Getting Started with Text Mining Archival Collections

Julia Frankosky, Michigan State University
Megan Badgley Malone, Michigan State University
Midwest Archives Conference, April 6, 2017

Learning Outcomes

After today you should be able to

- Understand what is text mining/text analysis
- Explain the research value of text mining/analysis
- Use Voyant and Overview, two freely available textual analysis tools
- Prepare your own documents for digitization and textual analysis projects

Julia Frankosky is the Government Information Librarian and Video Game Cataloger at Michigan State University. She has provided research and collection support for federal, Canadian, and international government publications since 2013. From 2011 to 2013, Julia worked as the Copyright Librarian at Michigan State University.

Julia has a Masters in Library and Information Science from Wayne State University with a graduate certificate in Records and Information Management. She received her B.A. in History from Michigan State University.

Megan Badgley Malone is the Collections and Outreach Archivist at the Michigan State University Archives & Historical Collections. Megan coordinates tours, teaches introduction to archives sessions, manages social media, answers reference inquires, and arranges and describes archival collections. She has worked at the University Archives since June of 2011.

Megan earned a Master of Library and Information Science degree with a graduate certificate in Archival Administration from Wayne State University. She also has a Bachelor of Arts in Secondary Education and History from Saginaw Valley State University.

Introduce Yourself

- Your name
- Institution
- Job title
- Why are you interested in text mining?

Text mining/text analysis

- These terms are often used interchangeably and for the most part they have very similar connotations and overlap, but there are some slight differences.
- Text Analysis involves looking at the document as a whole and creating things like word clouds and other similar visualizations to show word frequency, for example
- Text mining is more like finding a needle in a haystack; you have a large amount of text and want to find when and how a particular word or set of words are used

Benefits of text mining/text analysis

- Gain a better understanding of large documents
- Discover trends across a text that you may have failed to piece together
- See changes over time
- Find areas where additional research is necessary to gain a better understanding

Text Mining Projects

Flowers to Emma: an American Civil War Love Story

- Created by 3 MSU students for a DH class project
- MSU Archives Civil War Collections website
- Used Voyant for text analysis
- Used Neatline plugin in Omeka to create map

Mining the White House Project

- Using Voyant Tools to text mine inaugural addresses
- Online classroom resource
- Students can see:
  - Frequency of words & relative use to other words
  - Chronological Trends
  - Shifts in trends (e.g., war, peace)
- Small data set
- Good intro to text mining
- Encourages students to analyze data, develop research questions

Mining *the Dispatch*

- Exploring Civil War era Richmond, Virginia by text mining the *Daily Dispatch* (1860-1865)
- Used MALLET, a topic modeling program developed at University of Massachusetts, Amherst
- Fugitive Slave Ads
- Identifying clusters of words that often appear in the same document together

Nabokov's Favorite Word is Mauve

- Book by statistician Ben Blatt
- Thousands of classics & best selling books mined to research the craft of writing
- Do authors follow conventions regarding use of clichés, adverbs, exclamation points?
- Do men & women write differently?
- Can algorithms identify writers by prose style?
- Who mentions weather most in opening sentences?
- What are authors favorite words?

Text Mining in the Classroom

The Digital Middle Ages: Stanford University

"Medieval Studies is entering a phase of digital abundance. In the last five years, more medieval material has been put online than has ever been available for study at any point in the past. How can we engage with the growing mass of digitized material available to us? How does this sudden access impact the work we do, the types of questions we ask, the connections we make, and the audiences we write for?

"...This course examines and evaluates digital medieval resources and software that has been created for interacting with those resources. Students will have the opportunity to design and create an innovative project based on medieval primary sources held at Stanford, applying current digital methods in the analysis of those resources." http://blalbrit.github.io/courses/dlcl122_2016/

Sample Assignment

- Choose a small textual corpus related to the materials for the final project
- Use Voyant begin to analyze this corpus of texts
- Reflect on limitations and possibilities of this tool and this approach to analysis as they relate to the final project
- Present your analysis in a blog post following class discussion

Mining Black Culture: Georgia Regents University

“using technology to recover lost and under studied texts, to conduct textual analyses that link African American texts to African diasporic and continental texts, and to discover themes and linguistic patterns or commonalities. Not only does a technology such as text mining allow researchers to add data to existing qualitative studies, but it also generates new areas of scholarship because the software can reveal patterns and discontinuities within a body of literature previously unseen.”

https://miningblackculture.wordpress.com/about/syllabus-for-curs-4990/

Sample Assignment

- Mine, map, and analyze word patterns and occurrences in the poetry of Langston Hughes, Margaret Walker, Gwendolyn Brooks, Maya Angelou, and Alice Walker.
- Create a word database for the collected poems of five major American authors and make that database available and searchable for the public.
  - Goal: Provide a tool for scholars to produce new scholarship on African American poets, promote the inclusion of African American literature in the emerging field of digital humanities, and involve Georgia Regents University in the national and international conversations concerning digital humanities and text mining.

Roman Archaeology for Historians: Carleton College

- Introduce and explore key concepts in Roman archaeology
- Explore the tensions between historical and archaeological ways of knowing about the past
- Express history and archaeology in a way that takes advantage of the key affordances of digital media
  - “Digital media has transformed the way we learn about the world, and so I want students who complete this course to be able to develop some media literacy to express history/archaeology this way”

http://carleton.ca/grs/wp-content/uploads/CLCV3202a-HIST3101a-syllabus-Aug-21-Shawn-Graham.pdf

Sample Assignment

Use Voyant on 5 archaeological and 5 classical academic papers and be prepared to discuss the disciplinary differences in the language used to talk about the past.

Voyant-tools.org

Acceptable Text Formats

- Plain text
- HTML
  - Can add URLs to text you'd like to analyze
- XML
- PDF
- RTF
- Microsoft Word

Easy to share what you discover in your texts

- Create direct links
- Get embed code
- Export as a PNG or SVG image
- With links and embed code, those viewing can interact with the visualizations

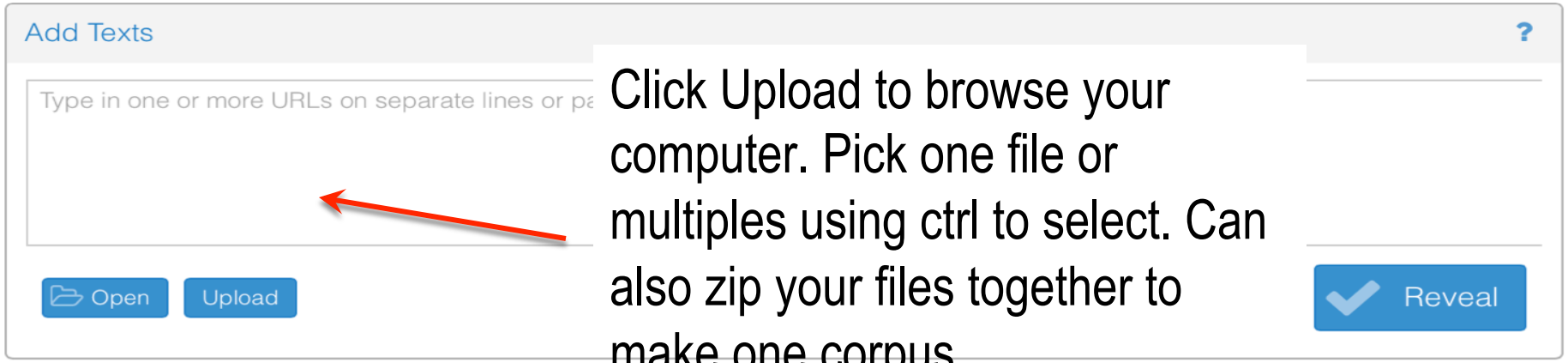
Limitations

- Using large texts or multiple texts can obscure some trends
- Also can take a long time for Voyant to process and for any modifications you make to take effect
- If using PDFs, the analysis is only as good as the optical character recognition (OCR)

