Fundamentals of Computational Text Analysis with Voyant

Dr. Wendy Perla Kurtz
Lecturer and Project Scientist
UCLA Program in Digital Humanities
wpkurtz@ucla.edu

TODAY'S WORKSHOP

- Introduction to computational text analysis (30 min)

    - What is it?

    - What can you do with it?

    - How you do it?

- Explore the Voyant dashboard (20 min)

    - default tools (aka "skins") & customization options

- Hands-on exploration with Voyant Tools (https://voyant-tools.org/) (remaining time)

THIS WORKSHOP HAS BEEN INFORMED BY…

- My work in the digital humanities

- Input from colleagues and students of digital humanities

- Scholarly articles and literature

- Work of other digital humanists

- Participation in numerous workshops (Digital Humanities Summer Institute, conferences day-long symposia, hour-long workshops)

- Product development and user experience research

WHAT IS COMPUTATIONAL TEXT ANALYSIS?

Text mining is a research practice that involves using **computational analysis** to discover information from **vast quantities** of digital, free-form, natural language, **unstructured text**.

COMPUTATIONAL TEXT ANALYSIS IS INTERDISCIPLINARY

"Text mining is an **interdisciplinary** endeavor that also borrows freely from **corpus linguistics** and **computational linguistics**, as well as **social-scientific traditions** like social network analysis…Humanistic text mining seeks to frame questions that contribute meaningfully to existing traditions of humanistic inquiry."
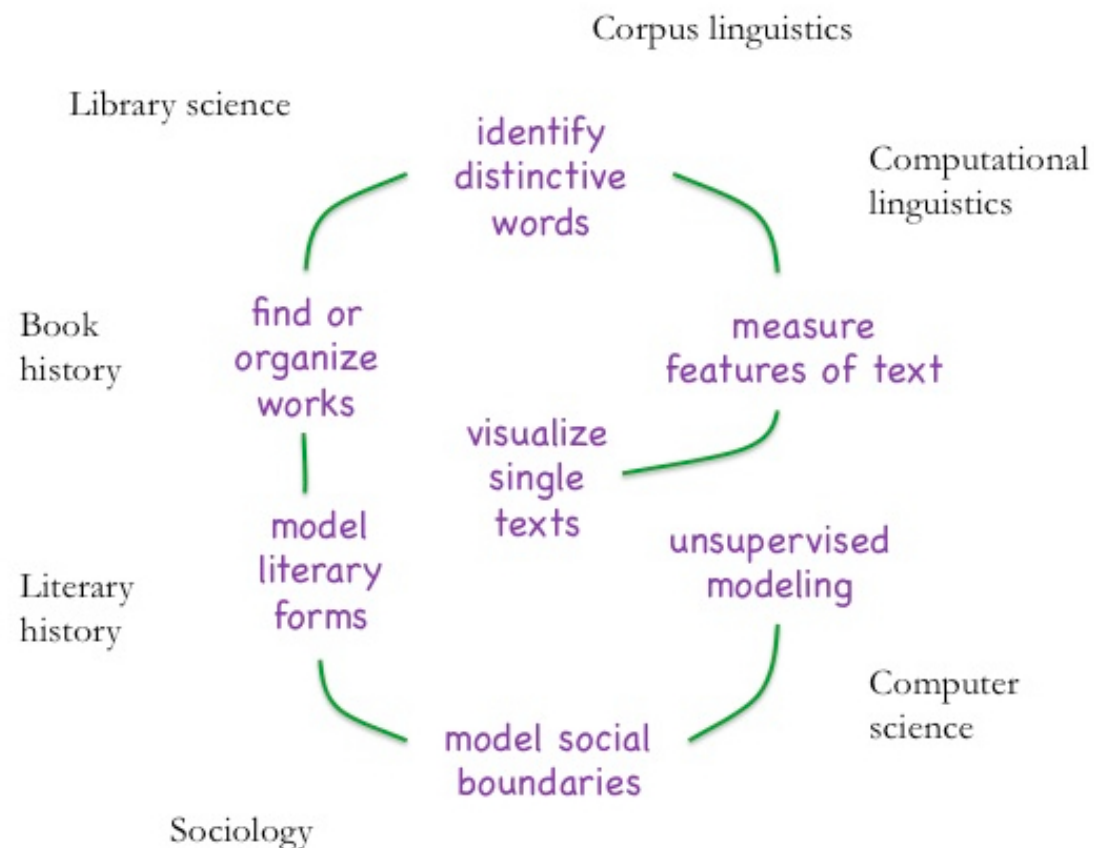
"Text-Mining the Humanities"
Matthew L. Jockers & Ted Underwood

"Seven ways humanists are using computers to understand text"
Ted Underwood

SYNONYMS?

