the most frequent bigram (nGram with two tokens or words), is ‘Communist Party’. The next few however deal explicitly with themes we’d readily expect - ‘Civil Rights’ ‘Board Education’ (likely for Board of Education), ‘Fourteenth Amendment’, ‘School Board’, ‘Equal Protection’, etc. All of these pertain to key legal battles surrounding the issues of race in the Civil Rights Era of the 1960s and early 1970s. It also suggests that the most frequent issues the US Supreme Court handled in regards to the Civil Rights era had to do with schooling and access to it, and segregation. Importantly though, despite the fact racial discrimination was

the heart of the issue, it comes much lower on the list following equal rights terminology. This suggests that while appellants were well aware of racial discrimination, they sought legal protection using equal rights arguments, rather than focusing on discrimination, as the basis for their legal filings. Remember, we didn't search for anything related to 'Civil Rights' or 'Discrimination' - these appeared as a matter of analysis.

Such outcomes confirm many of the things we know about the Civil Rights Era, and legal proceedings. But the Communist Party prominence suggests something that could be followed up - at least two NEW Research Questions:

1. Why does the 'Communist Party' appear so prominently in texts mentioning Black Americans in the US Supreme Court Records between 1950 and 1980?

2. What connections were perceived between the fear and conflict of the Cold War, and race, in the struggle for Civil Rights in mid-20th century America?

Research Outcomes

It's clear from this brief test project that large scale analysis of US Supreme Court Records provides insight into broad themes and concerns that we expect to find from Civil Rights era legal proceedings which mention Black, African, or Negro Americans. At the same time the results also suggest new questions we could consider.

Original Questions

1. The US Supreme Court Records and Briefs are clearly quite negative in tone. Even when divided by Document Type, the negative sentiment is striking. It fluctuates; and outside of Briefs and Petitions, tends towards slightly more positive tone over the 1970s.

2. 'communist party', 'civil rights', 'board education', 'fourteenth amendment', 'school board', 'equal protection', 'grand jury', 'high school', 'attorney general','rights act', 'civil rights act', 'racial discrimination', 'race color','title vii', 'constitutional rights'

3. This requires a more precise cleaning configuration to remove problematic OCR. Some of the topics we do see, however, do reflect Civil Rights themes - in particular segregation and discrimination, as well as equal rights.

4. This is not easily discerned from our outcomes. Topic Modeling suggests some possibilities, but needs to be clearer. nGrams also suggests dominant themes, but it is not as explicit as they could be.
5. In the nGrams we see the presence of the Fourteenth Amendment and Title VII - both of which expressly forbid discrimination on the basis of race. The former, as part of the Constitution, and the latter, as a section in the 1964 Civil Rights Act.
6. It would seem so, especially when it comes to nGrams. More exploration is required.

New Questions

1. Why does 'Communist Party' appear in the nGrams?
2. How does 'Document Type' affect the outputs of the analysis tools?

Revising the Content Set

In light of our initial outcomes, we can now think about revising or subdividing our original content set in order, to pursue our new research questions. One of the possible means of getting at the concerns of those within the legal establishment and the Civil Rights Era is to create sub-content sets derived from the original content set using various metadata. Legal Briefs and Petitions were requests and filings made to the US Supreme Court - they are specific genres, noted as Document Types in the DSL.

Perhaps the first question about 'Communist Party' will be clearer when we create sub-content sets consisting of different Document Types. Creating content sets which contains only briefs and petitions, on the one hand, and another with the rest of the Document Types, might allow us to gain a different view of our research question by contextualizing all of the analyses we've conducted with the issue of 'genre'. Genre brings certain kinds of questions that can shape how we think about the outcomes of the tools:

1. What is a legal 'brief' or a legal 'petition'? Are they the same? Who writes them, and why?
2. What kinds of information do Briefs and Petitions contain? As a genre, do they have particular characteristics?

3. How might knowledge of a Document Type - a genre, or a form of writing - shape our research inquiries? What can we learn by understanding the form of a document, and what it might contain textually?

Sub-Content Set - Briefs and Petitions

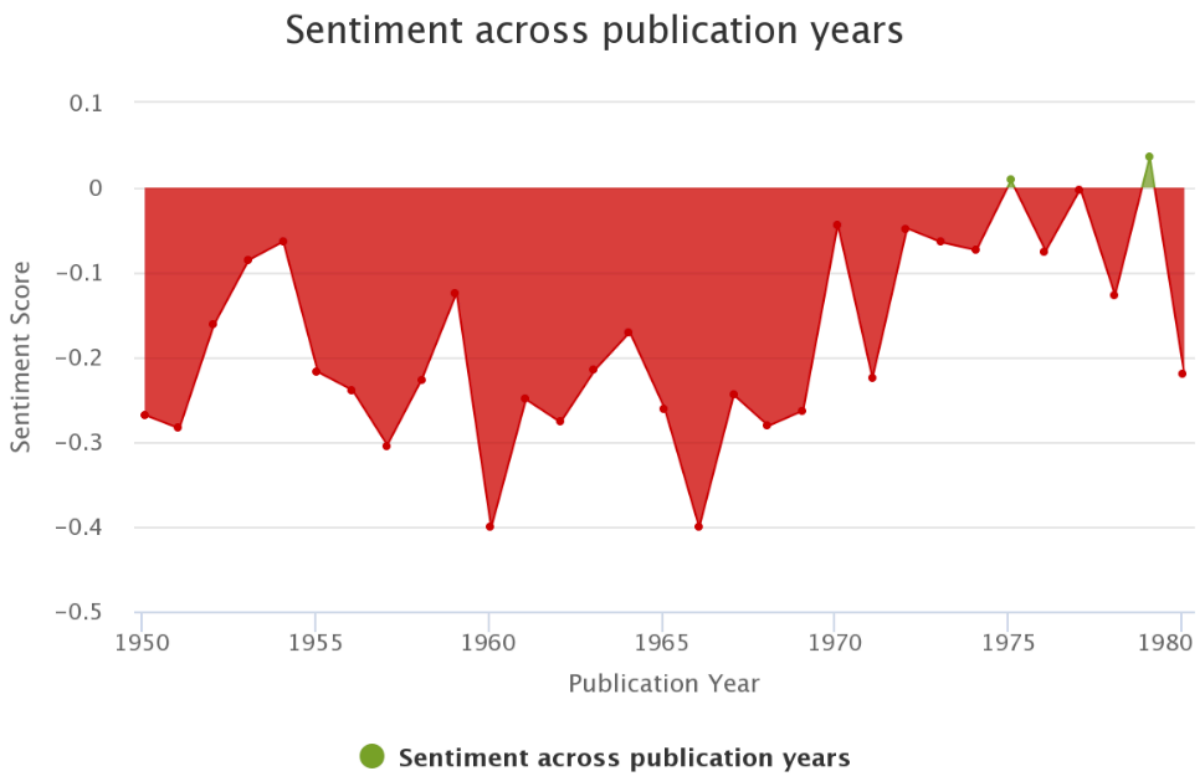
Topic Modeling

- jury state trial grand u.s county petitioner negroes death juror
- discrimination u.s title minority racial employment vii black cir program
- school negro race white schools state education public law equal
- school schools board plan education racial black students desegregation county
- city housing property public private state u.s park racial discrimination

This sub-content set seems to be more precisely concerned with the issues of the Civil Rights era than the main set.