Using Voyant

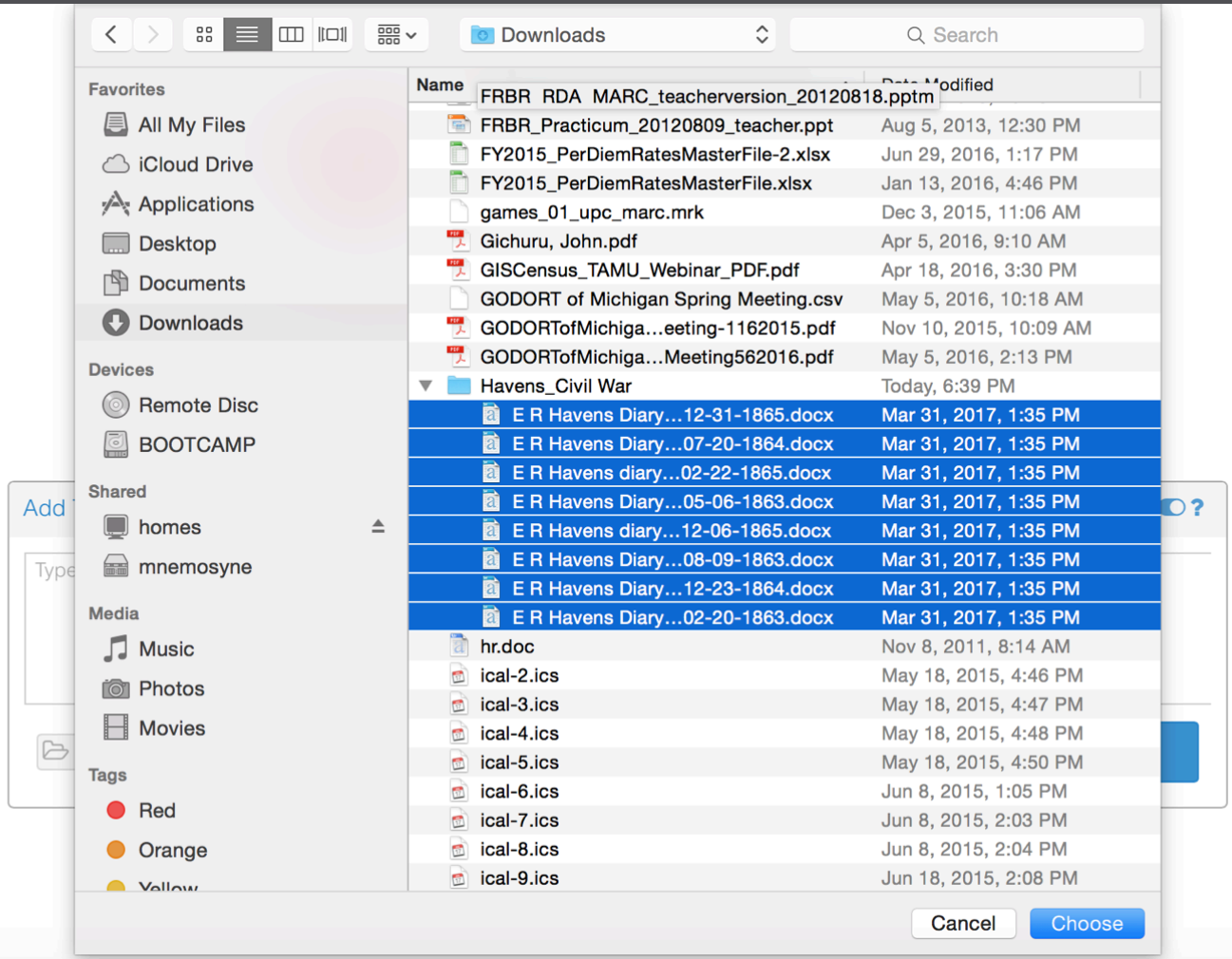
VOYANT

see through your text



Click Upload to browse your computer. Pick one file or multiples using ctrl to select. Can also zip your files together to make one corpus.

Voyant Tools is a web-based reading and analysis environment for digital texts – please visit Hermeneuti.ca for more information..


Downloads
Search
Favorites
All My Files
iCloud Drive
Applications
Desktop
Documents
Downloads
Devices
Remote Disc
BOOTCAMP
Shared
homes
mnemosyne
Media
Music
Photos
Movies
Tags
Red
Orange
Name
Date Modified
FRBR_RDA_MARC_teacherversion_20120818.pptm
FRBR_Practicum_20120809_teacher.ppt Aug 5, 2013, 12:30 PM
FY2015_PerDiemRatesMasterFile-2.xlsx Jun 29, 2016, 1:17 PM
FY2015_PerDiemRatesMasterFile.xlsx Jan 13, 2016, 4:46 PM
games_01_upc_marc.mrk Dec 3, 2015, 11:06 AM
Gichuru, John.pdf Apr 5, 2016, 9:10 AM
GISCensus_TAMU_Webinar_PDF.pdf Apr 18, 2016, 3:30 PM
GODORT of Michigan Spring Meeting.csv May 5, 2016, 10:18 AM
GODORTofMichiga...eeting-1162015.pdf Nov 10, 2015, 10:09 AM
GODORTofMichiga...Meeting562016.pdf May 5, 2016, 2:13 PM
Havens_Civil War Today, 6:39 PM
E R Havens Diary...12-31-1865.docx Mar 31, 2017, 1:35 PM
E R Havens Diary...07-20-1864.docx Mar 31, 2017, 1:35 PM
E R Havens diary...02-22-1865.docx Mar 31, 2017, 1:35 PM
E R Havens Diary...05-06-1863.docx Mar 31, 2017, 1:35 PM
E R Havens diary...12-06-1865.docx Mar 31, 2017, 1:35 PM
E R Havens Diary...08-09-1863.docx Mar 31, 2017, 1:35 PM
E R Havens Diary...12-23-1864.docx Mar 31, 2017, 1:35 PM
E R Havens Diary...02-20-1863.docx Mar 31, 2017, 1:35 PM
hr.doc Nov 8, 2011, 8:14 AM
ical-2.ics May 18, 2015, 4:46 PM
ical-3.ics May 18, 2015, 4:47 PM
ical-4.ics May 18, 2015, 4:48 PM
ical-5.ics May 18, 2015, 4:50 PM
ical-6.ics Jun 8, 2015, 1:05 PM
ical-7.ics Jun 8, 2015, 2:03 PM
ical-8.ics Jun 8, 2015, 2:04 PM
ical-9.ics Jun 18, 2015, 2:08 PM
Cancel
Choose

Voyant Tools

Cirrus | Terms | Links

Scale | Terms:

Reader

E R Havens Diary 01-01-1865 through 12-31-1865(mgrace)

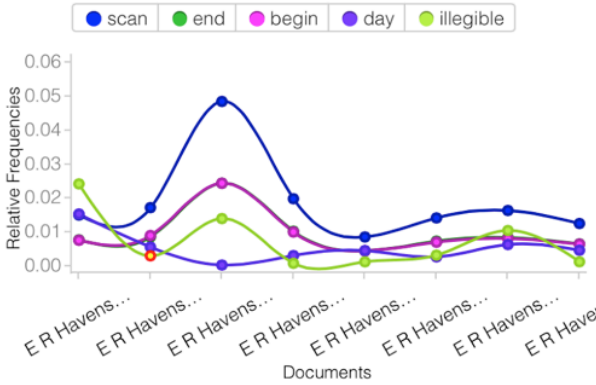
Collection ID: LC00016

Collection Name: Havens, Edwin R. Papers

Box #2, Volume #6

Jan 1, 1865 through Dec 31, 1865

Trends | Document Terms



Scale | Frequenci

Summary | Documents | Phrases

This corpus has 8 documents with 132,839 total words and 8,428 unique word forms. Created now.

