Distant Reading with Voyant Tools

Theme analysis

We gaan naar een erg humanistische activiteit kijken vandaag, anmelijk het analyseren van thema's in literatuur. En we gaan bekijken hoe een tool als *Voyant Tools* ons daarbij kan helpen.

In theme analysis, a literary critic identifies a meaningful pattern in a single text or across multiple texts. As M. H. Abrams elaborates in his landmark Glossary of Literary Terms, “Theme is sometimes used interchangeably with ‘motif,’ but the term is more usefully applied to a general concept or doctrine, whether implicit or asserted, which an imaginative work is designed to incorporate and make persuasive to the reader”.

The most common contemporary understanding of theme is an idea or point that is central to a story, which can often be summed in a single word (for example, *love*, *death*, *betrayal*). Typical examples of themes of this type are *conflict between the individual and society*; *coming of age*; *nostalgia* etc.

Traditionally, a critic identifies themes by reading carefully, assembling various clues that suggest how a certain topic organizes or underpins a text. These clues might include epigrams, allusions, sources of conflict between characters, key words in the narration or dialogue, prominent symbols, or issues repeated from the author’s other works or culled from the text’s historical contexts. This theme —this meaningful pattern— is then traced throughout the text(s) to form the foundation for a critical interpretation.

A theme may be exemplified by the actions, utterances, or thoughts of a character in a novel. An example of this would be the *flowers* in Mrs. Dalloway by Virginia Woolf, which are used to represent the passage of time and the inevitability of death, but also joy and beauty.

Alrightm now, how do we go about analyzing themes in literature?

Simply put, we have the following steps involved in theme analysis:

- You have some kind of qualitative data, in this case, a text or a set of texts.
- You identify a meaningful pattern or a code in the text or texts. And code it as a theme.
- You then trace this theme throughout the text(s) to form the foundation for a critical interpretation. This is an iterative process, and you may need to go back and forth between the text and your interpretation.

Instances of the theme are collected, collated, and supplemented with research. The critic then uses inductive reasoning to generate a thesis about the theme’s significance for the text as a whole. By documenting a pattern and analyzing it as a theme, what might have started as an isolated observation (“There sure are a lot of conversations about "flowers" in this text!”) is transmuted into a sophisticated argument.

Digital Theme Analysis

Throughout the history of literary criticism and theory, scholars have developed a **variety of methods** for **identifying and interpreting themes** in literature. These methods are often associated with particular schools of thought, such as New Criticism, Marxism, feminism, queer studies, ciritcal race studies.

You could call these methods **"interpretive modes"** for identifying themes; there is some kind of **preoccupation** with a particular aspect of the text, and the critic uses this preoccupation to craft a unified reading of the text.

For example, **New Criticism** dwelled on themes highlighted by the text's stylistic properties in order to craft a unified reading of the text; **form becomes the bearer of the theme**, telling the reader how to resolve ambiguity or dissonance in the text's thematic registers.

In other words, **thematic analyses of literature have always depended on some form of filtering** that efficiently processes texts -- it is all about the control of specific variables, and the iterative process of identifying those variables, creatively deform texts.

This is what Franco Moretti was also doing when he was trying to identify specific character interactions in Hamlet.

Theme analysis falls into the realm of digital when computational tools are employed to extract recurring patterns from texts in an automated manner.

Now, our question for today is: **can we use computational tools to Distant Read texts, and extract recurring patterns from texts in a kind of automated manner?**

Now, using computers to do this kind of work is not new. In fact, it has been around for a long time. But the digital age has brought new tools and methods to the table, and it's worth taking a **closer look at how we go about this analysis, the benefits and pitfalls it involves**, and how the digital age's approach to themes is both similar to and different from the traditional ways of doing things.

For example, as a useful approach to **completely unfamiliar texts**, digital theme analysis can aid in the earliest stage of analysis, namely, in the basic selection of the works by theme, which can then be used in a more sophisticated research project.

Alternatively, a critic who has already begun analog theme analysis might turn to digital methods to clarify a theme's relationship to other themes or discover moments in the text when the theme proliferates or attenuates.

Or a critic might use digital methods after drafting an article in order to create corroborating charts and figures.

All of these examples share a common practice: the critic at some point relies on computational tools that distant-read texts to locate, document, or analyze theme in a literary text or corpus.

Exploratory Data Analysis

- The analysis of datasets based on various numerical methods and graphical tools.
- Exploring data for patterns, trends, underlying structure, deviations from the trend, anomalies and strange structures.
- It facilitates discovering unexpected as well as conforming the expected.
- Another definition: An approach/philosophy for data analysis that employs a variety of techniques (mostly graphical).

