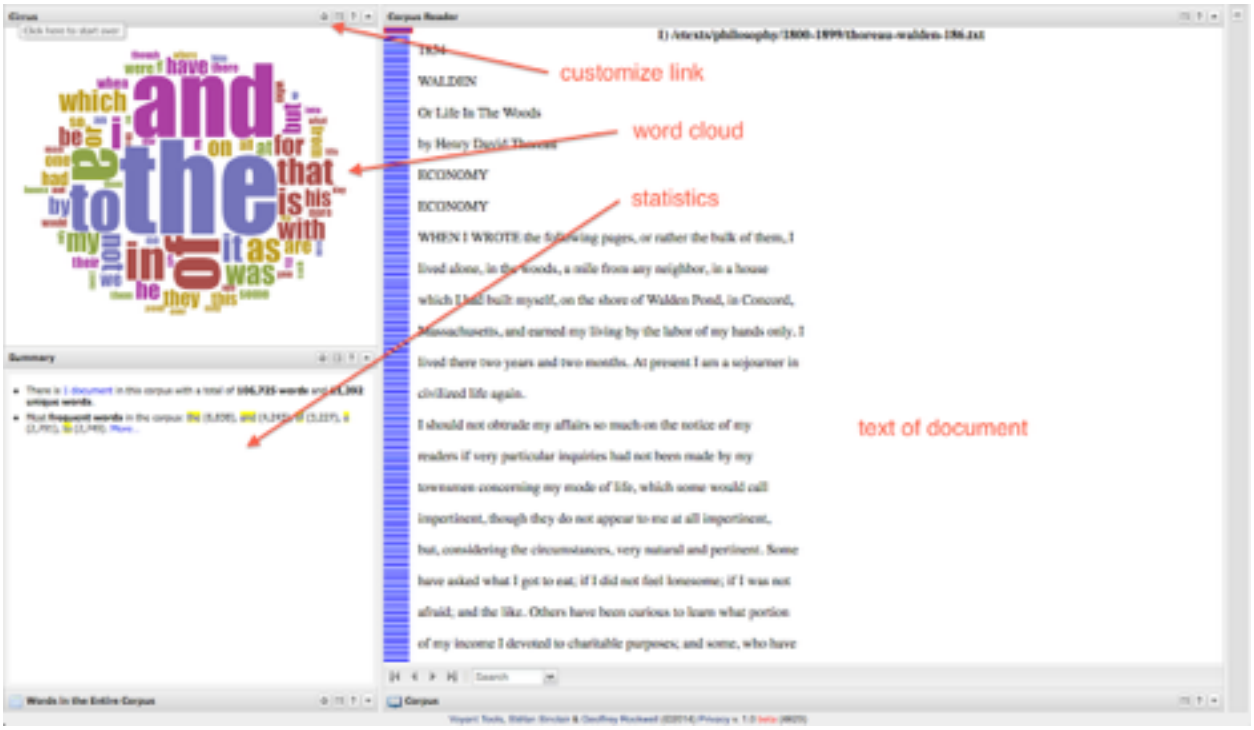
Simple text analysis with Voyant Tools

Voyant Tools is a Web-based application for doing a number of straight-forward text analysis functions, including but not limited to: word counts, tag cloud creation, concordancing, and word trending. Using Voyant Tools a person is able to read a document “from a distance”. It enables the reader to extract characteristics of a corpus quickly and accurately. Voyant Tools can be used to discover underlying themes in texts or verify propositions against them. This one-hour, hands-on workshop familiarizes the student with Voyant Tools and provides a means for understanding the concepts of text mining.