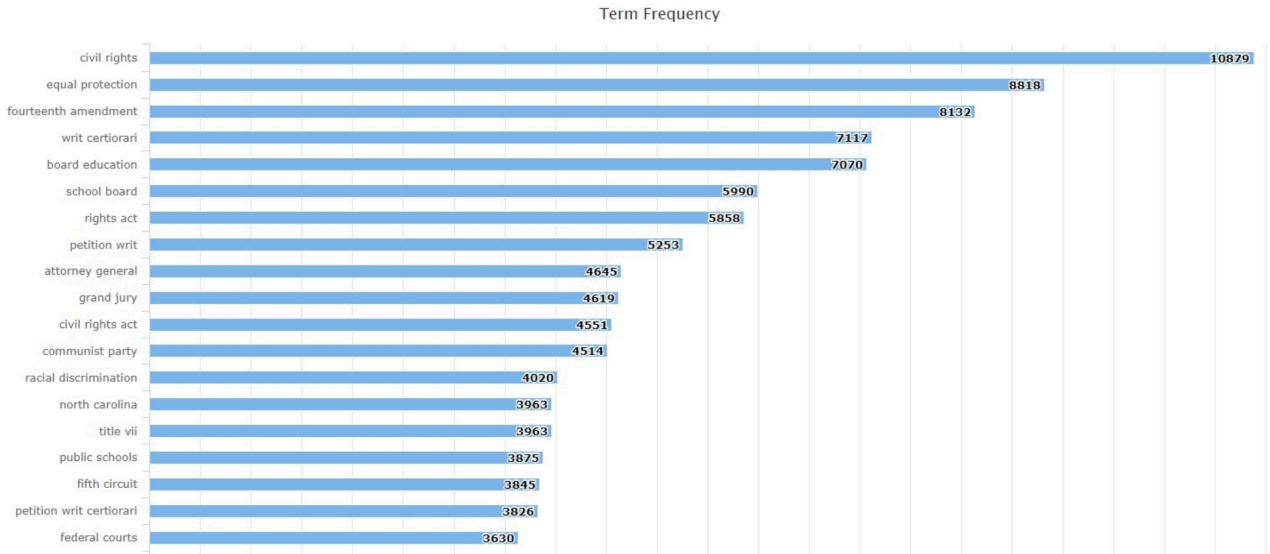
Sentiment Analysis

It would seem that Briefs and Petitions remain almost consistently negative when it comes to sentiment analysis, except for some small positive years in the mid to late 1970s. This could suggest that as genres, Briefs and Petitions are normally negative in tone or sentiment, and that perhaps the negativity isn't associated with Civil Rights Era issues. Or, it could suggest entirely the opposite. A good check would be running the same Sentiment Analysis on a new content set containing documents that aren't Briefs or Petitions in the main content set.



nGrams

We can see that the Communist Party no longer appears as the top nGram, and that the remainder of the top 10 or so nGrams are focused solely on Civil Rights Era issues. The communist party does appear as the 12th highest nGram, in a dramatic drop from first place. Clearly it retains some relevance, but is not as prominent as in the main content set which included memoranda.



Sub-Content Set - Statements, Memoranda, etc.

Let's see what the other Document Types look like with the same analyses.

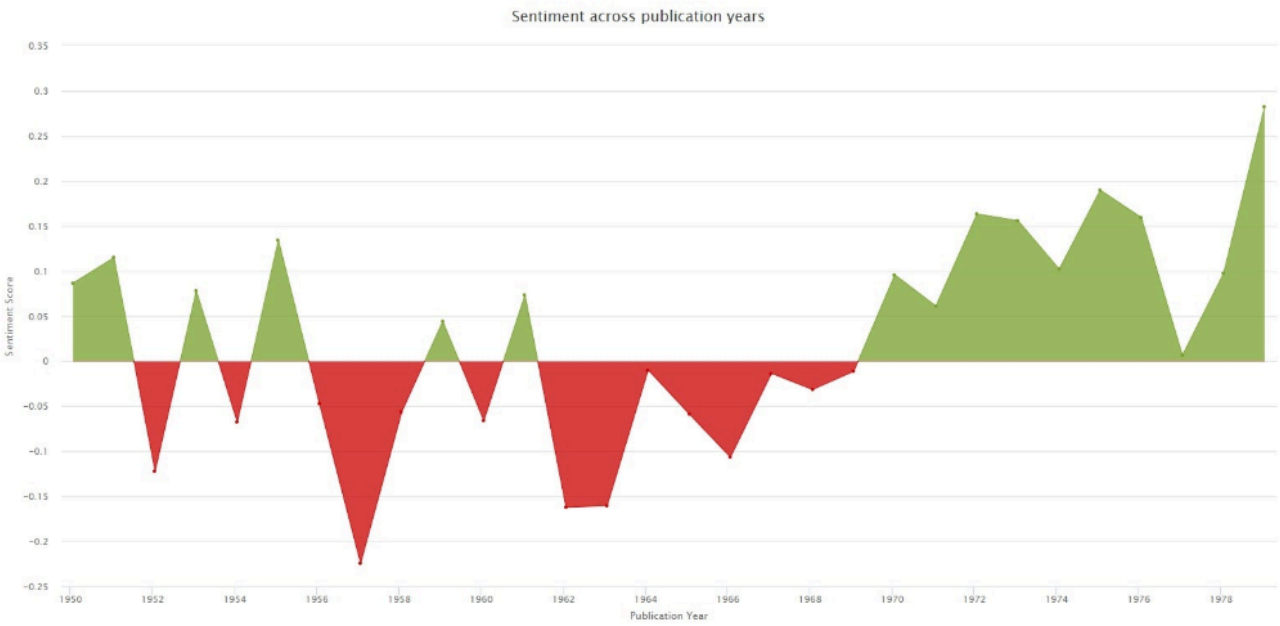
Topic Modeling

- school schools board plan education students high black white children
- party communist national exhibit mcgohey war sacher political objection class
- party communist bridges testimony harry committee union answer member donohue
- state defendants fol defendant plaintiffs alabama county plaintiff motion complaint

There are clearly some topics related to the Civil Rights era, but also those focused on the communist party topic, as well as other themes. This suggests a greater diversity of content in this sub-content set than the other.

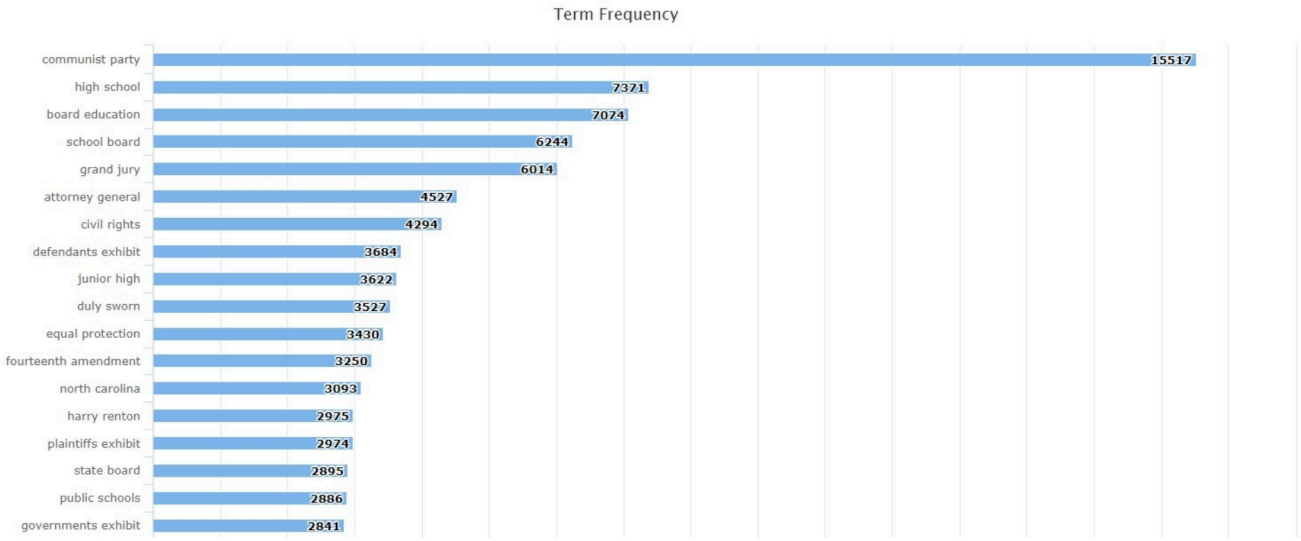
Sentiment Analysis

Sentiment Analysis is dramatically different. Clearly Statements and Memoranda had very different tones when it came to race earlier and after the 1960s. But the core Civil Rights Era of the 1960s was still overwhelmingly negative in sentiment.



nGrams

Notice, ‘Communist Party’ still appears dramatically in first place. It ranked 12th in Briefs & Petitions. Still many of the same nGrams appear, suggesting that Civil Rights era concerns remain dominant throughout the documents, regardless of their type.