Document Length:

- Longest: E R Havens diary 06-24-18… (35831); E R Havens Diary 10-18-18… (28987); E R Havens Diary 02-27-18… (19378); E R Havens Diary 01-01-18… (15916); E R Havens Diary 01-11-18… (12896)
- Shortest: E R Havens diary 02-01-18… (3075); E R Havens Diary 07-28-18… (6790); E R Havens Diary 07-06-18… (9966); E R Havens Diary 01-11-18… (12896); E R Havens Diary 01-01-18… (15916)

items:

Contexts | Bubblelines

		Left		Right
⊞	1)	1865 Document type: D [Begin	sc	1] January, Sunday, 1. 1865
⊞	1)	of interest in camp. [End	sc	1] [Begin Scan 2] January
⊞	1)	camp. [End Scan 1] [Begin	sc	2] January, Wednesday, 4. 1865
⊞	1)	and late in evening. [End	sc	2] [Begin Scan 3] January
⊞	1)	evening. [End Scan 2] [Begin	sc	3] January, Saturday, 7. 1865

1,875 context | expand | Scale

Cirrus Word Cloud

- Creates a wordcloud view of the most frequently occurring words in either the corpus or the document
- Hovering over the word shows the term frequency
- Clicking on the term may produce results in other tools


scan
begin
morning
original
end
day
illegible
camp
capt
went
night
men
miles
lieut
came
time
today
yesterday
horses
quite
near
left
1865
good
received
soon
p.m
col
afternoon
warm
house
tuesday
train
company
sunday
arrived
sent
evening
a.m
1st
monday
took
march
got
little
sergt
gen
cold
city
passed

Terms

- Terms is a table view of terms that appear in the entire corpus
- Can reorder by term and count
- Displays a sparkline graph of the term frequency trends across the corpus or the document
- Has additional columns such as relative frequency, distribution peakedness and skew
- Has a search box

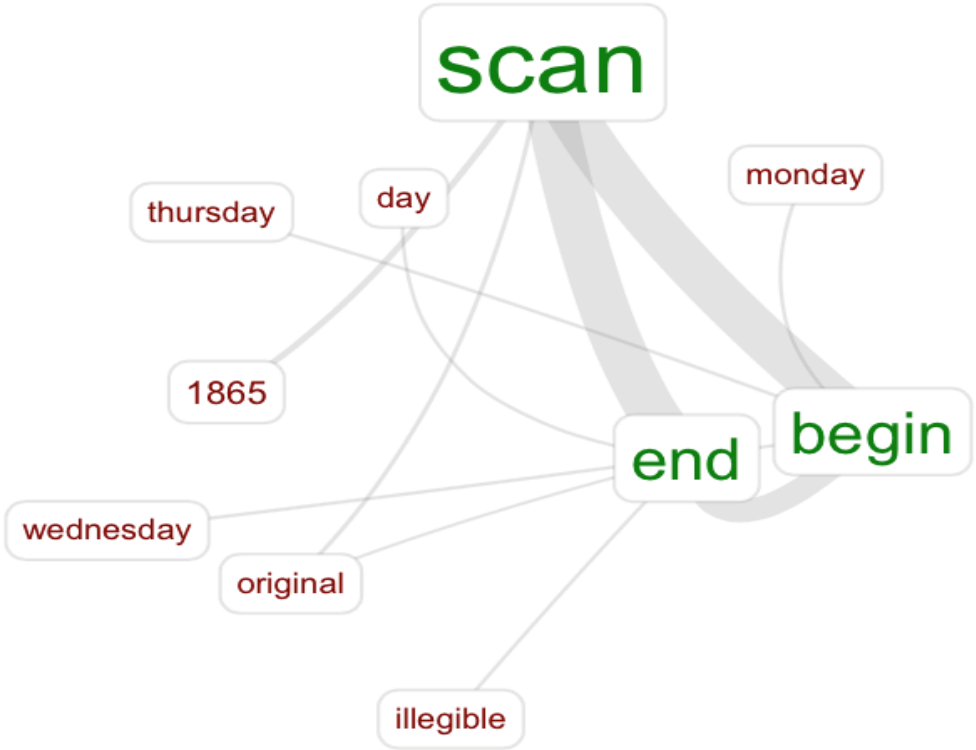
Terms

		Term	Trend
☐	1	scan	18
☐	2	end	94
☐	3	begin	93
☐	4	day	70
☐	5	illegible	62
☐	6	original	62
☐	7	morning	62
☐	8	camp	53
☐	9	night	46
☐	10	went	40
☐	11	capt	40
☐	12	men	38
☐	13	miles	31
☐	14	came	29
☐	15	today	25
☐	16	lieut	25
☐	17	1865	25
☐	18	time	24
☐	19	left	24
☐	20	yesterday	24

8,046

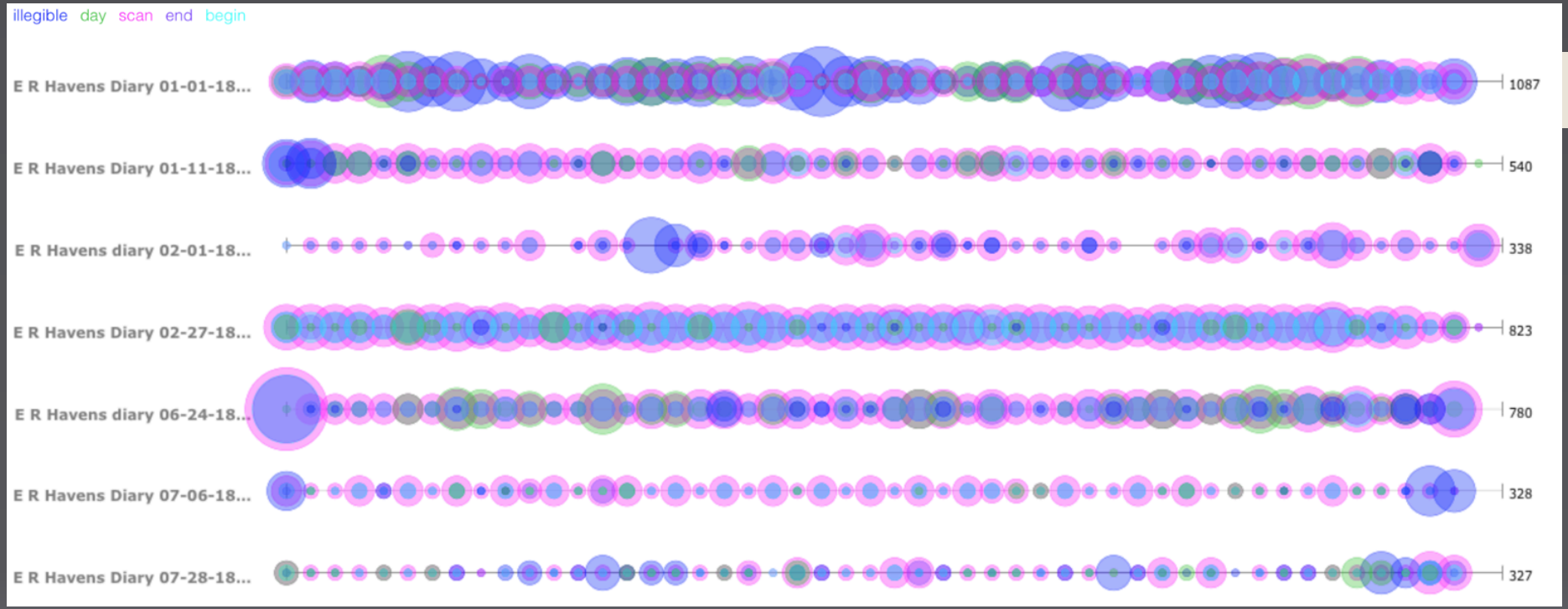
Links

- Shows a network graph of higher frequency terms that appear in proximity
- Keywords are blue and collocates (words in proximity) are orange
- Hovering over a keyword shows their frequency
- Hovering over collocates shows their frequncy in proximity (not their total frequency)
- Double clicking on any word fetches more results
- Has a search box


scan
monday
day
thursday
1865
begin
end
wednesday
original
illegible

Bubblelines

- Visualizes the frequency and repetition of a term's use in a corpus
  - Each document in the corpus is represented as a horizontal line and divided into equal-sized segments.
  - Each term is shown as a bubble and the size of the bubble indicates its frequency.


illegible
day
scan
end
begin
E R Havens Diary 01-01-18...
1087
E R Havens Diary 01-11-18...
540
E R Havens diary 02-01-18...
338
E R Havens Diary 02-27-18...
823
E R Havens diary 06-24-18...
780
E R Havens Diary 07-06-18...
328
E R Havens Diary 07-28-18...
327