- Quantitative study of literature (many)

- Algorithmic criticism (Ramsay)

- Digital literary studies (Siemens/Schreibman et al)

- Computer-assisted reading / literary analysis / interpretation (Rockwell/Sinclair)

- Distant reading (Moretti)

- Macroanalysis (Jockers)

- CLS (Computational Literary Studies) (Da)

- ...
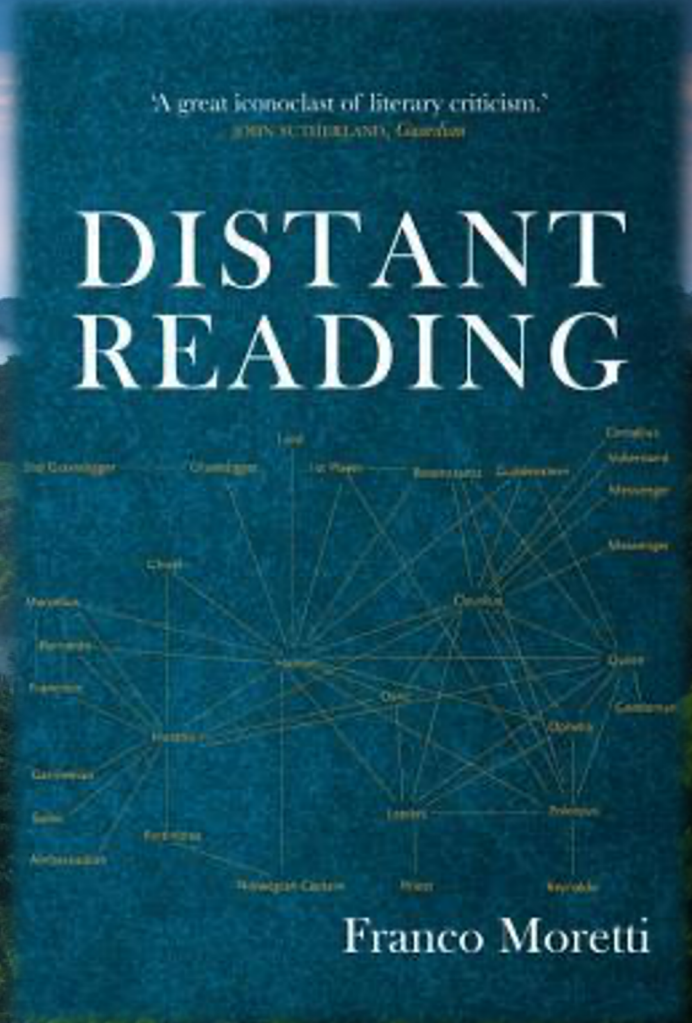
DISTANT READING VS. CLOSE READING
FRANCO MORETTI - *STANFORD LITERARY LAB*

WHAT CAN YOU DO WITH COMPUTATIONAL TEXT ANALYSIS?

Raw OCR text

Text
(digitized or born-digital)

Processed ("Clean") OCR text

Statistical output

Visualization output

YOU CAN…

- **Summarize** topics of interest in a group of texts

  Analysis method: Topic modeling & Clustering

- **Connect** common keywords among a group of texts

  Analysis method: Network analysis

- **Track** sentiment over topic, text source, time period

  Analysis method: Sentiment Analysis

- **Identify** names, locations, entities

  Analysis method: Natural Language Processing

- **Distinguish** texts in a corpus by a given author (i.e. who authored which federalist paper)

  Analysis method: Stylometry

- **Differentiate** poetry from prose

  Analysis method: Text Classification

- **Contrast** the vocabulary of different corpora

  Analysis method: Keyword/feature extraction

- **Categorize** documents

  Analysis method: Document/term clustering

Slide adapted from: UNLV "Introduction to Text Analysis"

APPLICATION FOR COMPUTATIONAL TEXT ANALYSIS
SAMPLE USE CASES

Text Analysis: Cultural studies

THE ACHILLES EFFECT

What Pop Culture is Teaching Young Boys about Masculinity

CRYSTAL SMITH

Crystal Smith

HOME | ABOUT | BOOKS | FREELANCE WRITING | BLOG

SEPTEMBER 20, 2017 · BY ADMIN

Word Cloud: How Toy Ad Vocabulary Reinforces Gender Stereotypes

This post went viral shortly after I published it on March 28, 2011. There was so much response that I had to write a preamble in April to address some of the questions and concerns readers had. It was really a simple exercise but garnered many thoughtful and important questions. The original comments were lost but it looks like most are visible via the WayBackMachine.

Slide adapted from: UNLV "Introduction to Text Analysis"

TEXT ANALYSIS: LITERARY STUDIES

AMERICAN FICTION 1774 - 1920

Positive adjectives and terms about family tend to dominate at the start of novels, and then tail off. Terms relating to death peak at the end of novels. There are some words (they've identified 200) that have a particular narrative "charge" (i.e. they dominate certain stages of a novel more than you'd expect).

David McClure
Stanford Literary Lab

Stanford Literary Lab
Director: Mark Algee-Hewitt

Distributions of words across narrative time in 27,266 novels

"athletic"