Reflections on Method

This project allows us to reflect on how we build content sets, and what iteration means when revising and cleaning them prior to analysis. The comparisons between our two sub-content sets - Briefs & Petitions and Statements, Memoranda, etc. - allows us to clearly see how the parameters or fields we use to build a content set can affect or shape the kinds of results we obtain from the analysis tools. Apparently, ‘Communist Party’ was a matter of discussion in memoranda and not briefs or petitions to the US Supreme Court between 1950-1980 in cases involving mention of Black, African, or Negro Americans.

Content Set Building

Building this method was fairly straightforward as our research question was focused precisely on one Archive source, and a clear and defined publication date window. The purpose was to use the DSL to explore and discover possible new research questions from a large set of documents, rather than pursue precise questions around a specific author, genre, or perhaps another variable. That said, it’s clear that as we worked with the Content Set, new questions emerged

which required refining the Content Set. We built two new versions of the original Content Set, dividing it up using the Document Type values. We could have done this at the outset, as well. Another possible avenue for more precise Content Sets - allowing us to explore more precise research questions - would be to build additional Content Sets based on authorship, or by case. Such precision requires increasing in depth knowledge and expertise with the subject matter itself. The DSL can build such Content Sets, but you, the researcher, need to have sufficient experience with the subject matter in order to define the parameters of your research question and how it relates to the Content Set you might build. Having a list of cases involving Civil Rights would offer a rich picture of views of African Americans during this Era. However, it would also mean determining what cases involved civil rights issues. Were they just those grappling with civil rights statutes? Or can we learn anything about how the legal system treated or viewed African Americans in cases that didn't address civil rights statutes directly? These are two different kinds of questions, and require different content sets.

Iteration

Iteration for this project focused more on the Cleaning Configuration than working through the Content Set itself. This is common - the Default Cleaning Configuration is only a base starting point. Each project will need to have at least one Cleaning Configuration of its own, if not more. And these require testing and rerunning until the results you receive from the Analysis are meaningful and uncluttered with 'noisy' data. Over the course of this project, it became increasingly clear that while the Content Set was usable for Sentiment Analysis without much tweaking of the Cleaning Configuration, nGrams and especially Topic Modeling required some tinkering to get the right Cleaning Configurations. This has much to do with how the tools work as with problematic OCR. Sentiment Analysis matches words it already knows, meaning if something is misspelled, it is ignored. Cleaning, in this context, only adds words into the mix; problematic OCR isn't analyzed. The other two tools work with the actual words within a document, and so if problematic OCR words have a significant enough presence, they'll show up like any other words. When it came to nGrams, there were nGrams which included correctly spelled words, but which had little usefulness to our research question: often legal documents contain many sections, and those section numbers showed up, along with abbreviations as they're in every document. Adding them to the stop word list removed them from being included in the nGram analysis. Finding them all took several tries, but it was possible to get a fairly clean and meaningful output after a few iterations.

Topic Modeling, however, took much longer as the nature of the tool is to collate words that statistically appear often with one another. While problematic OCR was usually ignored by the nGrams tool, it became quickly apparent with Topic Modeling because the tool returned results

suggesting that problematic OCR words were themselves constituted one or more Topics. Rerunning the Topic Model analysis, allowed the outcomes to be used to revise the Cleaning Configuration; and then the analysis was run again.

Another problematic element with Topic Modeling is finding the right balance between results we expect due to the genre of the document, and eliminating words through the stop word list so we might find things that could be meaningful. In the Topic Proportion view, the most prevalent topics were those concerned with procedural or legal terms, no matter how much we cleaned out OCR or lesser words (like prepositions, conjunctions, articles, etc.). But removing terms like 'statute' or 'federal' might alter the themes we're interested in investigating at the same time they often appear as unified topics. This is to be expected, but it's another kind of 'legitimate noise' - results we need to wade through in order to find the themes of the Civil Rights era we're looking for.

Understanding Outcomes

What do these outcomes tell us about our research interests? We can understand outcomes in a variety of ways: how they answer the questions we originally posed, and how they suggest new questions and new avenues for research.

It's clear that our outcomes corroborate many of the things we already know about the Civil Rights Era:

- It coincided with the height of the Cold War, and anxiety surrounding Communism
- Cases focused on discrimination on the one hand, and equality on the other
- School segregation was a primary concern of cases shaping the experience of the legal system among Black, African, or Negro Americans

What is unclear from our outcomes, however, is whether not the sentiment of the documents in our Content Set is a matter of their genre and context (ie legal documents are always 'negative', perhaps because they involve contestation or argument), or actually related to racism and discrimination. It's likely a combination, but there's no assured method of teasing these two apart.

We can also examine the outcomes as a way of understanding and reflecting on our methodology. What kinds of fields or content set building techniques might we use to create more focused or precise collections of documents that could better answer our questions? How could we manage whether a document actually discusses what its metadata says it does: in other

words, do the contents match what cataloguers or others, even perhaps the original authors, tell us? Closer reading of the documents prior to adding them to a document set will help us discern whether they should be included in a content set, or not.

Revising Questions

Once we saw our initial outcomes in this project using the main Content Set, it was clear we could revise our research questions somewhat, both making them more precise, but also possibly exploring a new question around the presence of something unexpected - the 'Communist Party' biGram. Where did this come from? Is there any way to isolate it to discern why it might appear so prominently in the Content Set? Why does it appear in documents which mention Black, African or Negro Americans in the mid-20th Century? Is there an overlap between the Cold War and fears of communism and the racial tensions of the Civil Rights Era?

As discussed in the guide, it's not only normal to revise your research questions after running analysis tools on a Content Set, it's an integral part of the research process. Often, analysis will turn up new questions which could lay beyond the scope of your current project. This is how researchers develop new projects and lines of scholarship - by following clues and new questions that come up while pursuing other research.

Limitations

As useful as these results might be, there are limitations to what kinds of cleaning and analysis that can be done with the DSL.

- Currently, there is no method within the DSL to compare all words against an English dictionary in order to identify problematic OCR. Iterating through Cleaning Configurations using Topic Modeling is a sound method for finding problematic words, it is a time consuming process. Replacements can be made easily, allowing problematic OCR to be fixed, but there's no method of finding all instances of misspelled words.

- This project did not build content sets using actual cases, nor were they built following a close reading of the documents included in the sets. A more precise content set could be built by determining, following examination of each document, whether or not it was appropriate to include in a Content Set focused on the specific parameters of the project.

- We haven't fully considered the difference between raw numbers or 'counts' and statistical measures as distinct ways of thinking about significance. Although the Topic output allows us to examine counts, the Latent Dirichlet Allocation method used by MALLET is a kind of

> prediction of the likelihood of words appearing with each other. It's suggestive, in other words, of something significant. The nGrams in contrast are raw counts across the content set. Having more documents, or longer documents - ie documents with more words - would increase those counts. Numerical presence, however, doesn't always translate into intellectual significance or meaningfulness.

Beyond the Lab

Presentations

All of the tool outputs can be downloaded as images to use in powerpoints, or embedded in webpages or other ways to present your work.

