19 April 2019

COLLECTIONS AS DATA

Text Analysis with Voyant

http://tiny.cc/texts_voyant

Sara Duke, @SaraDuuuuke96
MLS/MIS Candidate, Information & Library Science
Digital Outreach Specialist, Institute for Digital Arts & Humanities

Michelle Dalmau, @mdalmau
Associate Librarian, Head of Digital Collections Services
Co-Director, Institute for Digital Arts & Humanities

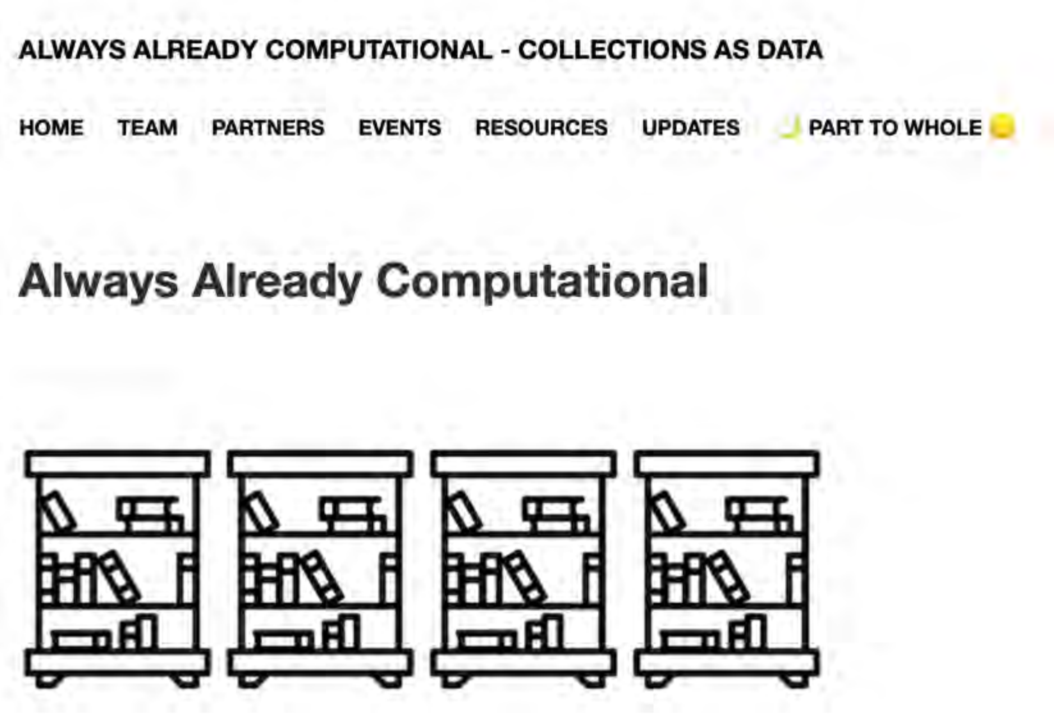
Externalize Affordances
of Digital Collections for Computation

Highlight Collections at IU and
Beyond **for Use/Re-Use**

Help you work with collections as data

Collections as Data

For **decades**, cultural heritage institutions have been **building digital collections**. Simultaneously, **researchers have drawn upon computational means to ask questions** and look for patterns. This work goes under a wide variety of names including but not limited to text mining, data visualization, mapping, image analysis, audio analysis, and network analysis. [With few exceptions], cultural heritage institutions have rarely built digital collections or designed access with the aim to support computational use.



ALWAYS ALREADY COMPUTATIONAL - COLLECTIONS AS DATA

Always Already Computational

https://collectionsasdata.github.io/statement/

Collections as Data (Con't)

"A collections as data paradigm seeks to foster an expanded set of research, pedagogical, and artistic potential predicated on the computational use of cultural heritage collections. Collections as data raises the question of **what it might mean to treat digitized and born digital collections as data rather than simple surrogates of physical objects or static representations of digital experience**."