"death" in 27k novels
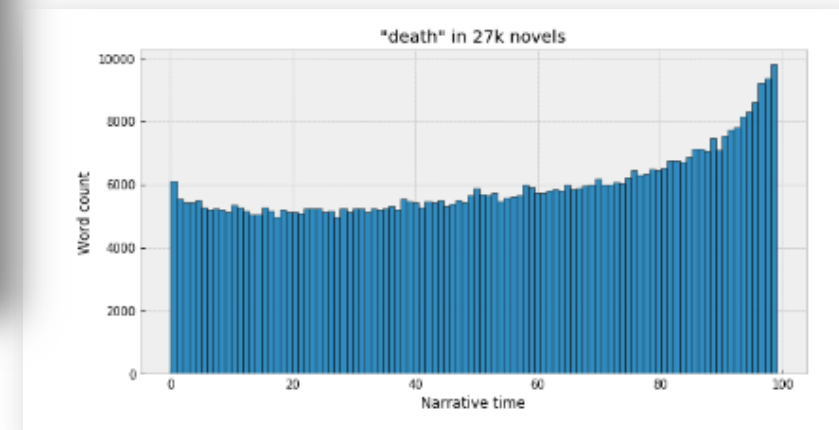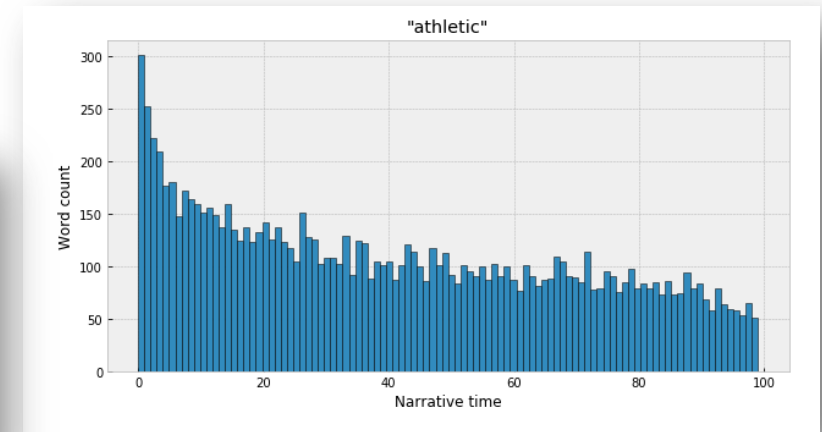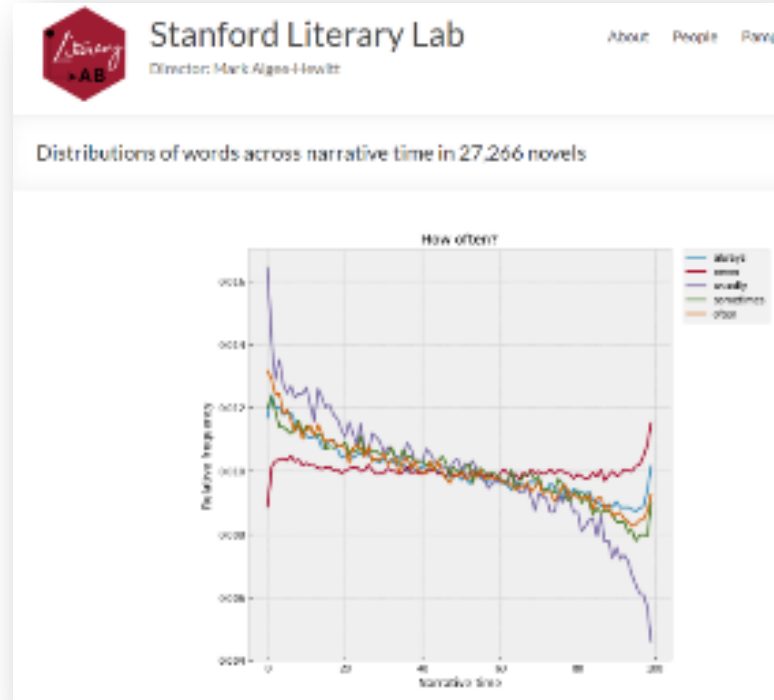
Text Analysis: Literary Studies

"Textual Reuse in the Eighteenth Century: Mining Eliza Haywood's Quotations"

Douglas Ernest Duhaime, University of Notre Dame

Digital Humanities Quarterly, Volume 10 Number 1, 2016.

Text Analysis:
historical newspapers

the language of british suffrage in the press

Dr Kat Gupta
University of Nottingham

Text analysis: historical newspapers

"Mining the *Dispatch*"

Robert K. Nelson, Director of the Digital Scholarship Lab at the University of Richmond