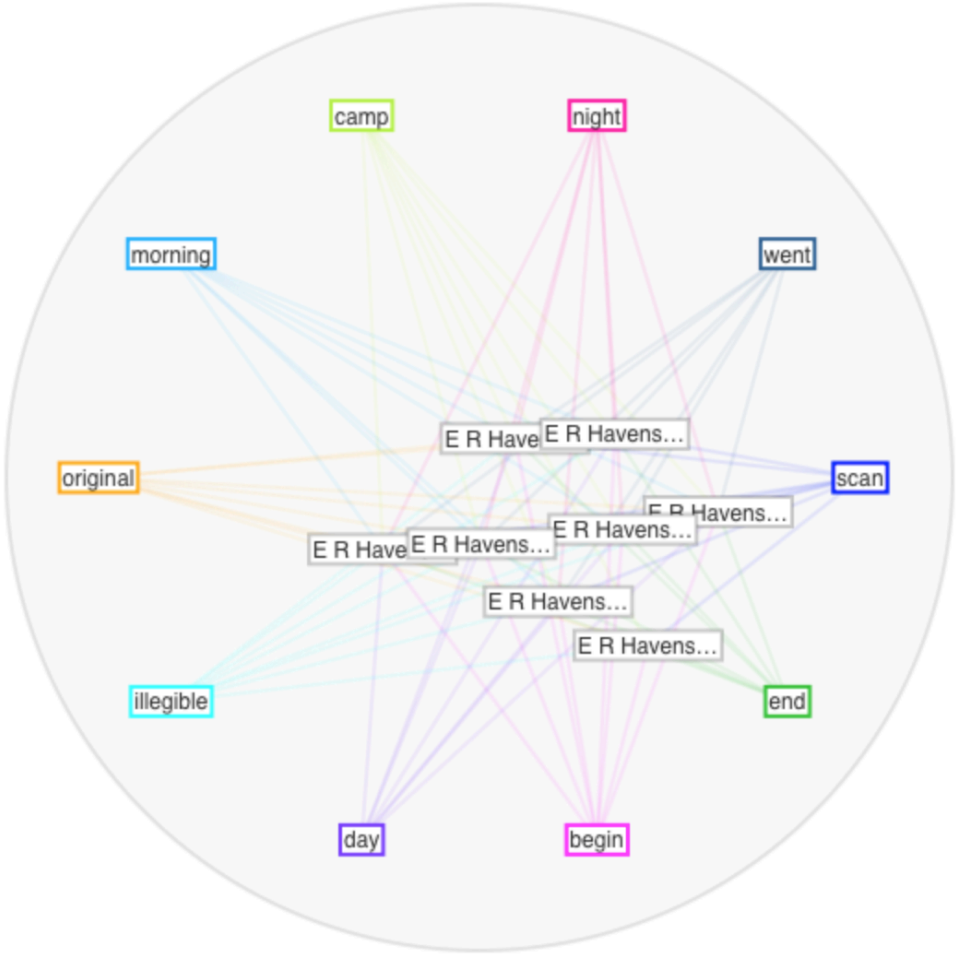
Collocates

- Table view of which terms appear more frequently in proximity to keywords across the corpus
- Can reorder by keyword, collocate word, collocate word count
- Search box

☐	night	467	scan
☐	day	704	clear
☐	scan	1875	tuesday
☐	scan	1875	thursday
☐	day	704	pleasant
☑	morning	620	went
☐	camp	534	went
☐	went	404	camp
☐	scan	1875	wednesday
☐	scan	1875	day
☐	scan	1875	camp
☐	day	704	scan
☐	camp	534	scan
☐	day	704	illegible
☐	went	404	morning

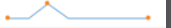
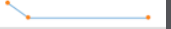
Mandala

- Conceptual visualization that shows the relationship between terms and documents
- Each term pulls documents toward it based on the term's relative frequency


camp
night
morning
went
E R Have
E R Havens…
original
scan
E R Havens…
E R Havens…
E R Have
E R Havens…
E R Havens…
E R Havens…
illegible
end
day
begin

Phrases

- A table view of repeating phrases in the entire corpus

Term			Trend
edwin r havens co a 7th mich cav 1st brig 1st division	2	12	
reveillie sounded at daylight and we took up the march at 6	2	12	
receipts from capt jast benedict c.s city point va table follows	2	11	
we again took up our march and went as far as	2	11	
a large amount of mail and several passengers among them	2	10	
and go into camp at 10 a.m with orders to	2	10	
commissioned and non commissioned officers were ordered to be present	2	10	
illegible in original written sideways on left side of feb	2	10	
parsons cs feb 10 by scholarship 30.00 12 sundries 11.40	2	10	
1865 reveillie at 3 a.m move at 5 30	2	9	
brigade 1st cavalry division dr table follows end scan	3	9	
brigade 1st division cav corps table follows end scan	2	9	
buy all the liquor from them that they could	2	9	
front cover edwin r havens co a 7th mich	2	9	
issues to third cavalry division table follows end scan	2	9	
we were ordered out to practice target shooting with	2	9	
at day light marched 28 miles and camped	2	8	
brigade 1st cavalry division table follows end scan	3	8	
e r havens cash a c 1867 jany	2	8	
empty pages precede this entry in his diary	3	8	

? 5,730 Length Overlap

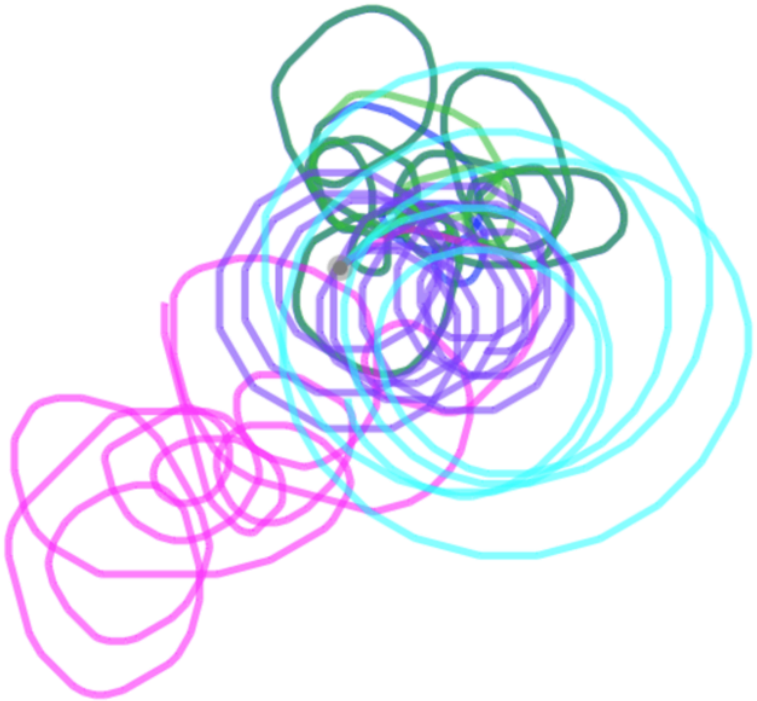
Knots

- A visualization that represents terms in a single document as a series of twisted lines. Each occurrence of a term is represented by a bend in the line, so the more twisted a line, the more a term repeats and straight stretches represent no occurrences.

By default Knots represents the most common terms in the first document of a corpus.

Knots

illegible original day scan end



Cl... Documents Speed: Start: Turn: sound

Reader

- Provides a view of the text
- Frequency information appears when hovering over a word
- Distribution information appears in a graph at the bottom when clicking on a word
- A bar graph at the bottom indicates the relative size of each document in the corpus
- Search box

side of Feb 21 entry over top of entry]

Three times three

were given hats tossed

[remainder of entry illegible in original]

Wednesday, 22.