New Visualizations

It's also possible to download the data which power the visualizations as comma delimited (CSV) or javascript object notation (JSON) files, allowing you to create and format your own visualizations. If you have the skills, it's possible to collate or create new visualizations that may combine outputs from similar visualizations into one, allowing you to compare and contrast in new ways that the DSL tool does not. The Topic Modeling tool downloads are especially rich with possibilities for new visualization. The Topic view download is large and contains results for each document and measure for the Tool - much more data than the Topic Model visualizations can currently display. If you're a programmer, this is the ideal place to start to explore the data created by the DSL using other tools and visualization designs.

Refining the Content Sets

Understanding the limitations of the DSL allows us to consider what can be done to both to build content sets, and to use the results produced by its tools. Building content sets using US Supreme Court case records and documents would provide a completely different method of considering the themes of Civil Rights and racial discrimination. At the same time, it would also isolate analysis to documents that are explicitly tied to such legal questions. We know, however, that these themes cannot be, and were not isolated to explicit cases.

Similar Projects

Appreciating how we can structure a project or line of inquiry can be shaped as much by the tools and content we have, as by modeling similar projects that explore similar kinds of documents. The Old Bailey Online project (https://www.oldbaileyonline.org/) involved large scale

analysis of court proceedings from the main municipal court in the City of London. Although its content has been encoded, and cleaned, and its platform doesn't contain the same kinds of tools, as a project it offers a way of considering how examination of the US Supreme Court records might be modeled. It provides approaches, questions, and methods which could be applied and tested using our current project's contents and tools.

Week 8 Discussion

This week's discussion post will focus on the tools you have used and analyses you've run over the week. To recap, we looked at Sentiment Analysis, NER and Topic Modeling within the DSL, with some discussion of the extensibility of the platform through the export of OCR texts and raw CSV/JSON data.

1. Include a summary of reading you completed this week. Aim to discuss 2-3 articles or projects.

2. Provide a complete record of your work this week, including processes, triumphs, research insights and questions. Focus particularly on teasing out the meaning of the analyses you have run. How much cleaning did you have to do to generate output that was workable? More often than not, analysis results can be less than fascinating; if this is the case with your project, how can you reframe your research question or reformat your content set to dig a little deeper? If you tried running analyses using tools outside the DSL, record your work in this section.

3. Update your Project One Pager to reflect the current status of your work.

Replies are due by Tuesday Week 9 by 11.59pm

WEEK 9a Timelines & StoryMaps

How to create a StoryMap

Instead of writing a final paper for this class, you will be building an interactive StoryMap with narrative, images, and links to completed work using Knightlab's StoryMapJS. The interface is straightforward, and your final submission will be a URL to the narrative map you have created.

Sample Projects

Bosch Garden: https://storymap.knightlab.com/examples/bosch-garden/

Hockey, Hip Hop and other Green Line Highlights https://www.minnpost.com/stroll/2014/06/hockey-hip-hop-and-other-green-line-highlights/

Non-NPL Superfund Sites in North Carolina https://uploads.knightlab.com/storymapjs/8022dd050b6f5683410e66e127965163/historic-map-of/index.html

How to Create a StoryMap

Video tutorial : https://youtu.be/X4gOXga-Q_w

Tips & tricks (from the StoryMap creators)

1. Keep it short. We recommend not having more than 20 slides for a reader to click through.
2. Pick stories that have a strong location narrative. It does not work well for stories that need to jump around in the map.
3. Write each event as a part of a larger narrative.
4. Include events that build up to major occurrences — not just the major events.

Media sources

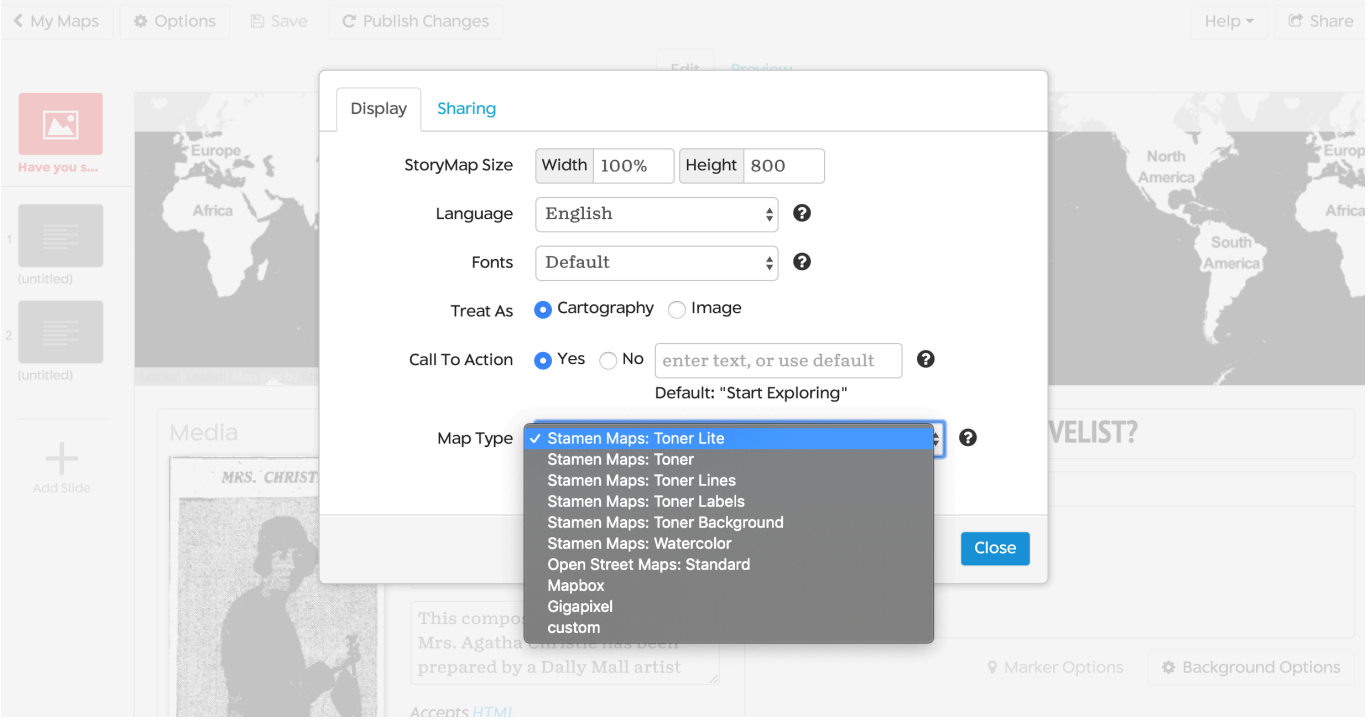
StoryMap JS can pull in media from a variety of sources. Twitter, Flickr, YouTube, Vimeo, Vine, Dailymotion, Google Maps, Wikipedia, SoundCloud, Document Cloud and more.