http://dsl.richmond.edu/dispatch/pages/home
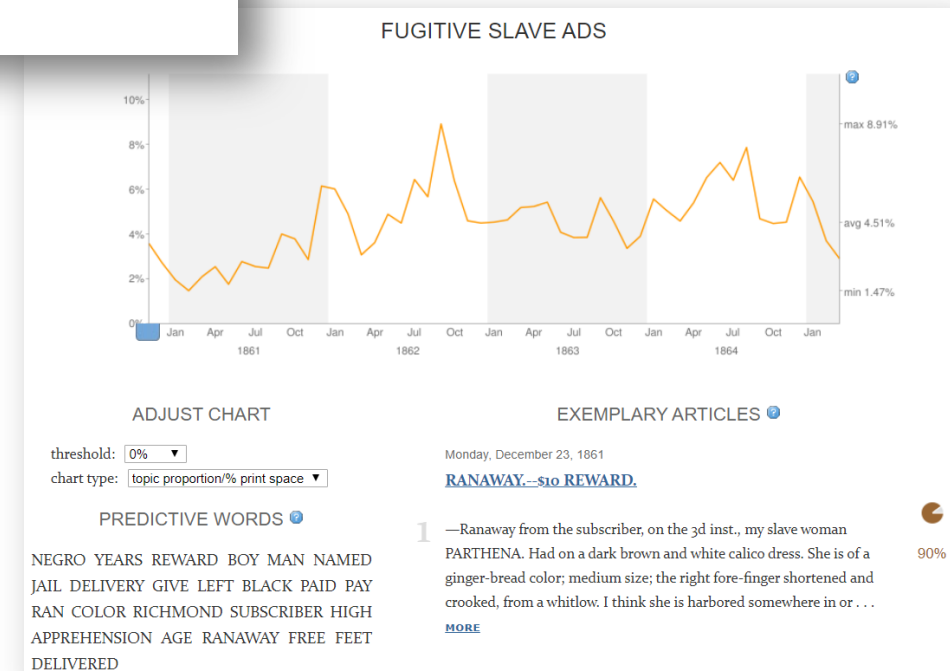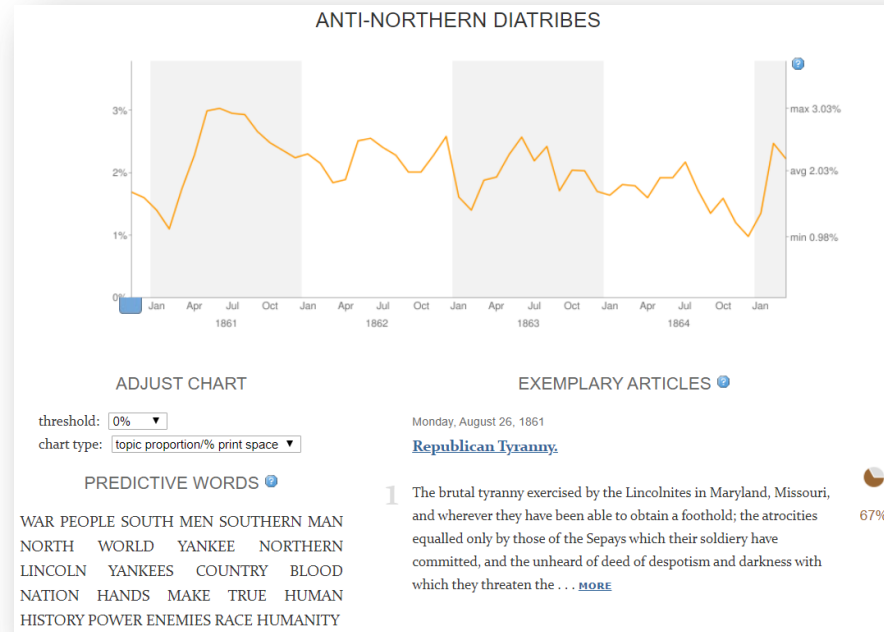
SO…HOW DO WE DO IT?

TYPES OF TEXT YOU CAN MINE

Digitized Texts

Physicals documents that are digitized and processed using optical character recognition or manually keyed to create a digital facsimile.