Morning fair and pleasant

Afternoon lowry and ev.g

rainy. Washington's birthday

and all unnecessary duties

dispensed with no drill. Salutes

fired near Winchester at sunrise

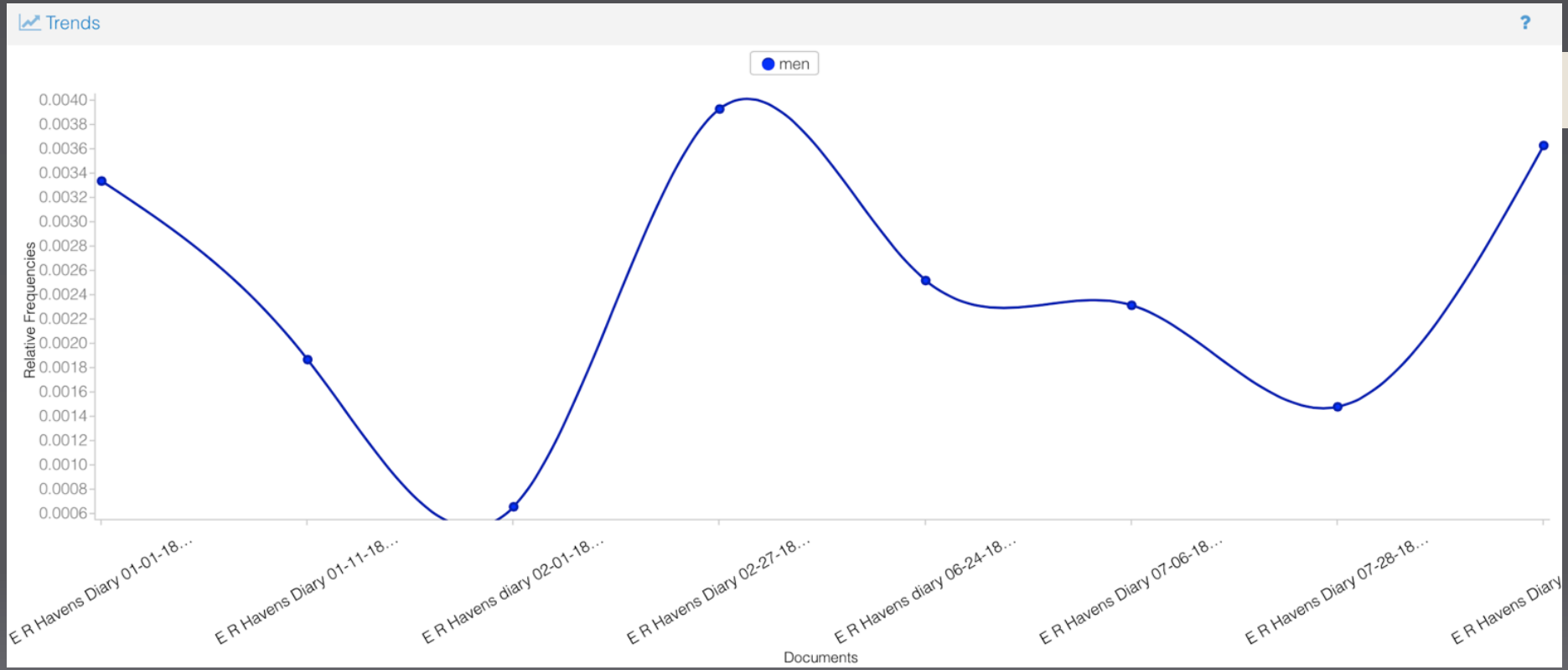
and sunset. Crocker on picket

Preparations, looking like a raid

Thursday, 23.

Trends

- A line graph of the relative frequency across the corpus or within a document


Trends
?
men
0.0040
0.0038
0.0036
0.0034
0.0032
0.0030
0.0028
0.0026
0.0024
0.0022
0.0020
0.0018
0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
Relative Frequencies
E R Havens Diary 01-01-18...
E R Havens Diary 01-11-18...
E R Havens diary 02-01-18...
E R Havens Diary 02-27-18...
E R Havens diary 06-24-18...
E R Havens Diary 07-06-18...
E R Havens Diary 07-28-18...
E R Havens Diary
Documents

Summary

- Provides general information about the corpus
- Total words and word forms
- Most frequently occurring terms
- Documents ordered by length and vocabulary density
- Distinctive words for each document

This corpus has 8 documents with 132,839 total words and 8,428 unique word forms. Created about an hour ago.

Document Length:

- Longest: E R Havens diary 06-24-18… (35831); E R Havens Diary 10-18-18… (28987); E R Havens Diary 02-27-18… (19378); E R Havens Diary 01-01-18… (15916); E R Havens Diary 01-11-18… (12896)
- Shortest: E R Havens diary 02-01-18… (3075); E R Havens Diary 07-28-18… (6790); E R Havens Diary 07-06-18… (9966); E R Havens Diary 01-11-18… (12896); E R Havens Diary 01-01-18… (15916)

Vocabulary Density:

- Highest: E R Havens diary 02-01-18… (0.263); E R Havens Diary 07-28-18… (0.219); E R Havens Diary 07-06-18… (0.186); E R Havens Diary 01-11-18… (0.165); E R Havens Diary 02-27-18… (0.147)
- Lowest: E R Havens diary 06-24-18… (0.116); E R Havens Diary 10-18-18… (0.119); E R Havens Diary 01-01-18… (0.142); E R Havens Diary 02-27-18… (0.147); E R Havens Diary 01-11-18… (0.165)

Average Words Per Sentence:

- Highest: E R Havens Diary 07-06-18… (27.5); E R Havens Diary 02-27-18… (27.0); E R Havens Diary 01-11-18… (25.5); E R Havens Diary 10-18-18… (24.7); E R Havens diary 02-01-18… (24.6)
- Lowest: E R Havens Diary 01-01-18… (13.3); E R Havens diary 06-24-18… (22.3); E R Havens Diary 07-28-18… (23.3); E R Havens diary 02-01-18… (24.6); E R Havens Diary 10-18-18… (24.7)

Most **frequent words** in the corpus: scan (1875); end (946); begin (937); day (704); illegible (627)

Distinctive words (compared to the rest of the corpus):

1. E R Havens Diary 01-01-18…: 1865 (111), entry (25), day (239), coach (32), reveillie (27).
2. E R Havens Diary 01-11-18…: 1864 (21), fredericksburg (12), belle (8), point (40), corps (24).
3. E R Havens diary 02-01-18…: sundries (40), 1.00 (34), 1865 (35), cash (34), feb (27).
4. E R Havens Diary 02-27-18…: drill (41), fairfax (28), april (15), walker (28), centreville (9).
5. E R Havens diary 06-24-18…: coach (106), 1865 (106), big (44), indians (43), o'clock (38).
6. E R Havens Diary 07-06-18…: hagerstown (9), gap (17), williamsport (8), kilpatrick (8), advanced (8).
7. E R Havens Diary 07-28-18…: 1864 (26), winchester (14), martinsburg (10), pr (13), sacks (9).
8. E R Havens Diary 10-18-18…: drill (78), mann (54), parade (33), furloughs (24), kellogg (33).

Documents

- Table view of the documents in the corpus
- Reorder by title, word count, word forms count, and ratio
- Search Box

Documents

	Title	Words	Types	Ratio	Words/Sentence
1	E R Havens Diary 01-01-1865 through 12-31-1865	15,916	2,260	14%	13.3
2	E R Havens Diary 01-11-1864 through 07-20-1864	12,896	2,129	17%	25.5
3	E R Havens diary 02-01-1864 through 02-22-1865	3,075	809	26%	24.6
4	E R Havens Diary 02-27-1863 through 05-06-1863	19,378	2,848	15%	27.0
5	E R Havens diary 06-24-1865 through 12-06-1865	35,831	4,171	12%	22.3
6	E R Havens Diary 07-06-1863 through 08-09-1863	9,966	1,854	19%	27.5
7	E R Havens Diary 07-28-1864 through 12-23-1864	6,790	1,485	22%	23.3
8	E R Havens Diary 10-18-1862 through 02-20-1863	28,987	3,461	12%	24.7

Contexts

- Shows each occurrence of a keyword with a bit of surrounding text (the context).
- Can be used to study how terms are used in different contexts
- Reorder documents by keyword or by left or right context
- Search box

Left		Right
original] [End Scan 17] [Begin	sc	18] February, Tuesday, 21. 1865
illegible in original] D. [End	sc	18] [Begin Scan 19] February
D. [End Scan 18] [Begin	sc	19] February, Friday, 24. 1865
Remainder to Harper's Ferry. [End	sc	19] [Begin Scan 20] February
Ferry. [End Scan 19] [Begin	sc	20] February, Monday, 27. 1865
3 miles beyond Stanton. [End	sc	20] [Begin Scan 21] Fishersville
Stanton. [End Scan 20] [Begin	sc	21] Fishersville March, Thursday, 2
eggs, [illegible in original] [End	sc	21] [Begin Scan 22] March
original] [End Scan 21] [Begin	sc	22] March, Sunday, 5. 1865
Hospital supplies Day beautiful [End	sc	22] [End Scan 23] Went
beautiful [End Scan 22] [End	sc	23] Went to [illegible in
importance occurred during march [End	sc	23] [Begin Scan 24] March
march [End Scan 23] [Begin	sc	24] March, Saturday, 11. 1865
Div ^ Warm and pleasant [End	sc	24] [Begin Scan 25] March
pleasant [End Scan 24] [Begin	sc	25] March, Tuesday, 14. 1865
Roads bad Country poor. [End	sc	25] [Begin Scan 26] March
poor. [End Scan 25] [Begin	sc	26] March. Friday, 17. 1865
to be sent away. [End	sc	26] [Begin Scan 27] March
away. [End Scan 26] [Begin	sc	27] March, Monday, 20. 1865
688?] men on board [End	sc	27] [Begin Scan 28] March

Activity 1: Using Voyant

Share Your Findings

Break

Overviewdocs.com

Overview

Overview helps you make make sense of big disorganized sets of documents (current maximum is 2,000,000 documents per set). It's a visualization and analysis tool designed for *sets* of documents, from dozens to millions of pages of material.