Padilla, Thomas. "Collections as Data: Implications for Enclosure 2018." *College & Research Libraries*, vol. 79, no. 6, 2018, http://crln.acrl.org/index.php/crlnews/article/view/17003/18751.

"In general, computer-assisted methodologies such as text analysis, visualizations, and data mining are just such tools, but they often **provide the view the magnifying glass** gives the user when he or she turns it upside down. These methodologies **defamiliarize texts**, making them unrecognizable in a way (putting them at a **distance**) that **helps scholars identify features they might not otherwise have seen, make hypotheses, generate research questions,** and **figure out prevalent patterns and how to read them.**"

Text Analysis, Data Mining, and Visualizations in Literary Scholarship

By Tanya Clement in *Literary Studies in the Digital Age: An Evolving Anthology*, 2013 (https://dlsanthology.mla.hcommons.org/text-analysis-data-mining-and-visualizations-in-literary-scholarship/)

- **select or collect texts** in order to explore an hypothesis;
- **look for patterns** (of words, ideas, symbols, rhetorical or formal structures, etc) within an individual text and/or within sets of texts;
- **discover relationships** (of development, dependence, seriality, association, intention, allusion, intertextuality, etc) between parts of texts, whole texts, or sets of texts;
- **interpret the significance** of these patterns, relationships, and texts;
- **develop arguments** for the larger significance of these interpretations

Text Analysis

By Natalie M. Houston in *Digital Pedagogy in the Humanities: Concepts, Models and Experiments*, 2016

(https://digitalpedagogy.mla.hcommons.org/keywords/text-analysis/)

Preparing the Text Corpus

COLLECTIONS AS DATA //

- Data Gathering
  - Downloading / Web Scraping
  - Filenaming conventions
  - Transcription/OCR

- Data Cleaning
  - Tokenization
  - Removing extra-textual info
  - Fixing OCR errors

- Data Segmentation/Chunking
  - Structure
    - Book => Chapters
  - Temporal
    - Newspaper => Articles
  - Thematic
    - Letters to Hamilton's wife

Always keep an original version of your data set!

Exercise 1: Data Cleaning

COLLECTIONS AS DATA //

About Our Corpus

- Source of corpus: https://founders.archives.gov/
- Hamilton's outgoing correspondence
- 1774-1804
- 3,508 letters total

Steps for Exercise 1

- Go to http://tiny.cc/texts_voyant
- Click on Exercise _1.md

Voyant for Text Analysis

Voyant is an online program that provides several different text analysis results on either one single text or a corpus of documents, including frequency, trends, correlations, and topic modeling.

https://voyant-tools.org/

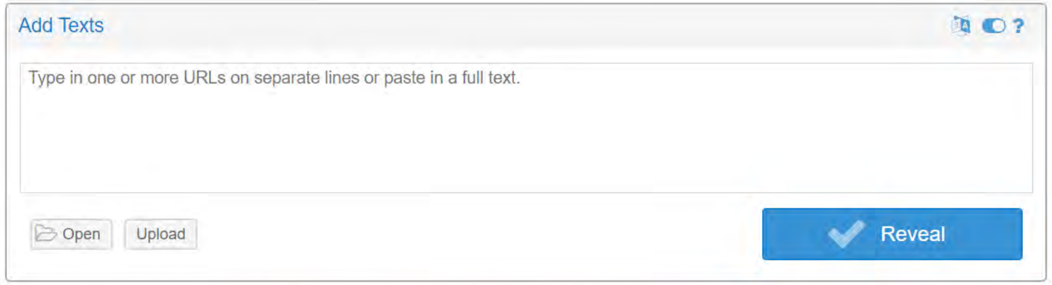
COLLECTIONS AS DATA //

Exercise 2: Using Voyant

- https://voyant-tools.org/
- Upload Texts
  - Click "Upload" and "Open"

VOYANT

see through your text



Add Texts

Type in one or more URLs on separate lines or paste in a full text.