- Based on insights developed at Belll Labs in '60
- Techniques for visualizing and summarizing data sets
- What can data tell us? In contrast to conirmatory data analysis.
- Developed by John Tukey, a statistician at Bell Labs
- He was interested in the data analysis process, not just the results
- Introduced many basic techniques, especially for visualizing distribution of data:
    - The 5 number summary: The five-number summary is a set of descriptive statistics that provides information about a dataset. It consists of the five most important sample percentiles:
- ```
  the sample minimum (smallest observation)
  the lower quartile or first quartile
  the median (the middle value)
  the upper quartile or third quartile
  the sample maximum (largest observation)
  ```
    - Box plot
    - Stem and leaf plot

Founding chairman of the Princeton statistics department in 1965.

While working with John von Neumann on early computer designs, Tukey introduced the word bit as a portmanteau of binary digit. The term bit was first used in an article by Claude Shannon in 1948. and the first published use of the word software.

"finding the question is often more important than finding the answer. Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there."

Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

The goal of EDA is to open-mindedly explore data. Tukey: “Unless the detective finds the clues, judge or jury has nothing to consider. Unless exploratory data analysis uncovers indications, usually quantitative ones, there is likely to be nothing for confirmation data analysis to consider.”

Here, judge or jury is a confirmatory data analysis Tukey: Confirmatory data analysis goes further, assessing the strengths of the evidence. With EDA, we can examine data and try to understand the meaning of variables.