Features

- Built in OCR
- Sophisticated search engine
- Word clouds
- Entity detection
- Topic-based document clustering
- Tagging and metadata support
- A variety of import and export formats
- Open source

Document formats

- PDF
- HTML
- Microsoft Word (.doc and .docx)
- Microsoft PowerPoint (.ppt and .pptx)
- plain text, and also rich text (.rtf)

- Only you can access the documents uploaded to your account, unless you share them.
- Overview currently supports documents in English, Spanish, French, German, Russian, Arabic, Swedish, Dutch, and Romanian.
  - They can add another language in a day or two if you have documents to test with.
- If all of your documents are in PDF form you can simply upload them directly. Note that documents scanned from paper must first be OCR’d to turn their images into searchable text.

Overview’s Main Screen

Divided into four parts:

- the folder tree
- search field
- tag list
- document viewer

You can navigate through the folders in the tree with the arrow keys, or by clicking. Each folder is labelled by the keywords that best describe the documents filed under that folder. The label also tells you if MOST, SOME, or ALL of the documents in that folder contain each keyword. A folder's sub-folders contain, collectively, all of the documents in the parent folder, broken down into increasingly narrow topics.

The document viewer shows either a particular document or a list of selected documents. Each document in the list is summarized by a list of keywords specific to that document.

If you know what you're looking for, enter your query in the "search" box and Overview will show you where documents containing that term appear in the tree.

The tree automatically expands and zooms to follow your selections. Or you can pan it by dragging with the mouse, and zoom using the +/- buttons or the mouse wheel. Folders marked with ⊕ can be expanded to show sub-folders, while ⊖ hides sub-folders.

Tagging

As you explore the folder tree, you'll run across individual documents or entire folders you want to remember. Enter a descriptive tag in the "new tag" field and press "tag." If you're currently viewing a specific document, Overview will tag just that document. If instead you're viewing the list of documents in a folder, Overview will tag the entire folder

Tags and folders have independent lives: each document can have any number of tags applied to it, and the same tag can be applied anywhere in the tree.

Once you've created a tag, you can add that tag to the current document or document list at any time by pressing the + button that appears when your mouse is over the tag name. Or press – to remove the tag.

Clicking on a tag name selects that tag, highlighting the tagged documents in the tree and loading them into the document list.

When you have a lot of documents, it pays to be systematic. Work your way through the folders in the tree from left to right — biggest folders to smallest folders. Select a folder then view a few of the documents in it to see if you understand what they have in common. If specific words appear in MOST or ALL documents in a folder, that's a sign that the folder contains a single meaningful topic. Otherwise there may be more than one important topic in the documents in that folder, so try opening child folders instead until you find a folder where all of the documents are similar. Then tag that folder with a descriptive label.

Use search to find specific documents of interest, but pay attention to which folders contain those documents. You may find other relevant documents in the same folder, even if they don't contain your search term.

As you proceed, you may find documents that talk about similar topics in different folders. Overview doesn't know what you want out of your documents, so it can't always guess how they should be arranged. You can apply a tag to any combination of folders and documents to create a set that is meaningful to you.

You may also discover that the documents in a folder are irrelevant to your work, in which case you can tag them with "read" and simply move on. Part of the power of Overview is being able to decide *not* to look at an entire folder.

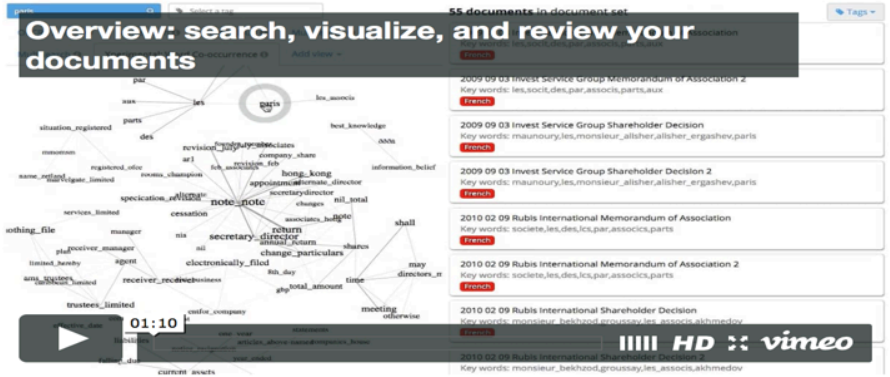
When you're finished this process, you'll have a neatly categorized tree, and a set of tags corresponding to all the interesting topics in your documents.

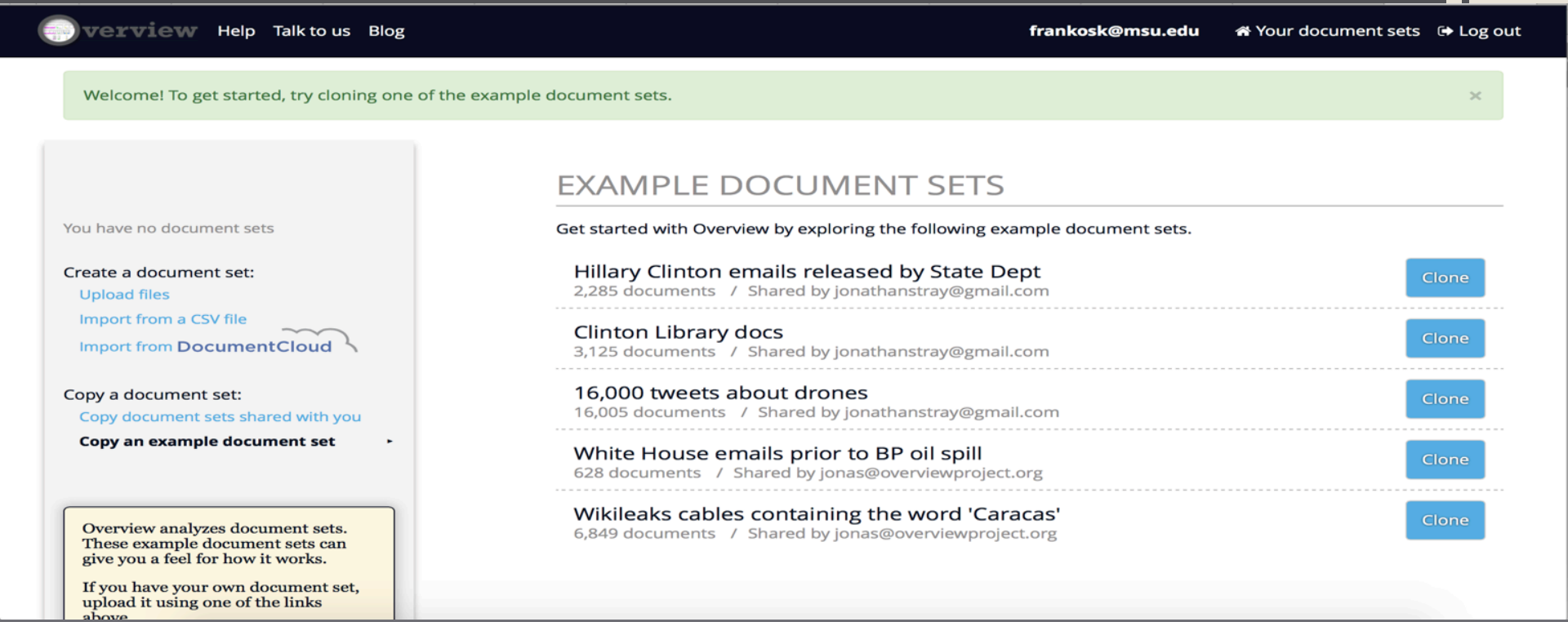
verview Help Contact us Blog

verview

Search, visualize, and review your documents. Up to hundreds of thousands of them, in any format.