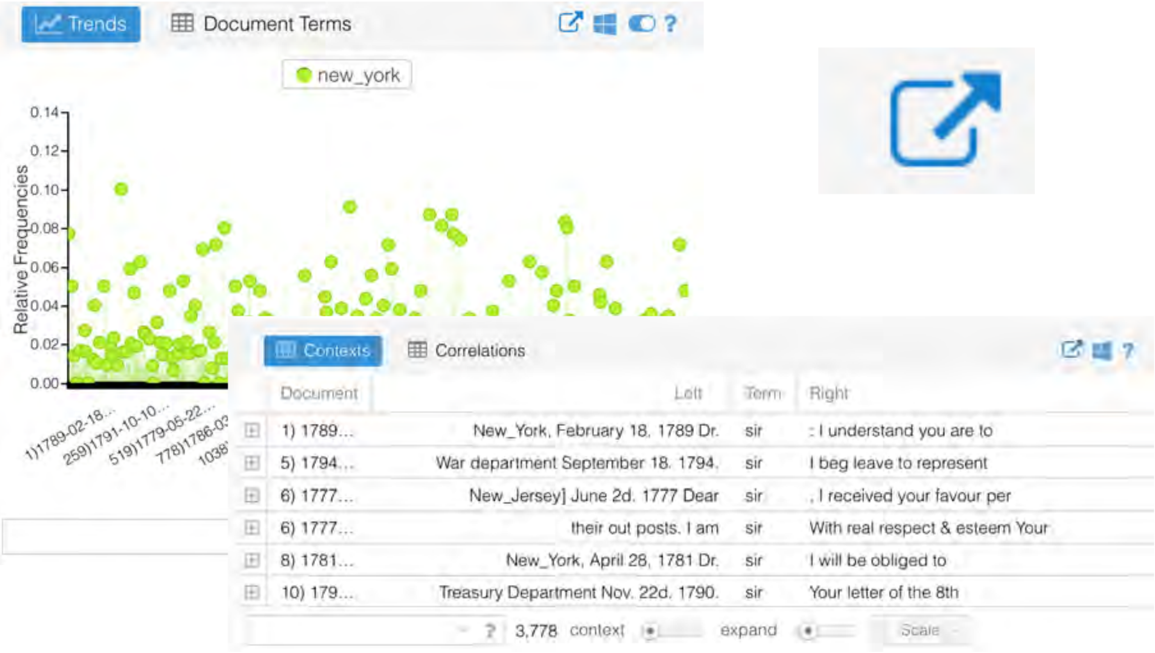
Voyant Tools is a web-based reading and analysis environment for digital texts.

Next Steps

- List of Voyant modules for text analysis with a brief description (just over 20 modules): https://digihum.mcgill.ca/voyant/tools/
-

Step-by-Step Voyant Tutorial: http://docs.voyant-tools.org/category/workshops/

- Export Graphs/Charts and Data from Voyant



	Document	Left	Term	Right
⊞	1) 1789...	New_York, February 18, 1789 Dr.	sir	: I understand you are to
⊞	5) 1794...	War department September 18. 1794.	sir	I beg leave to represent
⊞	6) 1777...	New_Jersey] June 2d. 1777 Dear	sir	, I received your favour per
⊞	6) 1777...	their out posts. I am	sir	With real respect & esteem Your
⊞	8) 1781...	New_York, April 28, 1781 Dr.	sir	I will be obliged to
⊞	10) 179...	Treasury Department Nov. 22d. 1790.	sir	Your letter of the 8th

3,778

Pedagogy & Research Consultations

COLLECTIONS AS DATA //

LIBRARIES

idah

Fall & Spring Semesters

Tuesdays:	10 am – 12 pm
Wednesdays:	2:00 pm – 4:00 pm

Schedule Us: https://idah.indiana.edu/

Email Us: idah@Indiana.edu