Maximize insight into a dataset Uncover underlying structure Extract important variables Detect outliers and anomalies Test underlying assumptions
  ```

EDA: no hypothesis at first, generate a hypothesis, use graphical methods mostyl CDA: start with the hypothesis, test the null hypothesis, use statistical models

Case Study using Voyant Tools

Google's n-gram viewer is a way to analyze the rise and fall of specific themes e.g. You could also use topic models, to automate the discovery of themes across tens, hundreds, or tens of thousands of texts.

In some other contexts, TEI-XML may be used to encode particular themes in a digital edition.

To begin elucidating interpretive moves a critic might make to engage in digital theme analysis, let us turn to one popular, easy-to-approach tool: Voyant Tools.

Interestingly, if you look up DH on wikipedia, Voyant Tools is shown as a key example of one of the tools.

The importance of Voyant to the Digital Humanities cannot be understated. While advanced scholars in the field prefer to use more targeted software packages, construct databases instead of submitting their textual data into a browser-based tool, and/or to develop their own methods of distant reading, Voyant Tools is a popular tool for the classroom, for the general public, and for entry-level DH students and scholars. (And some project developers even run Voyant locally to bypass the website altogether, neatly obviating this drawback.)

Advanced tools used for text analysis often connect with the language R, which is designed for statistical analysis and is also readily connected to libraries of visualizations. Any researcher serious about data mining and text analysis will need to commit to learning these programs, but to experience text analysis without the programming skills, Voyant is extremely helpful.

Voyant was developed by Stéfan Sinclair (McGill) and Geoffrey Rockwell (University of Alberta), whose book *Hermeneutica: Computer-Assisted Interpretation in the Humanities* documents and reflects upon their creation of Voyant.

Voyant Tools is a **free, open-source application** that works in-browser. It combines a variety of simple text-mining tools in a single, intuitive graphical user interface, divided into individual windows or panes, which the user may adjust at will. Users may change the relative size of the individual panes, their positioning on the screen, and even the tools displayed on the individual panes.

What is extremely useful is that **some tools dynamically respond to one another**, altering their displays to stay current with the particular term or textual phenomenon that the user is exploring on another pane. In addition, users can customize key features related to each tool, such as by defining stopwords (words left out of tool calculations, such as common articles, pronouns, and prepositions, because they are not relevant for analysis).

Anatomy of the Voyant Dashboard

The default view presents a word cloud on the top left (the “Cirrus” tool). **“Cirrus”** may be easily switched, with a single click, to

- **“Terms**,” which is a simple concordance,
- or **“Links**,” which generates a simple network graph that shows the relationships between the most frequently used words in the text or texts the user has uploaded.

- In the top middle is a simple copy of the text(s) the user submitted (the **“Reader**,” which will automatically navigate to relevant passages as the user engages with other panes);
- users can click quickly to change this pane to **“TermsBerry**,” a cute name for a kind of network visualization that depicts relations between common terms by representing common words as a berry-shaped collection of circles that become colorized when the user mouses over the term or a related one.
- At the top right is **“Trends**,” a time-series graph that displays in the form of a familiar X-Y scatter plot to reveal how the frequency of a term changes over the course of the text;
- it may be replaced with another single, simple click with **“Document Terms,”** a searchable concordance
- that also displays a thumbnail representation of the **“Trends”** data for the relevant term.
- On the bottom left is the **“Summary”** pane, which reports the total number of words, the number of unique word forms, the vocabulary density (a measure of the variety in the author’s diction, generated by dividing the total number of words by the number of unique word forms), the average number of words per sentence, and the top five most frequent words.
- It may be quickly replaced with **“Documents**,” which allows users to sort through multiple texts if they are working with a large corpus (that is, more than one text),
- or **“Phrases**,” a tool that, like topic modeling, identifies multiple groups of commonly co-occurring words.
- On the bottom right, the **“Contexts”** pane displays the words that immediately precede and proceed from the terms that the user is currently investigating on other panes; this means that users need not flip through the text on another browser tab or a physical book.
- “Contexts” may be replaced with **“Bubblelines**,” which uses proportionally sized circles that correlate to word frequency to create lines of varying thickness (it therefore visualizes the same kind of information as the “Trends” pane, but in a more colorful, graphics-based format),
- or **“Correlations**,” an advanced analytical tool that locates similar patterns of co-occurring words (that is, groups of words that gain or lose frequency roughly alongside one another).

Many more tools beyond these reliable favorites are available on Voyant, including the new Veliza chatbot. Based on Joseph Weizenbaum’s classic natural language processing program ELIZA, trained to respond to textual inputs as a psychotherapist would, Veliza will generate responses to particular lines in your text, revealing how a psychiatrist might react to the characters’ thoughts and problems.

Except for Veliza, most of Voyant’s constituent text-mining tools use wordfrequency tabulations to visualize a text—that is, they count the most frequent words (MFW) used in text and generate some sort of chart or graphic, often interactive, that represents this frequency (most familiarly, the word cloud).

Though a critic may feel tempted to use a tool like Voyant simply to confirm hypotheses and immediately turn back to drafting a written argument, the data generated depend on the critic’s each and every input and interaction with the results. Consequently, visualizations therefore may not be replicable, so always save results for future consultation. To put it another way, just as with any tool used for digital theme analysis, each of Voyant’s affordances requires specific actions on the critic’s part so that it does not become a liability. For example, librarian Megan Welsh has critiqued it for a number of difficulties, including its uneven loading times and the lack of standardization preventing easy exportation (96-97).