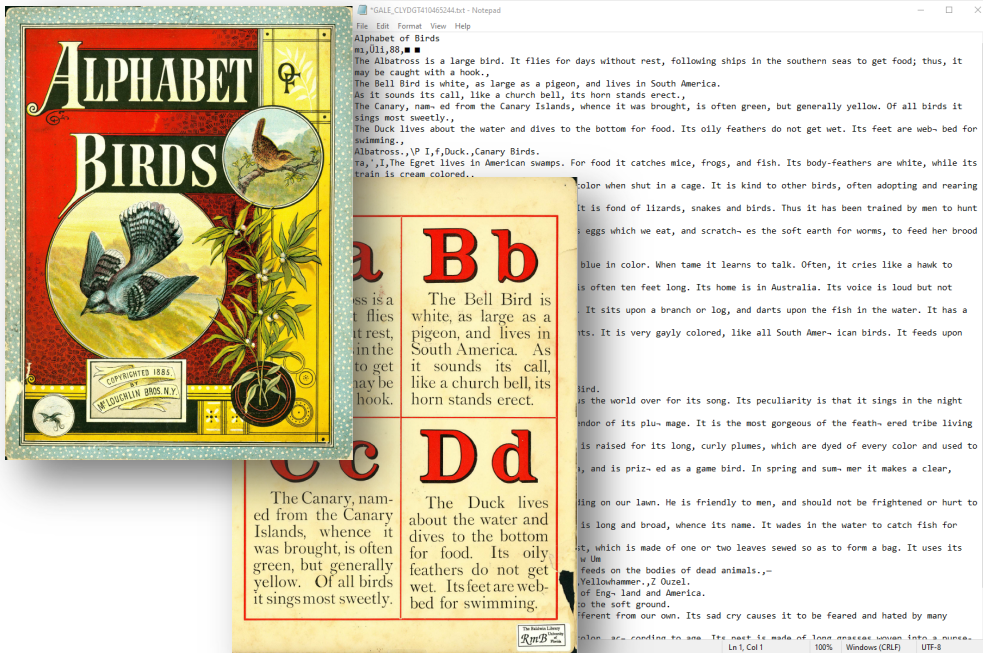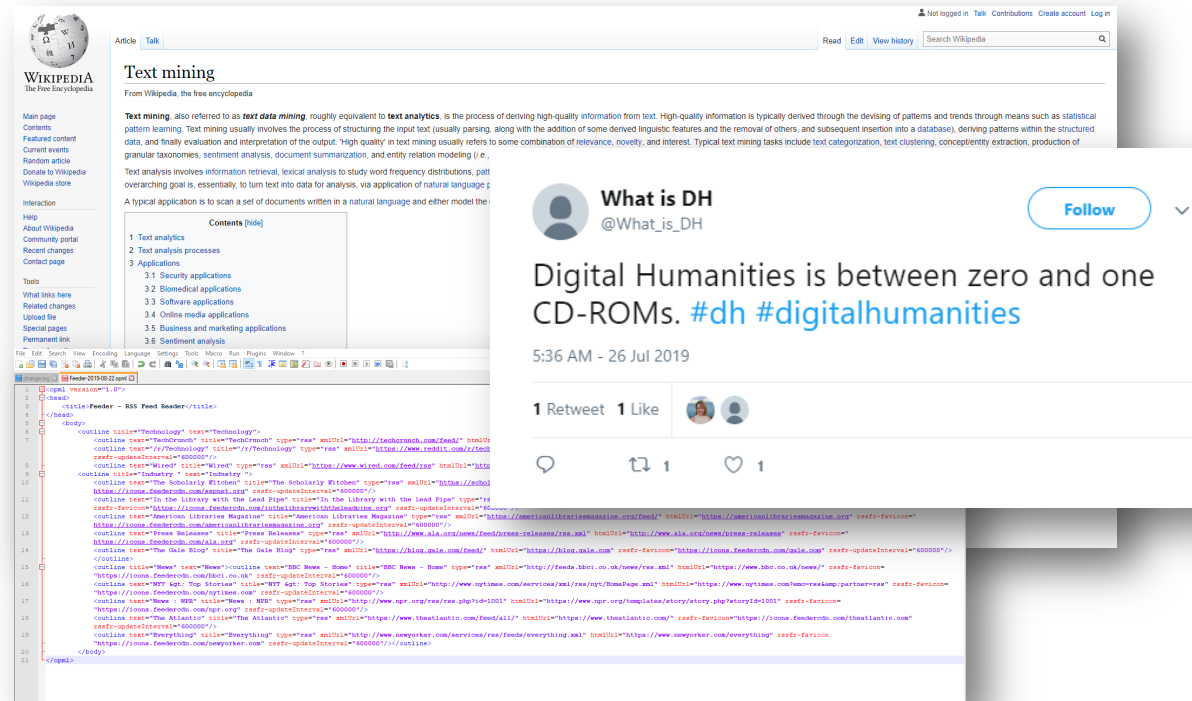## Native Digital Texts

Texts created in a digital format for the purpose of being accessed on an electronic device.

GETTING MACHINE-READY TEXT

- Machine-ready = machine-readable = capable of being processed by a computer

- Text must be OCR'd = optical character recognition
  - ABBYY Finereader is the standard for OCR (but not free)
  - Free options available but results will likely be noisy

- Not all things can be (easily) OCR'd
  - Handwriting, images, some older print fonts

PLACES TO GET TEXT

Digitized Texts

- Internet Archive
- Project Gutenberg
- Google Books
- Hathi Trust
- JSTOR Data for Research
- PubMed Open Access Subset
- Open American National Corpus

Native Digital Texts

- Email
- HTML
- RSS Feeds
- Twitter
- Wikipedia
- Data Liberation Front
- New York Times API

Dataset Repositories

- Kaggle
- English-corpora.org (BYU)
- Data is Plural (Jeremy Singer-Vine)
- DH Toychest (Alan Liu)

Slide adapted from: UNLV "Introduction to Text Analysis"

PLACES TO ANALYZE TEXTS

Programming Languages

- Python (Text Cleaning & Statistical Analysis)

- R (Statistical Analysis & Visualization)

- Javascript (Visualization)

- GeoJSON (Geo-mapping)

Software Libraries

- MALLET (Topic Modeling)

- spaCy (Natural Language Processing)

- Scrapy (extracting the data from websites)

- Transkribus

Out-Of-The-Box

- Voyant

- Lexos

- Juxta

- AntWord Profiler

- Textometrie (TXM)

- Textal

- Gephi

- Palladio

Other helpful links:

- TAPor

HTTPS://VOYANT-TOOLS.ORG/

VOYANT TOOLS OVERVIEW

- **What it does**: computationally processes large amounts of text (**corpus/corpora**) to produce a series of customizable visualizations for text analysis

- **What you need**: a corpus - single plain text doc or many docs combined (e.g. an author's entire works or mix-and-match works from different authors)

- **How it works**: upload your text and then start to explore!

ADDING A TEXT(S) FOR ANALYSIS

1. Upload documents (.txt, .doc, .csv, .html, .xml, .pdf, .zip, .json...)

2. Paste URLs (must link to a readable file format)

3. Paste plain text

4. Choose from existing

The Wonderful Wizard of Oz

Chapter I
The Cyclone

Dorothy lived in the midst of the great Kansas prairies, with Uncle Henry, who was a farmer, and Aunt Em, who was the farmer's wife. Their house was small, for the lumber to build it had to be carried by wagon many miles. There were four walls, a floor and a roof, which made one room; and this room contained a rusty looking cookstove, a cupboard for the dishes, a table, three or four chairs, and the beds. Uncle Henry and Aunt Em had a big bed in one corner, and Dorothy a little bed in another corner. There was no garret at all, and no cellar—except a small hole dug in the ground, called a cyclone cellar, where the family could go in case one of those great whirlwinds arose, mighty enough to crush any building in its path. It was reached by a trap door in the middle of the floor, from which a ladder led down into the small, dark hole.

