DSPS Summer Fellows 2017 - Computational Analysis of Text

An overview of computational analysis of text, foundations, and exploration of challenges and strategies.

Preparation

- **Please bring a laptop to support your explorations!** Bring your own or feel free to check one out at the Olin circulation desk. No special software will be needed. All exercises will be done through a Web browser, without any special plugins.
- **Suggested Reading**
    - The first three sections of Text-mining in Wikipedia. This will give us a common orientation to text analytics.
    - Ted Underwood's blogpost "Seven ways humanists are using computers to understand text"
    - Matt Jockers blogpost "A Novel Method for Detecting Plot"
- **Bring text samples** to explore with Voyant. Format can be plain text, PDF (with OCR), MS Word Document or URL (for analysis of web pages). Upload of material will be subject to the Voyant privacy policy, so bring text you can safely share.
- **Look through this lesson plan**, develop your questions - and bring them to class.
- *Optionally* - the intrepid may want to obtain a login to the HTRC Portal, create a workset and run a few algorithms in advance of this lecture. Documentation for obtaining a sign-on and documentation for the portal will be helpful.

Going down the rabbit hole: anatomy of a digital book

How is a digital book made? How does the structure relate to its function? What opportunities does this afford us in terms of text analysis?

(1) Consider this book: Alice's Adventures in Wonderland. Explore the controls on the right side of the page turner.

- Note the different views of the book. What types of digital files are likely to make up the parts of a digital book?
- How are these files likely to be made?
- For what types of uses are each suited?

(2) Now explore the controls at the top of the page turner.

- There is a box labeled "search in this text". What can you deduce about the book given this functionality?
- What do the other controls do? Is there a way to summarize this class of controls? What underlying logic might you predict that coordinates these functions? (Food for thought...download and display in a browser.)

(3) What might this page be? (It also has this view.) Is this also part of the book? When and how might it be used?

(4) Diving deeper into text. Optical Character Recognition (OCR) processes are not perfect. Consider some areas of special challenge:

- poor image quality (resolution, warp, skew, crop)
- books where characters vary from "standard"
    - handwritten manuscripts, block printed books
    - early font styles, many non-Roman alphabets (Panjabi example)
- character-based writing systems: Japanese done well, and not so well
- language specific challenges - Arapaho gospel of St. Luke
- Mixed languages: Latin/Greek, German/Greek
- Unexpected arrangement

Computational analysis of text

We count tokens - What is tokenization? Why tokenize? What are some strategies used to tokenize?

(1) Let's look again at the Arapaho gospel of St. Luke**.** Switch to text view.

- Is this OCR accurate to the visually captured page?
- What is a word? How would you define "word" to a computer?
- What isn't a word? How would you tell a computer to exclude these?
- Consider languages with which you are familiar. Can you think of cases where tokens might contain more than one word?
- What sets of rules would we need in order to tokenize effectively? Would these be ordered in any specific way?
- Is there a "right way" to tokenize?

(2) Discussion: What are the opportunity points that the structure and arrangement of a book afford?

- How do challenges with OCR intersect with strategies for computational analysis of text? What might be effective strategies to deal with these challenges?
- What exactly is the "text"? Can you think of parts of a book that you might not want to include in your analysis? Why or why not? If you would, how would you exclude these parts?

Introducing Control - Microanalysis with Voyant

Voyant is a low barrier text analysis tool that delivers a rich, interactive interface and a variety of visualizations based on token counts within a single or a few texts. Input format can be plain text, a PDF (with OCR), an MS Word Document or a URL for HTML analysis. Documentation will help you use this tool and it's many features. Upload of any material will be subject to the Voyant privacy policy. Sample texts and URLs for analysis are listed below for experimentation, but feel free to use other source data that interests you.

- Sample texts, courtesy of Project Gutenberg. Use the **plain text** version.
  - Harriet Beecher Stowe, 1811-1896. Uncle Tom's Cabin
  - Zora Neale Hurston, 1891-1960. Three Plays
  - Bret Harte, 1836-1902. Urban Sketches
  - Steven Levy, 1951- . Hackers, Heroes of the Computer Revolution
- Sample URLS: copy and paste into the Voyant upload browser window to get started.Sample URLS: copy and paste into the Voyant upload browser window to get started.
  - Economics of Crisis - http://www.economicsofcrisis.com/indications.html
  - Instructions to major John Sullivan. Washington, George, 1732-1799. The writings of George Washington from the original manuscript sources. Internet Archive.
  - Copyright Law of the United States of America and Related Laws Contained in Title 17 of the United States Code - http://www.copyright.gov/title17/92preface.html
  - From Politico - transcripts of the 2016 Presidential debate
    - First Debate (09/27/2016) http://www.politico.com/story/2016/09/full-transcript-first-2016-presidential-debate-228761
    - Second Debate (10/10/2016) http://www.politico.com/story/2016/10/2016-presidential-debate-transcript-229519
    - Third Debate (10/20/2016) http://www.politico.com/story/2016/10/full-transcript-third-2016-presidential-debate-230063
- Sample visualization: Dr. Martin Luther King, Jr. *I Have a Dream* speech.

(1) Visualization of derived data

- Explore visualizations in the "dashboard" that results from analysis of uploaded text. Explore changing the options for visualizations.
- Discuss the relative merits of the various visualizations.

(2) Exerting control

- Experiment with stopwords - documentation
- Experiment with the slider for word counts
- Consider raw vs relative frequencies

(3) Discussion

- We calculate frequency
  - We can express our counts simply (as counts), or we can express them as frequencies. Why calculate frequencies?
  - Is either representation misleading? If so, in what ways?
- What does exerting control do to our results? Does it change the validity of our assertions?
- How should method be explained when making assertions from results?
- Who determines what is "signal" and what is "noise"?

