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.

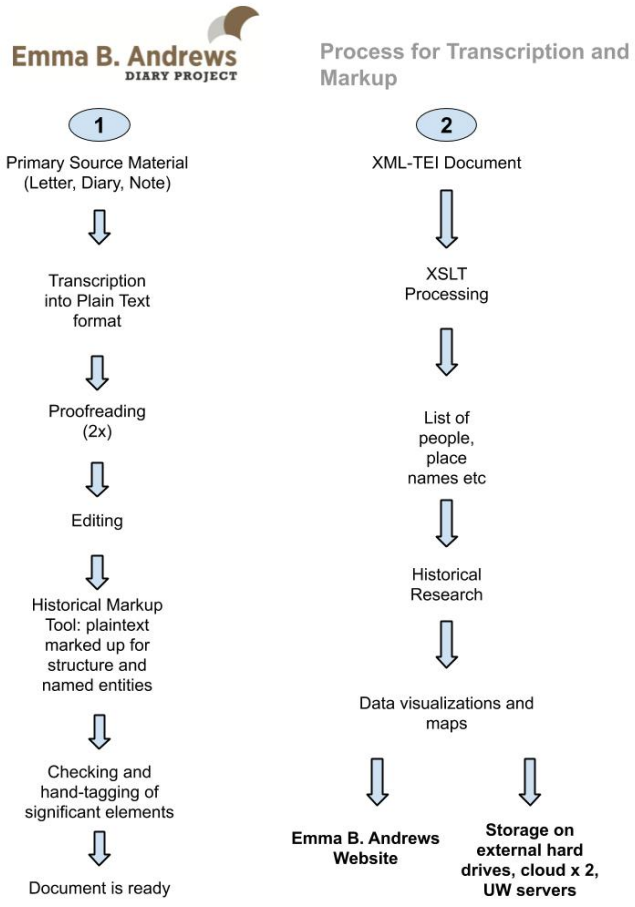
Emma B. Andrews Diary Project Transcription & Editing Process

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.
- 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

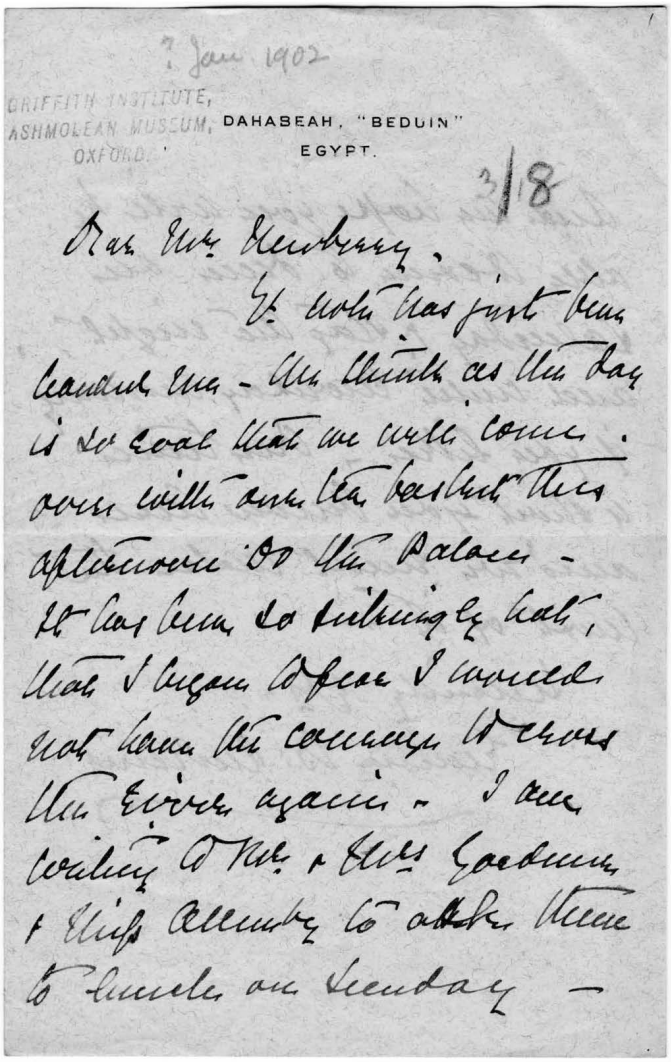
letterhead

sic (obvious typos by original transcriber or source misspelling)

notes added by a writer other than the author

TASK #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

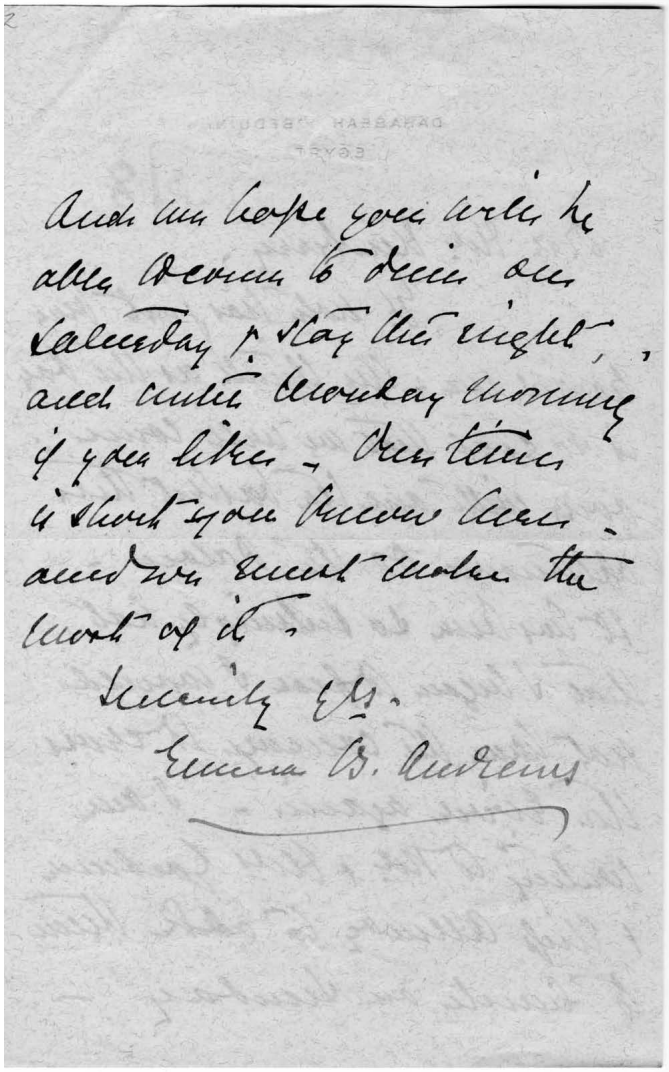
GRIFFITH INSTITUTE,
ASHMOLEAN MUSEUM,
OXFORD.

DAHABEAH, "BEDUIN"
EGYPT.

3/18

Dear Mr. Newberry.

Your note has just been handed me - We think 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 would 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, & with us Monday morning if you like - Then I shall see you before long - 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.

Readings

- 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 (in class) -

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?
```