When Dorothy stood in the doorway and looked around, she could see nothing but the great gray prairie on every side. Not a tree nor a house broke the broad sweep of flat country that reached to the edge of the sky in all directions. The sun had baked the plowed land into a gray mass, with little cracks running through it. Even the grass was not green, for the sun had burned the tops of the long blades until they were the same gray color to be seen everywhere. Once the house had been painted, but the sun blistered the paint and the rains washed it away, and now the house was as dull and gray as everything else.

When Aunt Em came there to live she was a young, pretty wife. The sun and wind had changed her, too. They had taken the sparkle from her eyes and left them a sober gray; they had taken the red from her cheeks and lips, and they were gray also. She was thin and gaunt, and never smiled now. When Dorothy, who was an orphan, first came to her, Aunt Em had been so startled by the child's laughter that she would scream and press her hand upon her heart whenever Dorothy's merry voice reached her ears; and she still looked at the little girl with wonder that she could find anything to laugh at.

Uncle Henry never laughed. He worked hard from morning till night and did not know what joy was. He was gray also, from his long beard to his rough boots, and he looked stern and solemn, and rarely spoke.

It was Toto that made Dorothy laugh, and saved her from growing as gray as her other surroundings. Toto was not gray; he was a little black dog, with long silky hair and small black eyes that twinkled merrily on either side of his funny, wee nose. Toto played all day long, and Dorothy played with him, and loved him dearly.

Today, however, they were not playing. Uncle Henry sat upon the doorstep and looked anxiously at the sky, which was even grayer than usual. Dorothy stood in the door with Toto in her arms, and looked at the sky too. Aunt Em was washing the dishes.

From the far north they heard a low wail of the wind, and Uncle Henry and Dorothy could see where the long grass bowed in waves before the coming storm. There now came a sharp whistling in the air from the south, and as they turned their eyes that way they saw ripples in the grass coming from that direction also.

Suddenly Uncle Henry stood up.

"There's a cyclone coming, Em," he called to his wife. "I'll go look after the stock." Then he ran toward the sheds where the cows and horses were kept.

Aunt Em dropped her work and came to the door. One glance told her of the danger close at hand.

"Quick, Dorothy!" she screamed. "Run for the cellar!"

Toto jumped out of Dorothy's arms and hid under the bed, and the girl started to get him. Aunt Em, badly frightened, threw open the trap door in the floor and climbed down the ladder into the small, dark hole. Dorothy caught Toto at last and started to follow her aunt. When she was halfway across the room there came a great shriek from the wind, and the house shook so hard that she lost her footing and sat down suddenly upon the floor.

Then a strange thing happened.

The house whirled around two or three times and rose slowly through the air. Dorothy felt as if she were going up in a balloon.

The north and south winds met where the house stood, and made it the exact center of the cyclone. In the middle of a cyclone the air is generally still, but the great pressure of the wind on every side of the house raised it up higher and higher, until it was at the very top of the cyclone; and there it remained and was carried miles and miles away as easily as you could carry a feather.

The Wonderful Wizard of Oz from Project Gutenberg
https://www.gutenberg.org/files/55/55-h/55-h.htm

PRE-PROCESSING YOUR DATA

After uploading your corpus, Voyant will:

- guess the file format

- ignore punctuation and symbols

- parse & index corpus text based on specified delimiters, or tokenize text

- apply stopwords dictionary to omit common terms unlikely to be significant (e.g. "the," "and," etc.)

Voyant will not:

- correct misspelled or merged words (via OCR)

- cluster spelling variants

- translate HTML entities or Unicode characters

But errors can be easier to discover in Voyant…

Add Texts

"My darling child!" she cried, folding the little girl in her arms and covering her face with kisses. "Where in the world did you come from?"

"From the Land of Oz," said Dorothy gravely. "And here is Toto, too. And oh, Aunt Em! I'm so glad to be at home again!"

Open   Upload   ✔ Reveal

Options

Input Format: Auto-Detect (recommended)

For more information on the advanced options below, see the documentation on *creating a corpus*.

Corpus
Text
XML
HTML
JSON
Tables
Processing
Text Language: Auto-detect
Tokenization: Automatic (highly recommended)
               Automatic (highly recommended)
               Simple Word Boundaries
               Whitespace Only
Access Manag

VOYANT DASHBOARD

1. <u>Cirrus</u>: a kind of word cloud showing the most frequent terms

2. <u>Reader</u>: an efficient corpus reader that fetches segments of text as you scroll

3. <u>Trends</u>: a distribution graph showing terms across the corpus (or terms within a document)

4. <u>Summary</u>: a tool that provides a simple, textual overview of the current corpus

5. <u>Contexts</u>: a concordance that shows each occurrence of a keyword with a bit of surrounding context

ANATOMY OF A VOYANT TOOL

Default tools for panel area
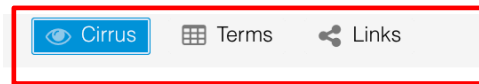
Tool details
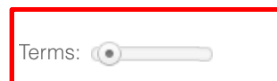
Tool-specific interactions