Choosing a Base Map

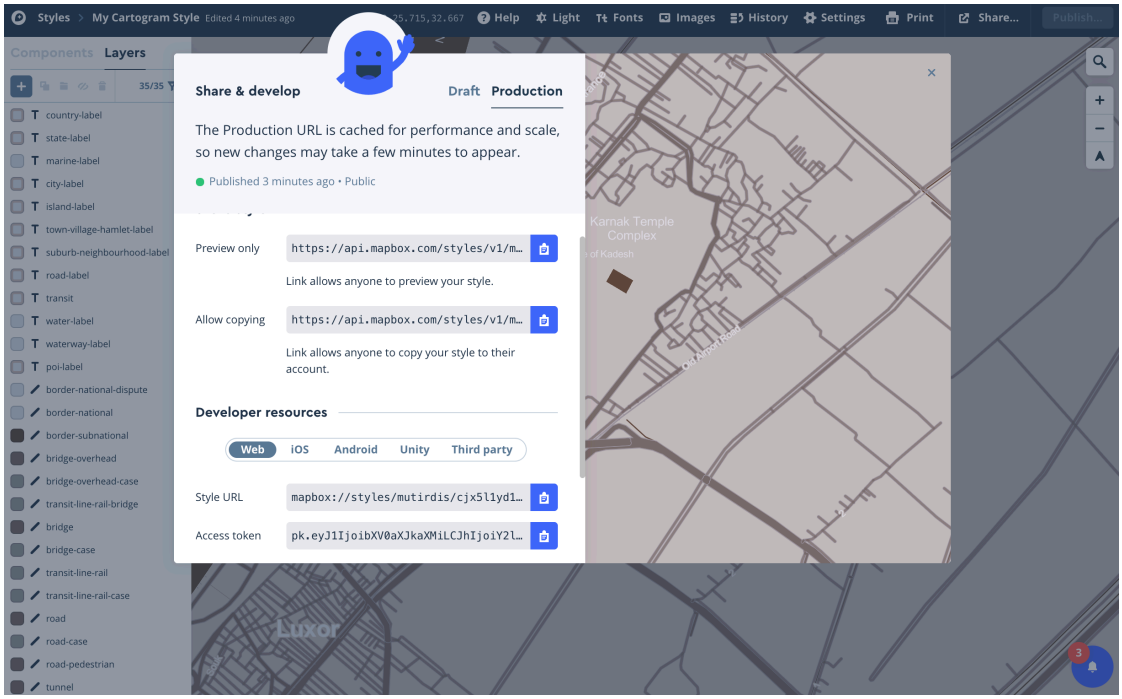
There is no requirement to do so, but if you wish to customize the base map, you can create a map at www.mapbox.com (sign up for a free account), and save the Mapbox map ID.

Then follow these directions for StoryMap:

Choose **options** from the top left of the StoryMap window, then from the **map type** menu choose the type you like.



Select **Mapbox** to enter the ID of your Mapbox map.

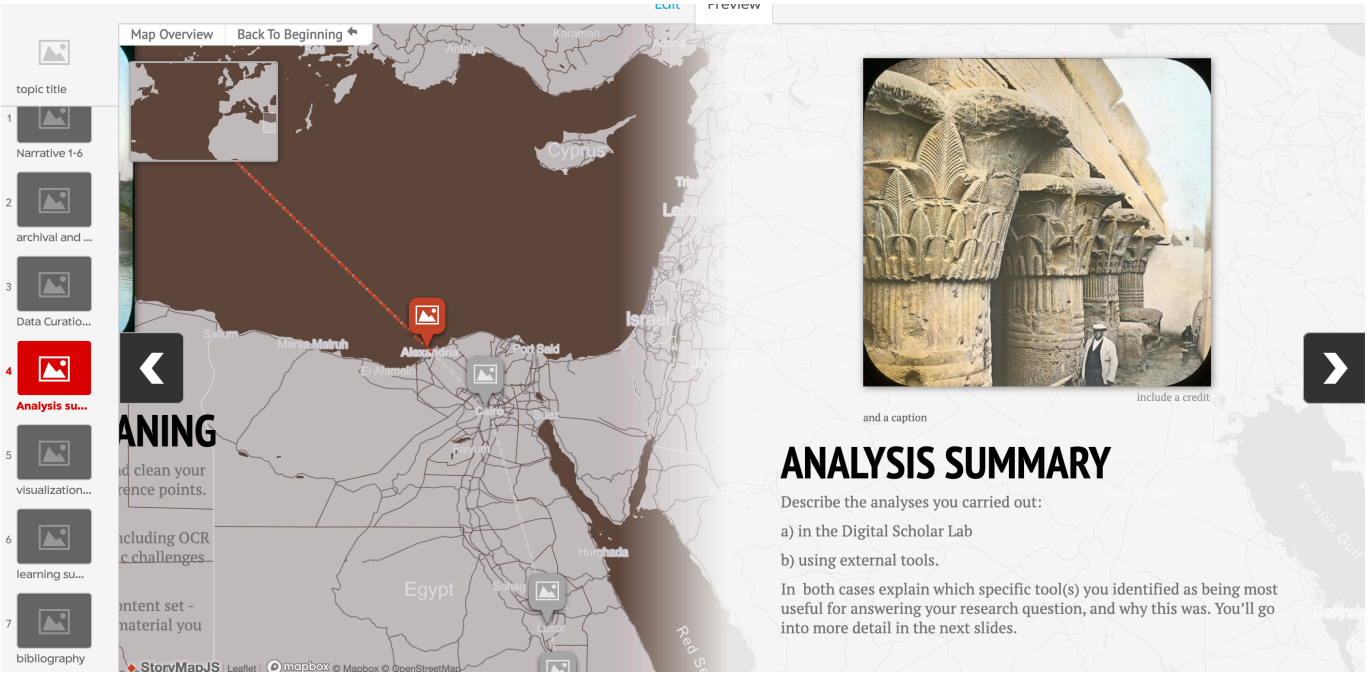


To get your Mapbox ID, select 'Share' from the top right of your Mapbox map, then select 'Web, Style URL' and copy it to your clipboard ready to paste into StoryMap. You'll also need to create a public access token which you can do from your account page.

Select **custom** to enter the URL for a tile server. If the server supports multiple subdomains, enter them as a single string in the Subdomain field (e.g. subdomains 'a', 'b', and 'c' should be entered as 'abc').

Final Project Rubric

You will be presenting your work as a visual narrative using StoryMapJS. To guide you, here's a visual template which you can use in conjunction with the step-by-step directions, detailed below.



General points:

Proofread your work carefully.

Include credits and a caption for each image. You can find citations for DSL material on the 'Document Explorer' page.

Choose an appropriate map background, or you can create a custom map in Mapbox.

You can also customize the background to the narrative portion by choosing 'Background Options'.

Here is a template for you to follow for your StoryMap. Ensure that you include all these elements, and feel free to add more if you feel inspired to do so.

```
Title slide
```

Give a high-level overview of your topic.

Describe your research question here. You can refer to your Project One-Pager to guide you in this section.

```
1.-6. Narrative - in at least FOUR slides describe your research topic, highlighting the main points, or sequences of events.
```

Over the next six slides (at minimum), summarize the main points of your research narrative. For example, for a topic related to 'sea serpents' in history, I might describe earliest recorded sightings, descriptions, reactions etc. Include images taken from Gale Primary Source material or other open source resources, with appropriate citation. Supplemental material can include video/ website content. Your tone should be engaging and informative.

```
7. Archival and research resources
```

This is where you will give an overview of the material that has informed your research. This will include original primary source documents, Gale Primary Source resources and any open source material you have used in building your project.

```
8. Data curation and cleaning
```

You'll include a summary of the steps you took to curate and clean your data here, using your work log and discussion posts as reference points. Include:

1. A summary of the quality of the OCR you worked with, including OCR confidence levels in the DSL, and discussions of any specific challenges you encountered.

2. A description of the strategies you used to curate your content set - what search terms did you use? A description of the other material you sourced, as relevant.

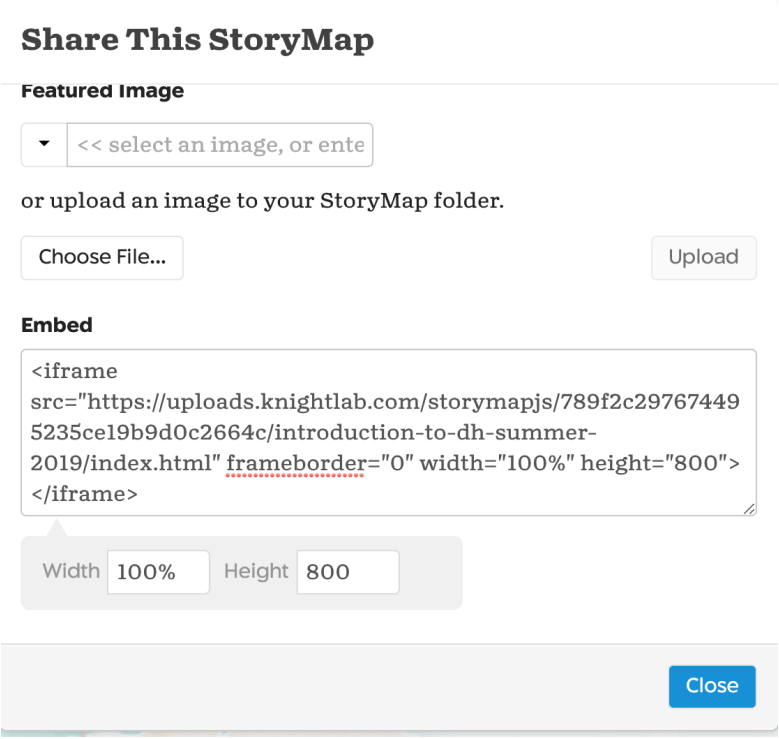
3. A description of your cleaning process. How effective was it? What specific steps did you take to clean your data? How did cleaning in the DSL differ from using, say, Lexos?

```
9. Analysis summary
```

Describe the analyses you carried out:

a) in the Digital Scholar Lab

b) using external tools.