Moving from Microanalysis to Macroanalysis (Google nGrams and HTRC bookworm)

nGrams are words or phrases, tokenized and counted in a defined corpus and displayed as a graph showing relative frequencies of those phrases as occurring over publication date. The two tools referenced below provide a basis for exploration of ngrams. Each tool is bound to secondary data derived from analysis of a different corpus, so results of the same nGram will not necessarily align.

Google's nGram Viewer. Use the links below as starting points; dynamic modifications can be made at any point. Rules for syntax can be found on the About page.

- a few racist depictions
- women authors of color
- computer and cybernetics - broken out by language
- enslaved * (This link will yield an error, but simply click on the blue button marked "search lots of books" and the error will resolve.)
- different types of alleys (This link will yield an error, but simply click on the blue button marked "search lots of books" and the error will resolve.)
- Seaports of the Mediterranean - note groupings by country

HathiTrust Research Center (HTRC) Bookworm (Tied to a corpus pre-1923). Again, consider these links as starting points. Rules for faceting ad controls are available on the HTRC wiki.

- mystery (contrast fiction and nonfiction)
- computer in fiction/non-fiction conference proceedings
- Does truth correlate with beauty?

Discussion

- How can these examples above be refined and improved?
- Compare the two interfaces, especially as to the affordances and the limits of each.
- What additional elements of control would be useful that aren't available?
- When we see unexpected or entirely expected wave forms, what do we make of these?
    - How much can we read into coorelations?
    - Do these constitute discoveries or represent errors? How can we distinguish?
    - Would the flaws be due to the data, the metadata, the algorithms?
- Are there lenses that we should be wary of?

More Macroanalysis: HTRC

HathiTrust Research Center (HTRC) is a collaborative research center (jointly managed by Indiana University and the University of Illinois) dedicated to developing cutting-edge software tools and cyberinfrastructure that enable advanced computational access to large amounts of digital text. A basic orientation of HTRC services is available, and features and steps for each are documented on the HTRC community wiki. We will be spending time in the **HTRC Portal** looking at the results of a few algorithms as a sampling of possibilities (links below will not render for all, but are parked to make sharing easier). Algorithms in the portal can

- Compare one collection of books to a second collection, and report the differences in frequency of tokens - ShksprDunning
- Extract entities from a set of books, and list out referents of where they occur
    - Person, location - WSPlaysEntityExtract
    - Dates - WSComediesExtractDates (Note the wide variety of what is considered a "date")
    - Dates visualized over timeline - WSComediesDateExtractSimileRemix
- Model "topics", or clusters of tokens that are statistically more likely to be found together - WSComediesTopics; WSTragediesTopics

Discussion

- In general, do the results look valid? Do any of these algorithms yield results that might be considered confusing or less than perfect?
- Note that results can be downloaded. What might be advantages of this portability?
- Are there things that the researcher would want or need to know about these algorithms when making claims about results?

On your own - network analysis and image analysis

Immersion

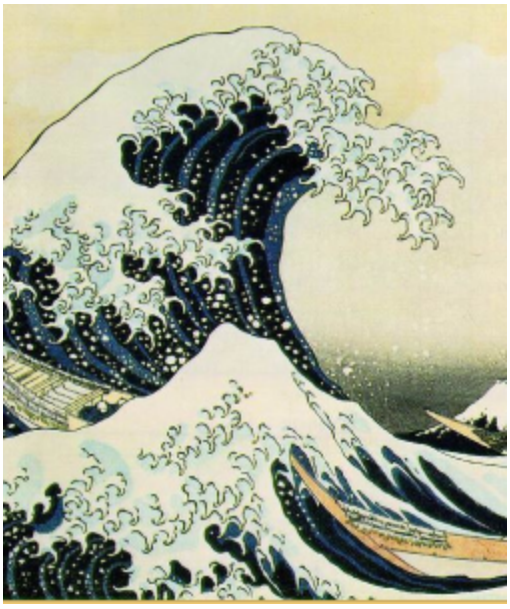
Immersion is a tool for discovering the connections in a corpus of email. It analyzes the flow data (information found in email headers) and represents these as a network of entities. The analysis is done in real time on the flow data for which you provide credentials. The display is rich and interactive. *By design, Immersion collects only header information (From, To, Cc and Timestamp).* However, using the actual flow data from your account may cause concerns regarding privacy - Be sure to read over the FAQs to understand what information you are granting access to, and how it will be used. If you do not like the terms of the tool, you can experience it with their demo data.

- Immersion
- Demo (If you would rather not use your own account)

Image analysis

Ukiyo-e.org is a database and image similarity analysis engine, created by John Resig to aid researchers in the study of Japanese woodblock prints. The data is over 213,000 digital copies of prints from 24 institutions, and their cataloging metadata. Metadata is indexed and searchable, as you might expect. (Details are noted in the about page.) But images are also searchable. Resig's Image search uses the TinEye matching engine to determine edges in an uploaded sample and compares with analyzed edges in database, returning probable matches (edge analysis). Use samples below to experiment.

Sample Images for search - click on desired image to display and choose either "download" or "save as..."






Resources

"Formatting Science Reports." *Academic and Professional Writing: Scientific Reports*. University of Wisconsin - Madison, 24 Aug. 2014. Web. 03 June 2016.

Jockers, Matthew L *Text Analysis with R for Students of Literature*. Cham: Springer International Publishing, 2014.

Underwood, Ted. *The Stone and the Shell*. Ted Underwood. Web. Accessed 03 June 2016.