CIRRUS TOOL

- word size based on frequency of the term's occurrence within the corpus

CUSTOMIZATION (STOPWORDS & APPEARANCE)

- Click the options in the panel to access

- Edit list to add corpus-specific terms

- Apply globally doesn't always work :-(

Options icon

Options

Stopwords:	Auto-detect	Edit List	☑ apply globally
White List:		Edit List	
Categories:	auto	Edit	
Font family:	Palatino		
Palette:	default	Edit List	

↺ Reset   ✖ Cancel   ✔ Confirm

CUSTOMIZING + REFINING RESULTS

Results can be customized/refined in a number of ways:

- Adjust # of terms/items you're seeing (except Reader)

- Stopwords to filter out noisy words

- Stemming + other wildcard queries to zoom in on specific terms/patterns
  - * to truncate or control for variance
  - ^ for exact match (if you only type a word, ^ is implied)
  - term01 | term02 for exact pair

TERMSBERRY

- a way of exploring high frequency terms and their collocates (words that occur in proximity)
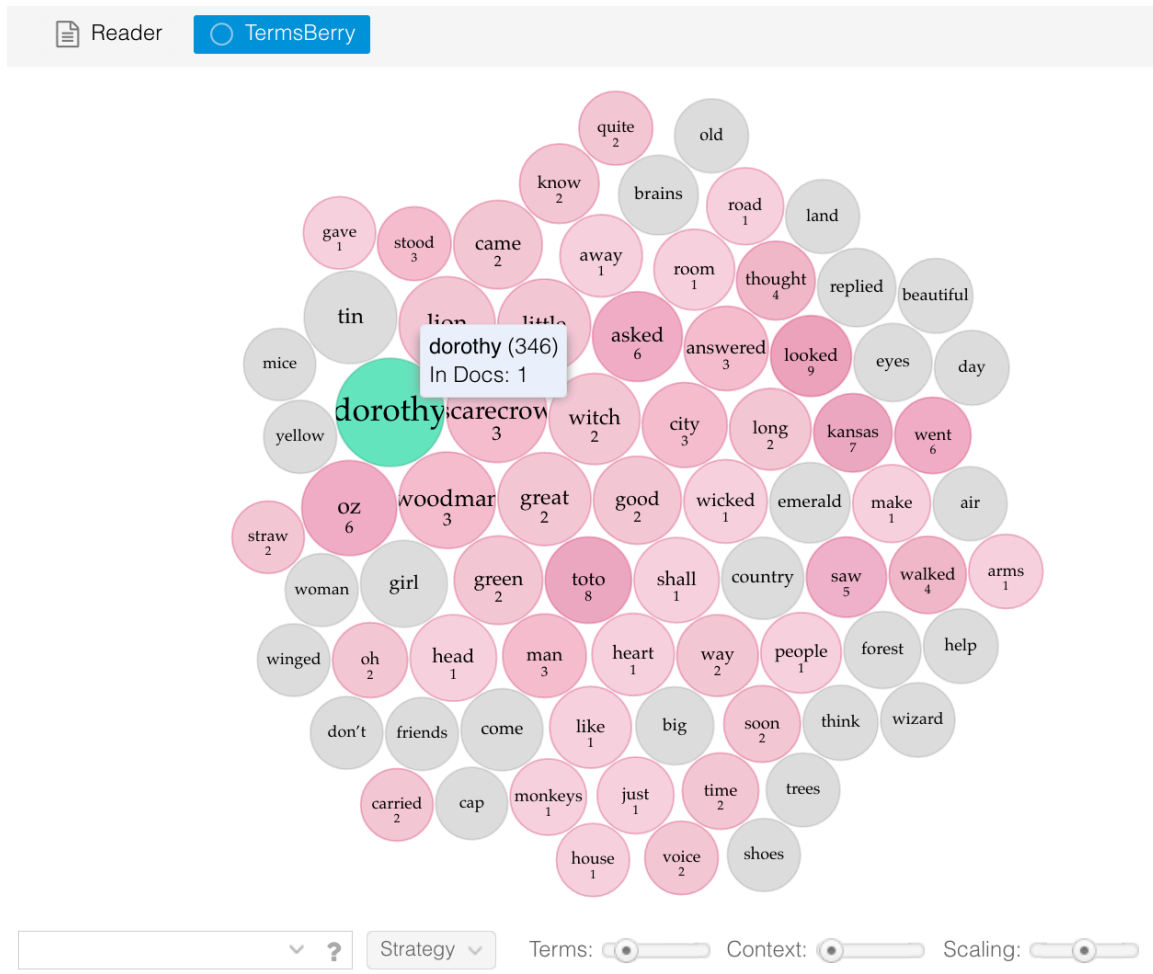
TRENDS TOOL