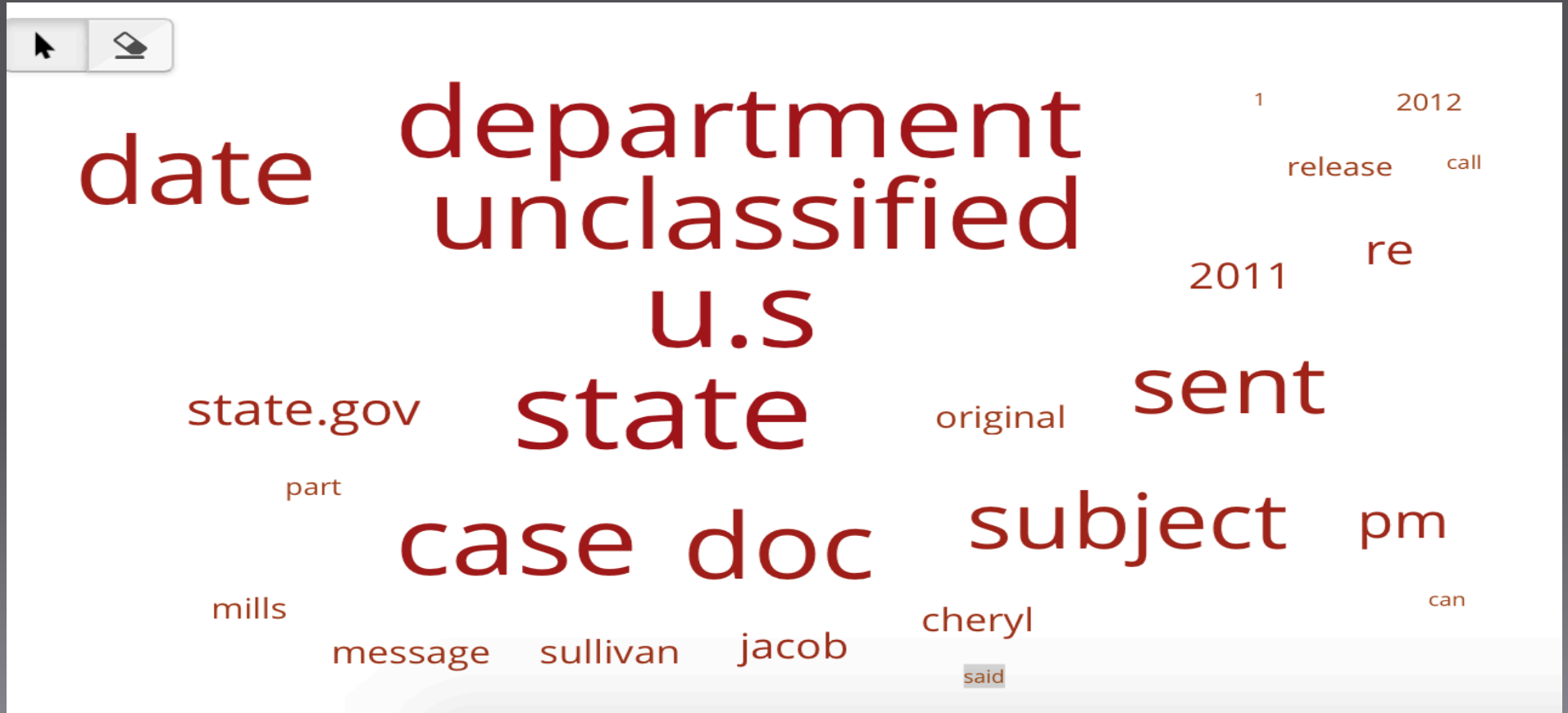
Learn more




verview Help Talk to us Blog
frankosk@msu.edu
Your document sets
Log out
Welcome! To get started, try cloning one of the example document sets.
You have no document sets
Create a document set:
Upload files
Import from a CSV file
Import from DocumentCloud
Copy a document set:
Copy document sets shared with you
Copy an example document set
Overview analyzes document sets. These example document sets can give you a feel for how it works.
If you have your own document set, upload it using one of the links above.
EXAMPLE DOCUMENT SETS
Get started with Overview by exploring the following example document sets.
Hillary Clinton emails released by State Dept
2,285 documents / Shared by jonathanstray@gmail.com
Clone
Clinton Library docs
3,125 documents / Shared by jonathanstray@gmail.com
Clone
16,000 tweets about drones
16,005 documents / Shared by jonathanstray@gmail.com
Clone
White House emails prior to BP oil spill
628 documents / Shared by jonas@overviewproject.org
Clone
Wikileaks cables containing the word 'Caracas'
6,849 documents / Shared by jonas@overviewproject.org
Clone

An account is needed before you can use this free tool.

Word Cloud


date
department
unclassified
u.s
state
state.gov
part
case
doc
mills
message
sullivan
jacob
cheryl
said
original
sent
subject
pm
can
1
2012
release
call
re
2011

Found **2,285 documents**

Tags

HRC emails/February 19/C05758561.pdf

...**UNCLASSIFIED** U.S. Department of State Case No. F-2014-20439 Doc No. C05758561 Date: 02/19/2016...

...BlackBerry Wireless Handheld B6 B6 **UNCLASSIFIED** U.S. Department of State Case No. F-2014-20439 Doc No. C05758561 Date: 02/19/2016...

HRC emails/February 19/C05758626.pdf

...B6 **UNCLASSIFIED** U.S. Department of State Case No. F-2014-20439 Doc No. C05758626 Date: 02/19/2016...

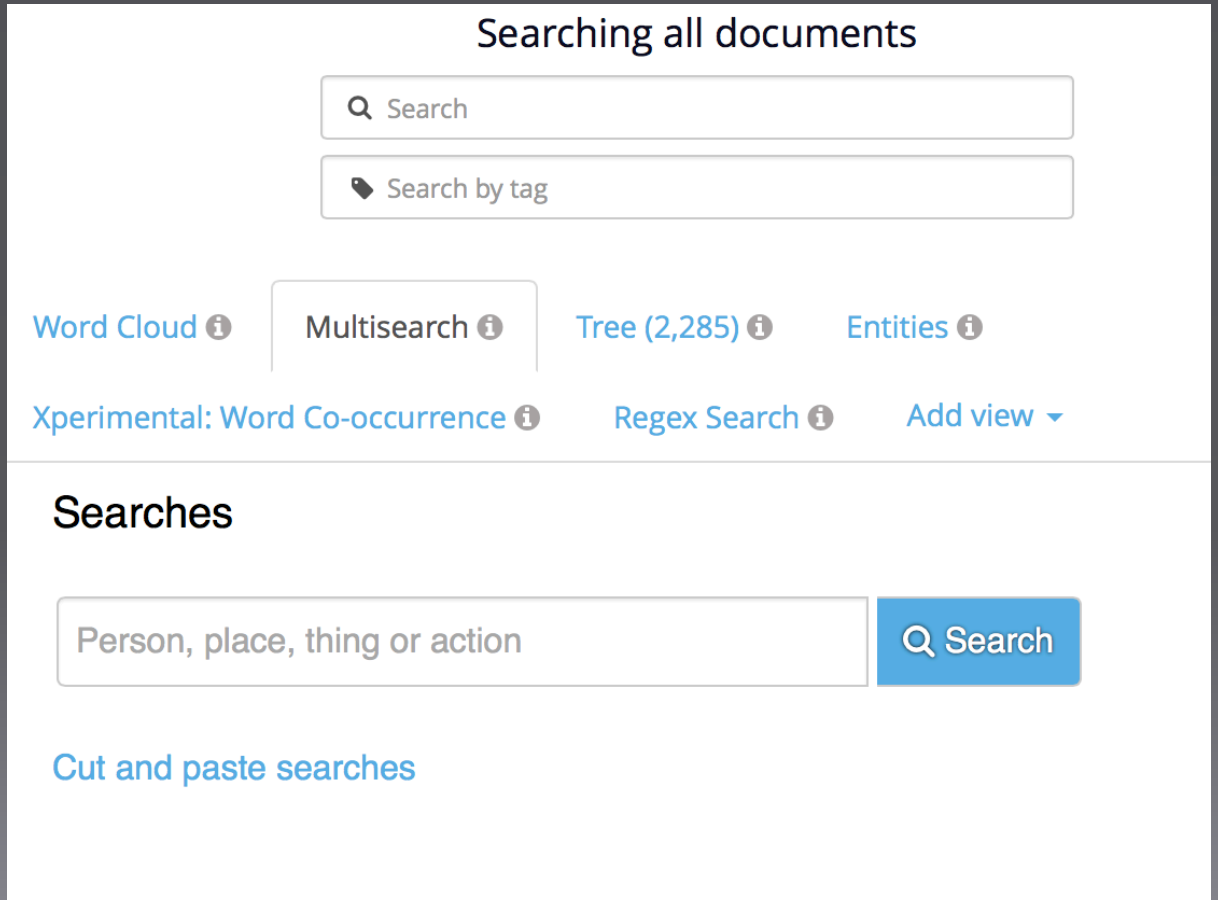
...Subject: one edit to our saturday agreement Thanks, Denis. **UNCLASSIFIED** U.S. Department of State...

