Computational Text Analysis in the Humanities, Arts and Social Sciences: Resources, protocols and opportunities

Tyne Daile Sumner, December 2018

Key messages:

- Text Analysis is a long-standing core component of Humanities research
- Text Analysis using computational methods encompasses a broad range of processes by which text and natural language documents can be organized and modified so that they can be analysed
- There are several distinct phases in a computational text analysis process, including text collection, cleaning and parsing, text summary and analysis, and text visualisation
- This guide focuses on analysis and visualisation of text using **Voyant Tools**

Learning objectives:

- Understand problems and questions related to digital HASS research in the area of computer-assisted Text Analysis
- Understand the challenges and barriers to entry for HASS researchers wanting to conducted computer-assisted Text Analysis in their research
- Examine and understand Voyant Tools as a useful starting point for computer-assisted text analysis in HASS research
- Learn effective research project design and outcomes when using Voyant Tools for Text Analysis in HASS research

Ingredients:

Pre-reading:

- Ted Underwood's "Seven ways humanists are using computers to understand text." Read here
- Sarah Jones's "When computers read: Literary analysis and digital technology." Read here
- Kath Bodes's blog post "A response to some responses." Read here
- Stéfan Sinclair and Geoffrey Rockwell's chapter "Teaching Computer-Assisted Text Analysis: Approaches to Learning New Methodologies" in Digital Humanities Pedagogy: Practices, Principles and Politics (ed. Brett Hirsch). Read here

Other resources:

- Duke University 'Text Analysis slides'
- Stanford Literary Lab Pamphlets
- Ted Underwood's "Where to start with text mining"
- Tooling Up for Digital Humanities - Text Analysis

Pre-workshop requirements:

(e.g. programs to be downloaded or registrations established)

- None

Background and context:

What is Text Analysis?

Text Analysis refers to the process conducted on a textual dataset to extract meaningful information from it. The purpose of Text Analysis is to create structured data out of text content in order to interpret it.

All Text Analysis projects in the Digital Humanities require the curation of a dataset. This might be a single author's oeuvre, part of a magazine collection, or a grouping of poems from across a number of centuries.

"Text analysis practices encourage reflection on the questions asked and formalization of queries"
(Geoffrey Rockwell)

Some key methods and techniques of Text Analysis include: word counts, trends, keyword density, correlations, and topic modelling.

As with all Digital Humanities projects, the question of what makes for meaningful information is always open for discussion. For this reason, a good digital Text Analysis project should always start with a clear, compelling research question.

Text Analysis tools

There are a large number of tools available for conducting computer-assisted Text Analysis, each of which performs different functions and requires different skills in users.

- Some downloadable applications that require little to no programing skills include NVivo, Tableau and Cowo
- Easy-to-use tools that don't require programming skills include Voyant Tools, JSTOR Lab's Text Analyzer, Netlytic, and Wordle.
- Other Text Analysis tools require small amounts of command line usage. Some examples include Stanford's CoreNLP and MALLET, a tool that generates topic models.
- Some platforms that require enhanced levels of programming skills include NLTK and Bookworm, which tracks word frequencies over time.

This tutorial focuses on one out-of-the-box Text Analysis tool, **Voyant**.

Voyant Tools is part of a larger project, Hermeneuti.ca, which is a collaborative project by Stéfan Sinclair & Geoffrey Rockwell to think through computer-assisted text analysis for humanists.

Voyant Tools is a web-based text reading and analysis environment. It is designed to make it easy for your to work with your own text or collection of texts in a variety of formats, including plain text, HTML, XML, Pdf, RTF, and MS Word.

Method:

Series of steps to follow

1. Open Voyant Tools and have a look around

2. Experiment with putting different types of text into Voyant
    a. Type text directly into the box
    b. Copy & paste text directly into the box
    c. Open one of the two existing Voyant corpora (i.e. Jane Austen or Shakespeare corpus) by clicking on the 'Open' button
    d. Upload a file from your computer (individually or in a zip file to upload a number of separate files at once, such as chapters in a book or political speeches)
    e. Copy & paste a URL into the box

3. After loading a text, explore the 5 default 'skins'
    a. Cirrus: a word cloud showing the most frequent terms in a corpus
    b. Reader: an efficient corpus reader that fetches segments of text as you scroll
    c. Trends: a distribution graph showing terms across the corpus (or terms within a document)
    d. Summary: a tool that provides a simple, textual overview of the current corpus
    e. Contexts: a concordance that shows each occurrence of a keyword with a bit of surrounding context

4. Explore some of the following features by sliding the various toggles on each skin or clicking on the '?' button in the upper right hand corner of each pane
    a. Remove 'stop-words' by heading to settings in the top right hand corner of the Cirrus pane and clicking on 'Define options for this tool.'
    b. Explore 'Trends' by selecting or deselecting particular words (do this by clicking on the coloured dots)
    c. Change the 'Trends' tool to 'Document terms' in order to reveal information about specific words in the text
    d. Have a go at changing the Corpus tool to *WordTree*
    e. Then *export the URL* to make it larger in a different window
    f. We could also examine such things as:
        - Word frequency
        - Collocation (words commonly appearing near each other)
        - Concordance (the contexts of a given word or set of words)

- Entity recognition (identifying names, places, time periods etc.)
- N-grams (common two-, three-, etc. word phrases)
- Dictionary tagging (locating a specific set of words in the texts)

Sample activities:

Activity 1: Load an entire text from Project Gutenberg

Objective of this learning activity:

- Start thinking about how to upload, organise and analyse textual data
- Consider the pros and cons of Voyant's quick visualisation (What don't we see? What information might be missing?)