Yet successful critics will anticipate such challenges and adjust their experimental design to reflect their needs and available resources, including text file format and reliability, relevance of producing “real-time” results (in pedagogical contexts), mode of dissemination (online versus hardcopy), and type of data

required (statistics, dataset, or image). Knowing the limits of any digital tool used for textual criticism is crucial for the early stages of digital theme analysis.

Regarding the present example, Voyant, the scholar should understand that word frequency lists are very good at suggesting the text's linguistic preoccupations, but less good at indicating word proximity, discerning between different denotations or connotations of the same word, or relating those preoccupations automatically to the text's plot or structure, to its historical context, or to word usages in comparison texts. This does not mean to avoid Voyant, but to supplement its results with other critical methods.

Using Voyant, and other text analysis tools, is most successful when applied to a text with which the researcher is familiar. Then the results can be gauged against prior understanding and assumptions.

Voyant Jargon

- Stopwords: words with little meaning ("and," "the," "a," "an") are usually filtered out when text mining. Stop words are the words in a stop list (or stoplist or negative dictionary) which are filtered out (i.e. stopped) before or after processing of natural language data (text) because they are insignificant.[1] There is no single universal list of stop words used by all natural language processing tools, nor any agreed upon rules for identifying stop words, and indeed not all tools even use such a list. Therefore, any group of words can be chosen as the stop words for a given purpose. The "general trend in [information retrieval] systems over time has been from standard use of quite large stop lists (200–300 terms) to very small stop lists (7–12 terms) to no stop list whatsoever".[2]
- Concordance: a list of words in a text, with their immediate context.
- Vocabulary density: a measurement of vocabulary usage in comparison to the length of a document. Think of how many words will be read on average before a new word is encountered. (For Moby Dick, a new word appears every 12 words!)
- Distinctive words: High frequency words relatively unique to a particular document (appears when comparing multiple documents).
- Correlation coefficient: comparison of relative frequencies of terms. Coefficient approaching 1 indicates that values correlate positively (they rise and fall together). Coefficient approaching 0 indicate little correlation. Approaching -1, terms correlate negatively (as one term rises, the other falls).

Case study: Jane Eyre

Our first case study shows how Voyant can be used to conduct a study of a text with a specific lens (theory) in mind. Let's say that a scholar is interested in an ecological reading of Charlotte Brontë's Jane Eyre and has already completed an initial round of analog close-reading. When submitted Voyant, with irrelevant data deleted (a preface not in the first edition, licensing information from the text available on Gutenberg) and stop words selected

the top 50 most frequent words apparently deal with interior themes—rather than ecological themes.

9. "room" (263 instances)
10. "house" (182 instances)
11. "door" (182 instances)

However, consulting the **“Word Trends”** box, which produces a line graph for individual word frequencies over the course of the book, yielded an interesting pattern in which “house” and “door” rise and fall rhythmically, with three distinct troughs that indicate that certain portions of the book that might deal more with the outdoors—suggesting not only that the novel fluctuates in privileging the outdoors, but also that the scholar should look in those “troughs” for relevant passages to close-read.

Before devoting time to closely examining these passages, it is necessary to filter out irrelevant passages by considering the multiple meanings of such superficially simple words. Does Brontë use “door,” a liminal object located on the boundaries between indoor and outdoor spaces, more frequently to indicate entering or leaving a house? Does she use the term figuratively or literally? Does she focus in interior or exterior doors?

Voyant’s **“Context”** window (which reveals the handful of words both before and after the instance of the keyword in the critic is interested) can help answer these questions. One of the most powerful functionalities in Voyant, this window allows the critic to combine distant and close reading easily and quickly. Consulting this window reveals that Jane is often being blocked from a space to which she desires entrance.

Thematically, doors are related to violence and alienation for Jane - a negatively connoted word not associated with freedom or happiness but with feelings of isolation and exile.

So far, if the critic wishes to make an ecological reading of the novel, Voyant suggests that it is not possible to support such a reading by establishing Jane’s unequivocally positive relationship to the outdoors. Connecting Voyant’s results to salient passages in the text shows that it is probably not possible to champion Jane as an environmentalist.

Reflecting on the novel’s major transitional moments, the critic then locates passages in the novel regarding this proposition: the book opens with Jane being pleased that **“there was no possibility of a walk that day”** (Brontë 2006, 9), and Brontë underlines the horror of Jane’s exile from Rochester by showing Jane weak and hungry from her few homeless nights spent on the moors.

Further down the word frequency list are “home” and “nature” in close proximity, yet inspecting the sentences in which such words are found, “nature” appears to denote temperament, personality, type, or a “state of nature” rather than to animals, plants, or the outdoors.

119. home (80 instances)
120. nature (78 instances)

It is only further in the word frequency list that a sense emerges of Jane Eyre’s focus on nature as an elemental force and as a subject for artistic representation:

176. air (60 instances)
177. wild (59 instances)
178. wind (48 instances)
179. sky (43 instances)
180. moon (43 instances)
181. wood (40 instances)
182. trees (38 instances)
183. rain (35 instances)
184. sun (32 instances)

185. garden (32 instances)
186. flowers (32 instances)

At this point, only after carefully sifting through Voyant's word frequency visualizers, a mature theme analysis of nature in Jane Eyre—one that eschews the politically inflected idealism that analog theme analysis might seem to support—begins to emerge.