HRC emails/February 19/C05758627.pdf

...B6 **UNCLASSIFIED** U.S. Department of State Case No. F-2014-20439 Doc No. C05758627 Date: 02/19/2016...

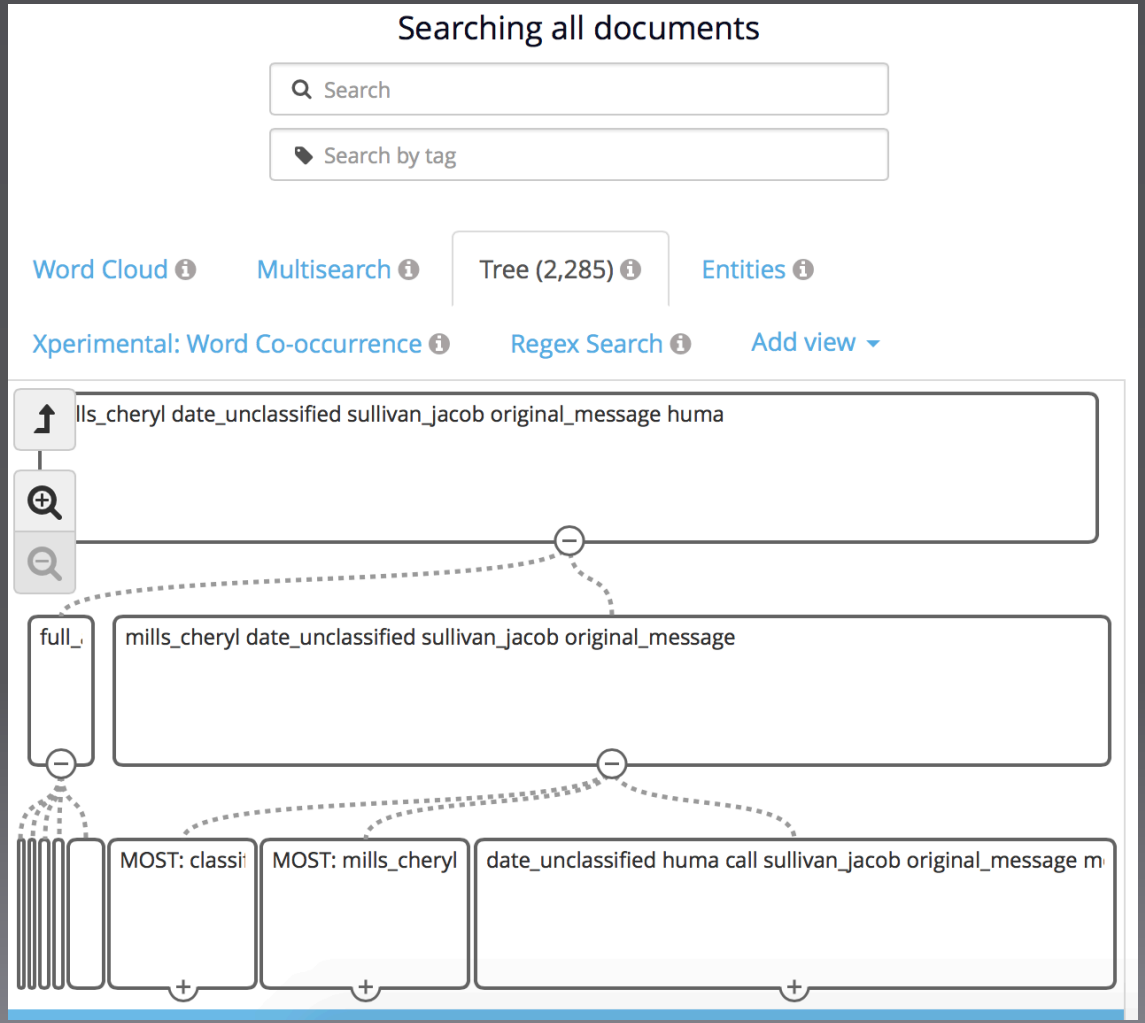
...2009 Subject: one edit to our saturday agreement B5 B6 **UNCLASSIFIED** U.S. Department

Mulitsearch



Tree

Organizes words into folders and subfolders. The size of the block shows how frequently words occur. You can see in what combination words appear together. You can tag these folders to make them more useful to your project. You can also search for specific terms.


Searching all documents
Search
Search by tag
Word Cloud
Multisearch
Tree (2,285)
Entities
Xperimental: Word Co-occurrence
Regex Search
Add view
lls_cheryl date_unclassified sullivan_jacob original_message huma
full_
mills_cheryl date_unclassified sullivan_jacob original_message
MOST: classi
MOST: mills_cheryl
date_unclassified huma call sullivan_jacob original_message m

Activity 2: Using Overview

Preparing Print Documents for Digitization and Text Mining

Overview of Basic Digitization

- Plan first so you don’t need to redo work
- Determine the best type of scanners
- Create your scanning workflow (and test it!)
- Use standard vocabularies/metadata to describe your scanned items
- File structure and naming systems
- Technical Specifications

Preparing Documents for Scanning

- Remove all staples, paper clips or other fastening devices.
- Repair all torn or damaged documents.
- Remove creases or folds for the pages so that no information is covered or lost.
- Remove extraneous documents.
- Identify and locate missing or misfiled documents.
- Identify and redact any confidential or sensitive data.

Arrange the documents in the order in which they are to be scanned

Overview of Scanning Steps

1. Align material on clean scanner bed.
2. Preview the scan.
3. Crop image. Leave sufficient margins.
4. Set to proper resolution and scale.
5. Scan.
6. Save at high resolution.
7. Perform editing for clearest image possible.
8. Create derivatives.
9. Save all derivatives.

Overview of Digital Camera Steps

1. Prepare room. Turn off overhead lights, close curtains.
2. Check settings for object you are capturing.
3. Clean lens.
4. Align objects directly under camera.
5. Preview through viewfinder or on computer, if possible.
6. Save high resolution image.
7. Perform editing for clearest image possible.
8. Create derivatives.
9. Save all derivatives.

Physical Environment

- Important if using digital camera
- Walls should be neutral – gray
- Use a white/light colored background
- Try to avoid glares or harsh lighting
- Do not use flash

Technical Considerations

- Resolution: at least 300 dpi for paper documents
- Color vs. Grayscale
- File Formats
- Target Size

Adjusting Scanned Images

- **Misconceptions**
- Image files saved directly from a scanner or digital camera are pristine or unmolested in terms of the image processing
- You should not perform any post-scan or post-capture adjustments on image files because the image quality might be degraded
- Deskew pages so word lines are horizontal
- Adjust pixilation and blurring

OCR

- Optical character recognition (or reader): the process of transforming images of characters in a document to the equivalent ASCII code for those characters
- How to OCR a pdf in Adobe Acrobat Pro
- Document > OCR Text Recognition > Recognize Text Using OCR
- Options in Adobe Acrobat Pro
- Batch Processing
- Document > OCR Text Recognition > Recognize Text in Multiple Files Using OCR
- Add Files or Folders
- Searchable Image vs. Clear Scan

OCR Programs

- Adobe Acrobat Pro (Readers only do not do OCR)
- ABBYY FineReader
- OmniPage Standard
- Presto! OCR
- Readiris Pro

Free

- Online OCR - http://www.onlineocr.net/
- Free OCR - http://www.free-ocr.com/

OCR Limitations

- Not 100% accurate.
- Some fonts convert better than others.
- Perceives some text as images (e.g., headers for *MAC Record*, *the Eagle*)
- Scan quality is important
- Handwriting, especially cursive, is problematic
- Some print publications wrap words at end of column to save space. (Lan –sing). OCR see them as two separate words.

OCR Demo

Wrap up and Questions