In both cases explain which specific tool(s) you identified as being most useful for answering your research question, and why this was. You'll go into more detail in the next slides.

`10.-11. Visualization and Analysis Result 1-2`

You should have at least one, preferably two visualizations and associated discussion.

Give a detailed description of the specific analysis tool you used to answer your research question, and the types of questions the tool is best suited to answer (based on the reading you've completed for this class).

Include your downloaded visualization.

1. If using a DSL visualization, what configuration options did you apply? If using an external tool, describe the setup you used in generating the visualization.

2. Discuss the analysis results in your visualization. How well do they answer your research question? Are the useful and/or intelligible? How much additional data curation did you have to do to add meaning to your analysis?

`12. Learning Summary`

What have found most surprising, significant or unexpected in your text mining investigations? What have your learned in this class? What additional questions do you have?

```
14. Bibliography
```

Include your bibliography here.

SUBMISSION

You will submit the iframe URL found in 'Share this StoryMap' as your final project submission:

Also include a sentence in your submission confirming whether or not you are agreeable to your StoryMap being displayed on the Emma B. Andrews Diary Project webpage. I'll send you the link to vet before it goes live.

Week 9 Discussion

This is the final worklog of the quarter. Plan to compile a complete record of the work you carried out during the week, which may include ongoing cleaning, curating, and analysis.

We have examined topic modeling and clustering, looked at the new update to the DSL, and walked through the StoryMap framework.

1. Include a summary of reading you completed this week related to topic modeling and clustering. Aim to discuss 2-3 articles or projects.

2. Provide a complete record of your work this week, including processes, triumphs, research insights and questions.

3. Complete the final tweaks to your Project One Pager as you prepare to include it in your final project.

No replies are necessary this week; focus on completing your StoryMaps.

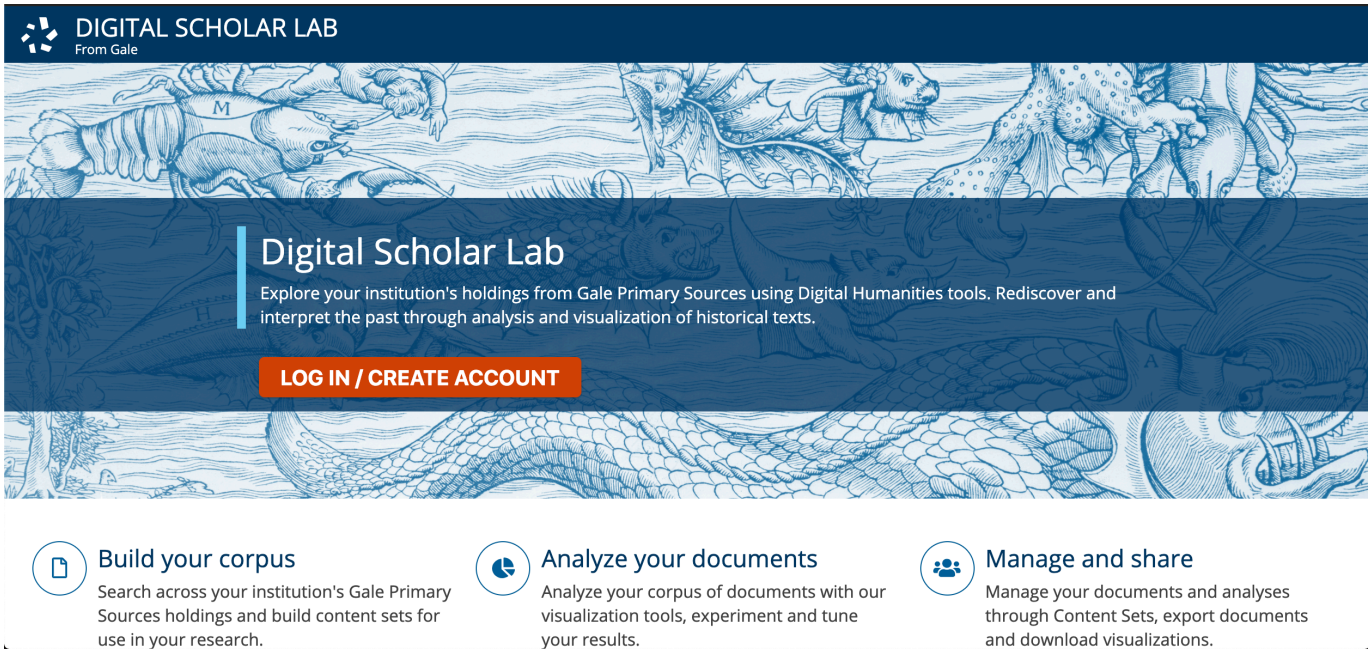
Gale Digital Scholar Lab Reference Guide & Glossary

Class URL: https://infotrac.gale.com/itweb/dslabwa?db=DSLAB

Password: uwash

Log in with Google or Microsoft credentials

This will take you to your own 'personalized workbench' environment, where you can start to build your content sets for this class based on the research topic that you choose.



Learning Resources

1. Help documentation linked from the footer of the Landing Page.

2. Introductory videos.

- Landing page, workflow, help documentation https://youtu.be/7ZUwwdYp2Y8
- Search and create content set https://youtu.be/wbZcXpEkOE8
- Text cleaning https://youtu.be/nDNyM6KPxcA
- Tool overview https://youtu.be/teYDkMUHGlg
- Topic modeling and sentiment analysis https://youtu.be/ZUxTfrKxVis
- My content sets https://youtu.be/fV3n0ak0pUE

Search Tips

Hyphen

A hyphen (-) used between two words is ignored. However, if you are searching for a word or phrase that normally contains a hyphen, you may include it:

"e-mail"
"dot-com"
Note that hyphens are also range operators for dates.

Apostrophe

Apostrophes should be used when searching contractions. For possessives, the apostrophe may be used in search phrases because the search engine will return results containing the words from the query.

can't
Evolution's Darling
Bush's cabinet

Ampersand
Ampersands may be used. For best results enclose the search term in quotes:

"AT&T"
"M&Ms"

Period
A period (.) used between two words is ignored by the search engine. However, if you are searching for a word or phrase that normally contains a period, you may include the period, as in gale.com.

Capitalization
The search engine is not case sensitive. That is, use of capitalization does not affect the results of a search. For example, the following keyword searches are considered the same:

Plants and animals
PLANTS and AniMAls
plaNts AND animALS

Wildcards
Sometimes you might want to find more than just exact matches to a search term. Wildcards let you substitute symbols for one or more letters.
With wildcards, you can match

- both the singular and plural forms of a word
- words that begin with the same root
- words that can be spelled in different ways
- You can even match words that you're not sure how to spell!

There are three wildcard operators:
***** An asterisk (*) stands for any number of characters, including none, and is especially useful when you want to find all words that share the same root. For example, pigment* matches pigment, pigments, pigmentation, etc. Note that you must enter at least three (3) non-wildcard characters. So, a search on o* is not allowed; rather you need to enter: oba*. An asterisk can also be used within a word, but the other wildcards are more precise for this kind of use.