To pursue Jane's ecology of aesthetics, the scholar could then use *Voyant* to search for "window" and "painting," as both focus on natural subjects as something to perceive, as well as their derivatives ("windowseat," "bow-window," "paint," "draw") and a word relevant to both: "frame."

- window
- painting
- frame

This investigation yields a constellation of significant passages to analyze, which the critic will then connect to other critical readings of the text, and it also yields ideas for further digital analysis, such as topic modeling (to check whether natural imagery and artistic/window imagery cluster), statistical analysis (to verify the statistical significance of these word counts), or n-grams (to compare Jane Eyre to other texts).

Significantly, these words were all suggested not directly from these Voyant results, but rather from dipping in and out of the text, reading passages indicated by Voyant, emphasizing the degree to which closely interacting with the text and existing literary criticism is still necessary to transform the quantitative data into mature theme analysis.

Case study: Frankenstein

Our second case study illustrates how one may approach a text **even if one has not engaged in any systematic close reading** of it.

Consider Mary Shelley's novel *Frankenstein*, which has been adopted by modern culture through countless remakes. This novel is colloquially associated with the story of the "monster," the "ungodly" evil creature that turns against his creator, Victor Frankenstein.

We would therefore expect to find words pertaining to monstrosity, science, horror, and revenge. To test these (ostensibly) predominant themes, the critic might run the text of the novel through Voyant.

What immediately stands out are the numerous repetitions of words relating to sentiments, which significantly outnumber those related to horror, science, or alchemy. In fact, the predominant words, when viewed in the "Context" window, portray a humanoid with deep emotions (or at least a strong desire for them) rather than a violent drone.

The following MWF capture these emotions: "felt" (79), "feelings" (76), "heart" (76), "dear" (72), "love" (59), "feel" (50), "hope" (50), "happiness" (49), "happy" (46), "joy" (41), "affection" (40), "good" (37), "pleasure" (37), "soul" (36), "spirit" (34)"gentle" (34), and "kind" (34). Solely considering the plot of Shelley's Frankenstein yields a story of murder and loss, yet the word counts stressing this theme are, in fact, relatively minor.

The critic's next step is to consult extant criticism. One representative example is Maurice Hindle's "Vital Matters: Mary Shelley's Frankenstein and Romantic Science," which argues, "Few I think would fail to use

the word ‘scientist’ to describe the monster’s creator, whether they had read Mary Shelley’s novel or not. Yet if one turns to the text of the book, this word is nowhere to be found [...]. The fact is, the word ‘scientist’ had not even been coined in 1818” (Hindle 1990, 29).

This is not a fluke but rather the first step in recognizing that we cannot import an ahistorical reification of science as we know it today. A Voyant user could advance Hindle’s argument by identifying a different theme for analysis—one that is justified by digital theme analysis. For example, the two most frequently used words are “man” (131) and “life” (115), while the words “creature” (66) and “monster” (32) are evoked far less frequently than “man” (131) or”human” (71).

At the same time, it quickly becomes apparent that nature, through human sensory perception, is one of the creature’s primary sources of acquiring knowledge: “country” (54), “nature” (53), “sun” (45), “world” (45), “scene” (44), “ice” (42), “light” (37), “mountains” (37), and “earth” (36) — an empirical approach, to be sure, but one that lacks a sci-fi element associated with the genre when tracing the creature’s origins and the language used to describe his worldview.

Rather, one of the MFWs used by the creature when contemplating his origin is “father” (113), indicating Victor’s role as the absent “father” from his life. This is closely followed by the word “eyes,” (102) where by looking in the “Context” window, the critic finds that it is the primary lexical connection between creator and creation.

This is even more poignant because the creature’s inability to engage with others renders sight the primary basis of his social life. Looking at Voyant’s density graph, which traces the frequency of abstracted word repetitions across a text and allows the critic to experiment with chunking the text (divide it into sections for easier interpretation), it becomes evident that many similarities link Victor and his creature.

Strikingly, many quotes are almost identical, such as one of the creature’s closing sentences: “I cannot believe that I am the same creature whose thoughts were once filled with sublime and transcendent visions of the beauty and the majesty of goodness” (Shelley 1992, 200).

Victor indulges in similar contemplations because he blames himself for creating a “monster,” emphasizing how the two defend the same humanistic values.

This digital theme analysis shows that Shelley’s Frankenstein does not prioritize the language of horror and science, but a language of humanism. Further, it does not suggest that Frankenstein is a horror story in the modern sense; in fact, there is a noted lack of words related to violence. The horror is located in the monster’s appeal to our emotions and logic and in the unexpected closeness of his state of mind and intentions to Victor’s, which makes the critic question who the monster really is—demonstrating how digital theme analysis moves quickly between quantitative data and the larger questions of traditional criticism.

- In groups of 4 Pick a dataset
    - Historical Cookbooks
    - Collected Works of Shakespeare
    - Jane Austen's Novels
    - Inaugural President Speeches
    - SF literature
    - Taylor Swift Lyrics (Lady Gaga, Rihanna, Beyonce, Katy Perry)
    - Social Media

    - Postcards

- Task: perform a distant reading of the dataset

    1. Do some exploratory data analysis. play around with the different panels. Become familiar with the data.
    2. Identify the most common words, phrases, and words in the context of other words
    3. Find a research question! What are you interested in? What do you want to know about the dataset?
    4. Consider the relevance of three different panels: X, Y and Z
        - How do they inform? Helpful or not?
    5. Add a stop word list
    6. Share your findings but more importantly, your experience!

For people who did't attend:

- Watch YouTube video on Voyant Tools
- Explore the Voyant Tools website
- Choose one of these two datasets:
    - Jane Austen's Novels
    - Inaugural President Speeches
- Think about the following questions:
    - What are the most common words?
    - What are the most common phrases
    - What are the most common words in the context of other words?
    -