- compare relative frequencies of multiple terms in a corpus or across segments in a document
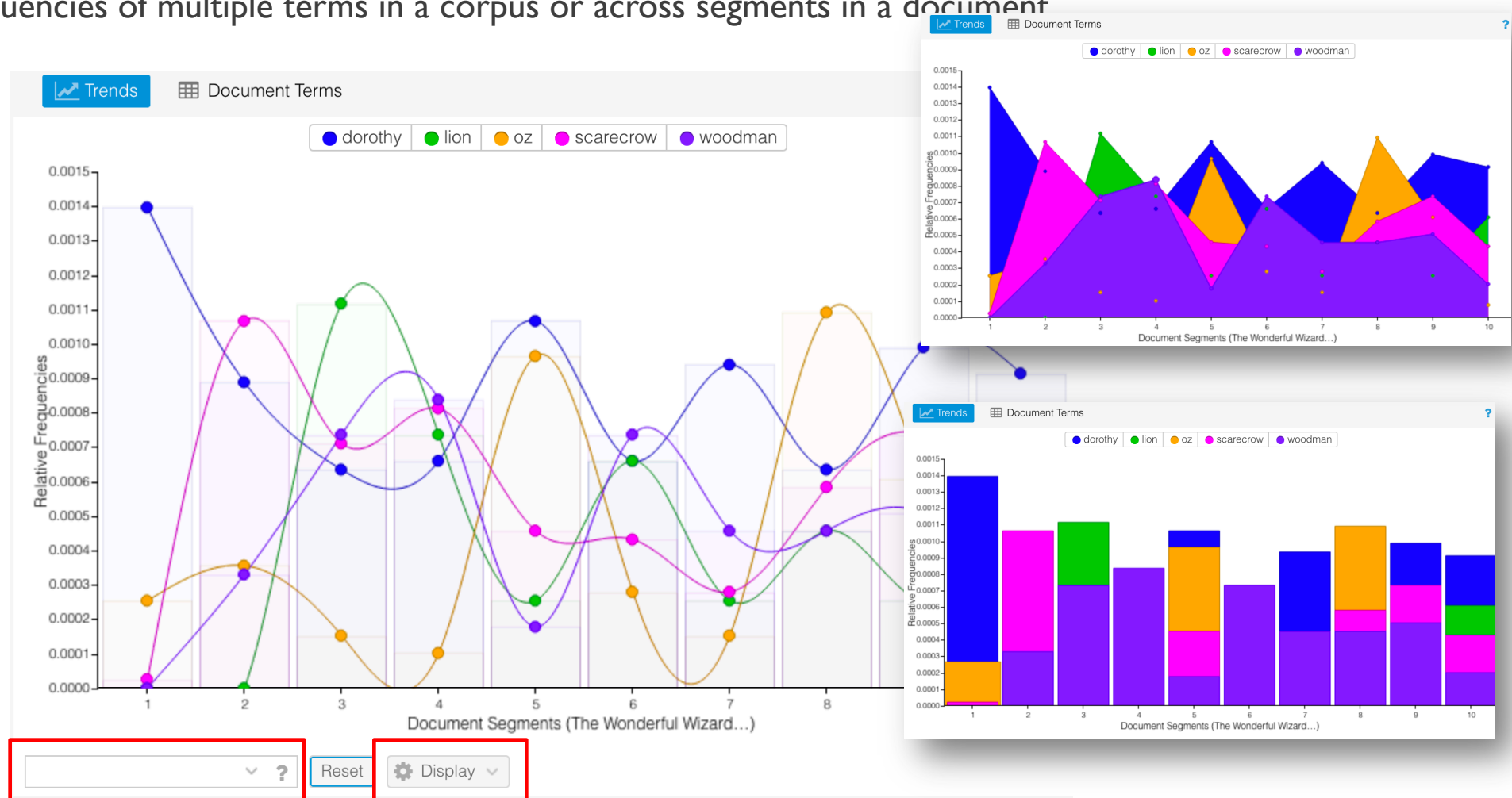
CONTEXTS TOOL

- active word in the contexts in which it appears (re: sense)

EXPORTING

- You can export a URL, an embeddable tool (interactive), data, or a bibliographic reference

- Applies to the entire Voyant project or a single particular tool ("skin")

  - You can return to your corpus and continue to work on it later, but:

    - ensure that you bookmark the URL. It is the ONLY way to access your corpus again

- You can also export a static .png of visualizations (a screenshot may have better image quality)

CHANGING THE TOOLS ("SKINS")

All of the panels/tools are resizable + interchangeable

- Tools ("skins") have baked-in options to switch tools
  - i.e. switching from *Reader* to *TermsBerry*
- But you can also change to other tools
- Options are categorized (but overlapping)

some are beta

- Corpus Tools
- Document Tools
- Visualization Tools
- Grid Tools
- Other Tools

HOW TO PICK A VIS TOOL ("SKIN")?

1. Figure out what you're trying to do and why
    a. Does context matter?
    b. Collocation? Relationship between terms?
2. Determine the type of tool (corpus, document, visualization, grid)
    a. Do you need to see something visualized?
3. Read about each tool and its intended uses to determine if the use fits your need: voyant-tools.org/docs/#!/guide/skins

LET'S EXPERIMENT!

SOURCES & INSPIRATION (also cited throughout)

- "DHSI 2019 Introduction to Computation for Literary Criticism," Randa El Khatib and David Joseph Wrisley

- "DHSI 2019 Making Choices About Your Data," Paige Morgan and Yvonne Lam

- Priscilla Finley and Chelsea Heinbach, University of Nevada, Las Vegas, "Introduction to Text Analysis: Using Voyant to encounter new questions."

- Pam Lach, San Diego State University, "Voyant for Text Analysis."

- Devon Mordell, McMaster University, "An Introduction to Textual Analysis With Voyant Tools."