Five minute activity

Head to Project Gutenberg and locate the online book (HTML) of the full text of *Moby Dick*

Copy & Paste entire text or copy & paste URL and load the text into Voyant Tools

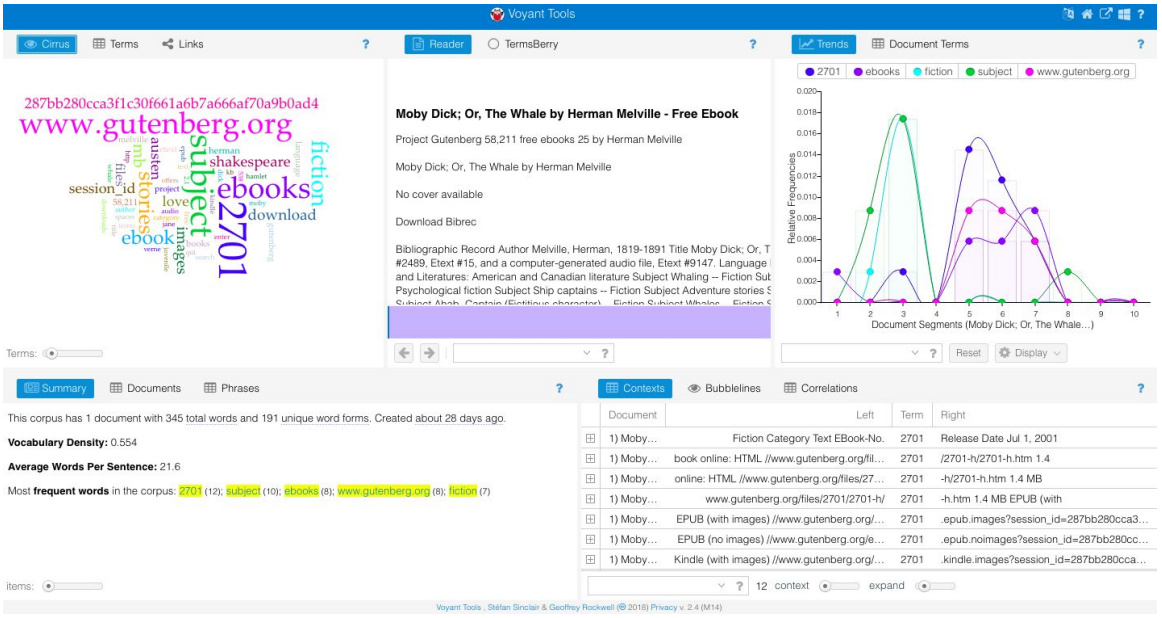

Fig. 1: Viewing the entire text of *Moby Dick* in Voyant Tools

Instructor to call on several participants to share first impressions/observations with the group. This encourages an environment of exploration and 'play' insofar as there are no immediate, correct answers.

Have participants examine the following components of *Moby Dick* in Voyant Tools:

- Look at the Word Cloud. What does this visualisation tell us immediately? What might it be suggesting that isn't entirely correct?
- Which features are metrical? (i.e. based on measuring the text in some way)
- How are the other features generated?
- What scope can you perceive to tweak the tool for better results?

Have participants reflect on this process. The instructor should also encourage participants to critique the tool itself. This discussion might be encouraged by asking the following questions or prompts:

- What information is missing?
- What information doesn't seem quite right?
- What kinds of conclusions can we draw? Are these supported by evidence?
- How would we explain this process to someone else?

Activity 2: Use Voyant to compare different texts

Objective of this learning activity:

- Learn how to use Voyant Tools to develop a 'compare and contrast' analysis of related but different corpora
- Consider additional steps that could be taken to maximise results (e.g. modifying the corpora, adding more texts, restructuring etc.)

Fifteen minute activity

Head to the tinker.edu.au environment and find the Voyant test instance.

Use the 'Open' tab on the Voyant test instance homepage to load the Tinker curated datasets. Select 'Election Speeches' and load the corpus.

Have participants examine the following components of Election Speeches in Voyant Tools:

- Explore the 'Summary' skin and discuss the significance and utility of the 'Documents' section
- Explore the 'Trends' skin in more detail now that we have different texts within our corpus. Are the texts chronologically structured? If so, what new analysis does this afford?
- After participants have explored the speeches in Voyant, have them report back on the perceived limitations of this dataset
- Have participants report back to the group (in pairs) on what they would add to the corpus (and how) in order to produce more sophisticated results

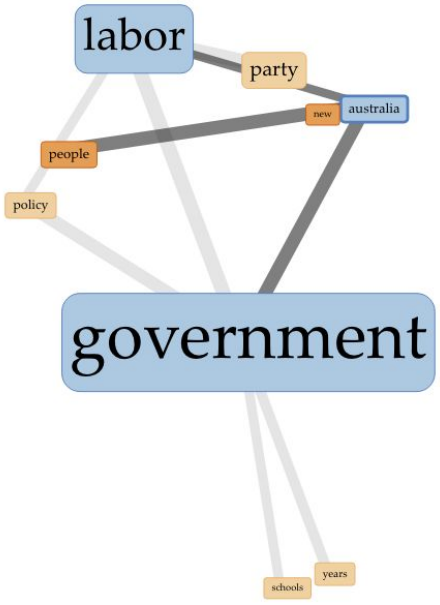

Fig. 2: Using the 'Links' function for visuaing Election Speeches in Voyant Tools

Next steps:

Further research data management resources available through Tinker include:

10 HASS Data Things

An adaptation of the ARDC 23 Research Data Things, this document gives researchers the opportunity to take the next step in learning about research data, with discipline relevant examples and self-guided learning activities. Dip into the 10 HASS Data Things here.