? A question mark (?) stands for exactly one character and is especially useful when you're uncertain of a spelling. For example, a search like relev?nce means you can match the word relevance even if, like many of us, you can't remember whether it's spelled with ‘ance’ or ‘ence.’

A question mark is also useful for finding certain words with variant spellings. For example, defen?e finds both defense (American) and defence (British and Canadian). Multiple question marks in a row stand for the same number of characters as there are question marks. For example, psych????y matches either psychology or psychiatry but not psychotherapy.

! An exclamation point (!) stands for one or no characters and is especially useful when you want to match the singular and plural of a word but not other forms. For example, product! matches product and products but not productive or productivity. The exclamation point can also be used inside a word to match certain variant spellings. For example, colo!r matches both color (American) and colour (British). If you see a message about a search being invalid, try adding more letters before the wildcard character.

Logical Operators

Logical operators create relationships between search terms, between a term and a result set and between two result sets. They allow you to find the result of the intersection of two search terms or result sets, the combination of two terms or result sets, or the exclusion of a term or result set from a search.

There are three logical operators:

and The *and* operator specifies that both words on either side of the operator must occur in the part of a record you're searching for that record to match. For example, alcohol and pregnancy finds only those records in which both the word alcohol and the word pregnancy occur.

or The *or* operator specifies that one or the other or both words on either side of the operator must occur in the part of a record you're searching for that record to match. For example, dreams or daydreams finds records in which either the word dreams or the word daydreams or both occur.

not The *not* operator specifies that the word before the operator must occur but the word after the operator must not occur for a record to match. For example, crime not murder finds all records in which the word crime occurs except the ones in which the word murder also occurs.

Logical operators in a search expression are evaluated in a particular order:

not and and
or

If you want to change the order of evaluation, use the nesting operators.

Nesting Operators

The search system follows a particular order of evaluation when there are two or more operators in a search expression. First, wildcards are evaluated. Next come proximity operators, which are tightly bound to the words on either side of them. Finally, the logical operators are evaluated: first not and and, followed by or.

You can change the evaluation order of the logical operators by using nesting operators (parentheses). When you nest entries, the search system performs the operation within parentheses first, then merges the result with the part of the entry outside the parentheses.

Examples

The search expression race or color and discrimination specifies that you want to find records that contain either the word race or both the words color and discrimination. This expression is equivalent to the expression race or (color and discrimination).

The search expression (race or color) and discrimination specifies that you want to find records that contain either or both of the words race or color and that also contain the word discrimination.

Proximity Operators

Proximity operators are used between two search terms to indicate that the terms must occur in a record within a specified distance of each other for that record to match. Words that are close to each other are more likely to be related than words that are far apart.

A proximity operator has two components:

- A letter that indicates the direction
- A number that indicates the distance in words

The proximity operators are:

Nn The N (near) operator specifies that the words on either side of the operator must occur within n words of each other in either direction for a record to match. For example, the search expression memory n5 repressed matches any records in which the words memory and repressed occur within five or fewer words of each other in either direction.

You can use the proximity operator only when searching indexes made up of individual words, such as a title index. They are most useful in indexes of large areas of text, such as keyword and full-text indexes.

Note that proximity operators can be used only between two words, not between a word and an expression within nesting operators (parentheses):

Invalid expression: fleas n10 (dogs or cats)
Valid alternative: fleas n10 dogs or fleas n10 cats

Quotation Marks
Enclosing your search terms in quotation marks yields results in which the words appear in the specified order adjacent to one another. This may be helpful for keyword and full text (entire document) searches, especially when you are searching for an exact phrase. For example, a search on "Wild Bill" is the same as searching wild W1 bill (using the W proximity operator). That is, the word wild must be followed by the word bill, in that order, with no other words in between.

If the phrase contains the word or not, and you want those words used literally, not as logical operators, then you must enclose your phrase in quotation marks. For example, if you typed sink or swim, the word or would be treated as a logical operator. However, enclose the phrase in quotation marks as: "sink or swim" and the system will search for those three words together, in the order listed.

For database collections that include Subject Guide Search and/or Publication Search: Note that Subject Guide Search and Publication Search ignore quotation marks.

Advanced Search Options

See a 21 page listing in the separate PDF.

Plaintext to XML-TEI using the Historical Markup Tool

The purpose of marking up our transcribed text documents into XML-TEI is to create a machine-readable edition that can be displayed on our website. We 'tag' specific topics of interest in our historical texts, including people's names, place names, hotels, boats and events. Capturing this information serves two purposes: it enables us to connect this information across the many documents in our digital archive, which in turn provides a template to guide our historical research: for example, people's names which occur frequently across the corpus of material are those which we target for our biographical research and Emmapedia database.

This set of instructions will guide you through the process of encoding the Helen Winlock archive of letters.

Historical Markup Tool

During the 2018-2019 school year, Audrey Holmes created a 'historical markup tool' for her Masters thesis in Computational Linguistics. It will save a considerable amount of work, since it encodes our plain text documents automatically. It will also prevent errors, which are easily made when coding by hand.

The tool is available on a standalone website: http://historical-markup.com/

and is also embedded in the Emma B. Andrews project website: http://www.emmabandrews.org/project/historical-markup-tool

---

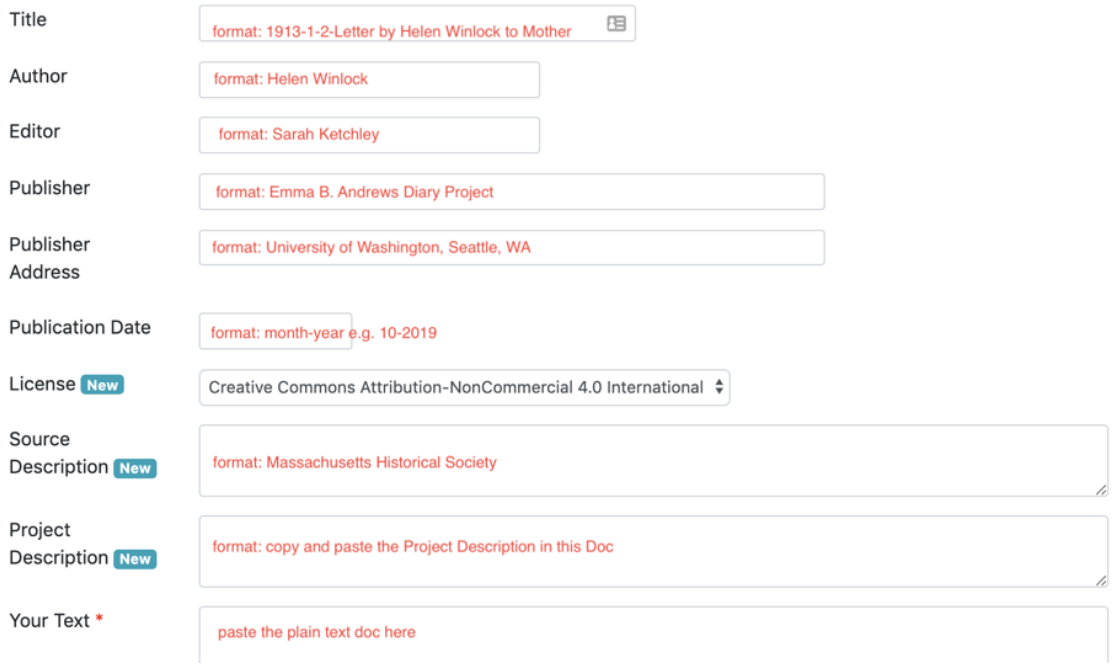
How To

Choose one of the completed plain text files from the Helen Winlock archive, here.

https://www.dropbox.com/sh/67p06yll085csle/AAApOPh_RrcV8lVHbECj8LF1a?dl=0

Open the historical markup tool, and **complete** each field:

Paste this text into the 'Project Description' box:

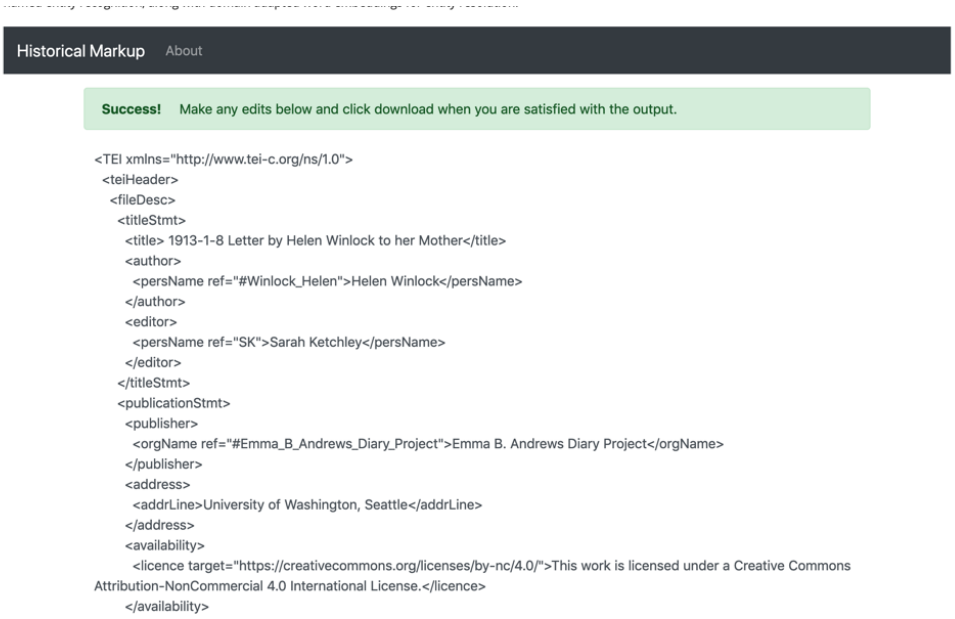


Established in 2010, the goals of the Emma B. Andrews Diary Project include the transcription and digitization of a wide range of primary historical material from the 'Golden Age' of Egyptian archaeology, at the end of the 19th and beginning of the 20th centuries. A founding partners of Newbook Digital Texts (www.newbookdigitaltexts.org), the EBA Diary Project offers undergraduate and graduate digital humanities education and internships at the University of Washington.

Emma B. Andrews is best remembered for her association with the millionaire lawyer turned archaeologist/art and antiquities collector, Theodore M. Davis. Traveling to Egypt with him between 1889 and 1912, she kept detailed journals of these voyages along the Nile, including his important yet under-reported excavations of over 20 significant tombs in the Valley of the Kings. Emma provides a vital commentary on the archaeology and pioneering Egyptologists of the time, painting a revealing picture of the lives of the colonial gentry and the cultural and scientific literati in at the dawn of the twentieth century. Analysis of the content of her diaries, along with a broad range of additional primary source material, will afford scholars information about important historical resources for the first time.

Paste the plain text for the letter you're working on into the "Your Text" field.

Click **submit**. You'll get this message once the tool has completed its run:



Carefully check the output, looking at the people and place names in particular.

Replace any page numbers in the plain text with the following XML:

<pb n="2"/>

(where "2" is the relevant page number)

EXAMPLE:

Download the output, and name it following the convention of the original document. It will end in .xml

Put the XML file into this Dropbox folder:

```
<div>
<p>THE <placeName>UPPER EGYPT</placeName> HOTELS COY.
ETC<placeName>LUXOR</placeName> WINTER PALACE LUXOR <date
when="1913-01-02"> January 2, 1913</date></p></div>

<div><p>Dearest <persName>Mother</persName> and
<persName>Father</persName> and <persName>Sister</persName> and
<persName>Brother</persName>,</p>
<p>Here we are once more in <placeName>Luxor</placeName> after much delay.
Got up here a couple of days or so ago and I have hardly been out of the hotel
grounds since and haven't seen the house <name>El Assassif</name> its called
meaning the Holes' or 'Ditches' being descriptive of the ground round about. I am
not sure whether that name appears on the note paper or not, <persName>Mr
Carter</persName> across the way calls his plan "the Hill of Flies" in the language
of course and that heads his paper so we must not be behind. I am much better
every day I can feel the change, but I haven't</p>
<pb n="2"/>
<p>dared ride so far and as all the furniture hasn't come it would mean riding back
again. <persName>Herbert</persName> of course has been over and has stabled
the two horses, the dog, the guard and the stable boy in the dining room tent right
across the river, everything handy, and now he has a cold and has been in bed for
the day, mainly to get well quickly as <name>J.P.</name> starts on the 5th and
nothing has been done to get ready for him. <persName>Herbert</persName> will
be alright tomorrow though sure. Our opinion is castor oil but the doctor hasn't
done anything of the sort said it didn't matter. However perhaps he knows. We had
a hectic day yesterday as we had a front room which was cold as the tomb which
the poor boy couldn't stand so we up and </p>
<pb n="3"/>
<p>shifted to the top floor back and the difference is enormous, though there isn't
so much room for trunks of which we seem to have enormous quantities. I trimmed
my tough looking black straw hat that you hate with the cord out of pajamas that
<persName>Mamie</persName> donated to tie my wrapper together with. It has
tassels and goes round three times and looks very cute. Then the green veil goes
round everything and hangs down behind & I take a swell walking stick a pair of
```

https://www.dropbox.com/sh/zcah22mgdnby7dt/AABjGCsUq__8jnNJVBIWniSaa?dl=0

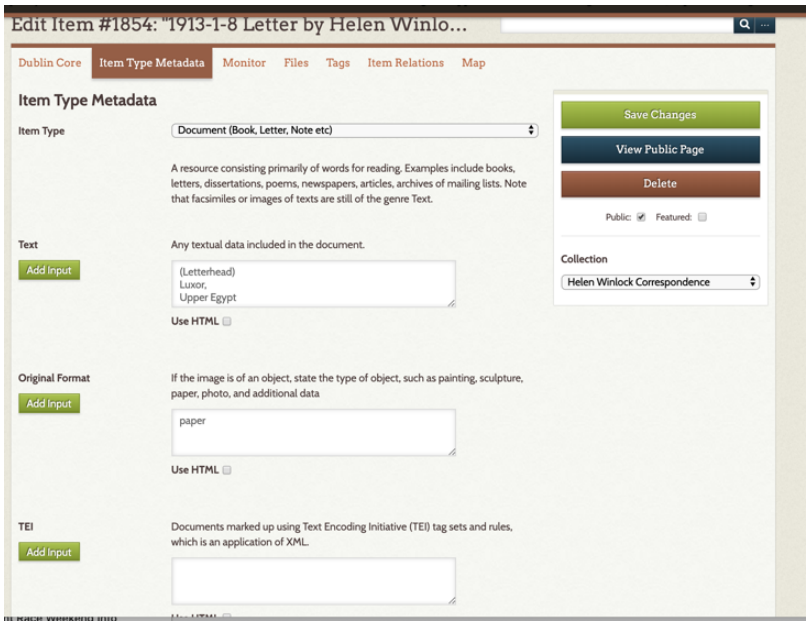
Finally, add a checkmark to this spreadsheet, in the ‘Markup Tool’ column, when you have finished.

https://docs.google.com/spreadsheets/d/1wNpBr3PDsbxzqJH3ImpPSzd2YqdHNYJKNjHbWMdKdmQ/edit?usp=sharing

What happens when I’ve finished?

1. The XML will be checked/edited if necessary.

2.It will be uploaded to our website in Omeka, along with the jpeg files and plain text transcription.



All three formats will be available to our website visitors by clicking on the relevant letter file.