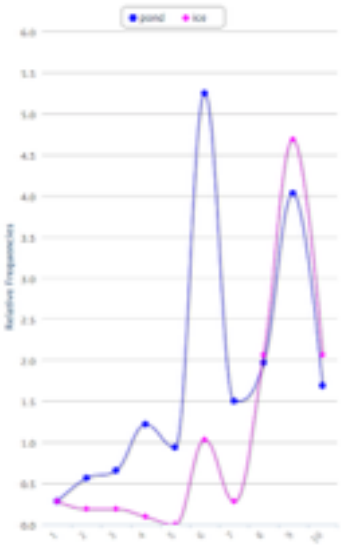
Getting started

Voyant Tools is located at http://voyant-tools.org, and the easiest way to get started is by pasting into its input box a URL or a blob of text. For learning purposes, enter one of the URL’s found at the end of this document, select from Thoreau’s Walden, Melville’s Moby Dick, or Twain’s Eve’s Diary, or enter a URL of your own choosing. Voyant Tools can read the more popular file formats, so URL’s pointing to PDF, Word, RTF, HTMl, and XML files will work well. Once given a URL, Voyant Tools will retrieve the associated text and do some analysis. Below is what is displayed when Walden is used as an example.



In the upper left-hand corner is a word cloud. In the lower-left hand corner are some statistics. The balance of the screen is made up of the text. The word cloud probably does not provide you with very much useful information because stop words have not been removed from the analysis. By clicking on the word cloud customization link, you can choose from a number of stop word sets, and the result will make much more sense. Figure #2 illustrates the appearance of the word cloud once the English stop words are employed.

By selecting words from the word cloud a word trends graph appears illustrating the relative frequency of the selection compared to its location in the text. You can use this tool to determine the consistency of the theme throughout the text. You can compare the frequency of additional words by entering them into the word trends search box. Figure #3 illustrates the frequency of the words pond and ice.



Once you select a word from the word cloud, a concordance appears in the lower right hand corner of the screen. You can use this tool to: 1) see what words surround your selected word, and 2) see how the word is used in the context of the entire work. Figure #4 is an illustration of the concordance. The set of horizontal blue lines in the center of the screen denote where the selected word is located in the text. The darker the blue line is the more times the selected word appears in that area of the text.

What good is this?

On the surface of things you might ask yourself, “What good is this?” The answer lies in your ability to ask different types of questions against a text -- questions you may or may not have been able to ask previously but are now able to ask because things like Voyant Tools count and tabulate words. Questions like:

- What are the most frequently used words in a text?
- What words do not appear at all or appear infrequently?
- Do any of these words represent any sort of theme?
- Where do these words appear in the text, and how they compare to their synonyms or antonyms?
- Where should a person go looking in the text for the use of particular words or their representative themes?

More features

Voyant Tools includes a number of other features. For example, multiple URL's can be entered into the home page's input box. This enables the reader to examine many documents all at one time. After doing so many of the features of Voyant Tools work in a similar manner, but others become more interesting. For example, the summary pane in the lower left corner allows you to compare words across documents. (Consider applying stop words feature to the pane in order to make things more meaningful.) Each of Voyant Tools' panes can be exported to HTML files or linked from other documents. This is facilitated by clicking on the small icons in the upper right-hand corner of each pane. Use this feature to embed Voyant illustrations into Web pages or printed documents. By exploring the content of a site called Hermeneuti.ca (http:/hermeneuti.ca) you can discover other features of Voyant Tools as well as other text mining applications.

The use of Voyant Tools represents an additional way of analyzing text(s). By counting and tabulating words, it provides a quick and easy quantitative method for learning what is in a text and what it might have to offer. The use of Voyant Tools does not offer "truth" per se, only new ways at observation.

Sample links

[1] Walden - http://infomotions.com/etexts/philosophy/1800-1899/thoreau-walden-186.txt
[2] Civil Disobedience - http://infomotions.com/etexts/philosophy/1800-1899/thoreau-life-183.txt
[3] Merrimack River - http://infomotions.com/etexts/gutenberg/dirs/etext03/7cncd10.txt

---

Eric Lease Morgan <emorgan@nd.edu>
University of Notre Dame

January 17, 2014