The Digital Humanities Coursebook

An Introduction to Digital Methods for Research and Scholarship

Johanna Drucker

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

7 Data mining and analysis

7a Data mining and text analysis

The idea of data mining seems far removed from the humanities and its use for analysis of literary and aesthetic objects has prompted many immediate and strong responses, such as the claim that "literature is *not* data" (Lamarche 2012). Made as part of a larger diatribe against the digital humanities, the defensive posture suggests that data mining was meant to replace methods of reading that have a long history in cultural practice. More reasoned and informed discussions from within the practice of digital humanities have argued for the value of these techniques, particularly given the unprecedented scale of cultural materials available in digital formats (Kirschenbaum n.d.). What can data mining offer to the humanities that augments traditional methods without replacing them?

Data mining is an automated analysis that looks for patterns and extracts meaningful information in digital files (Underwood 2017). While it is not limited to analysis of so-called "big" data, it is particularly useful at large scales. Data mining has long been incorporated into the natural and social sciences. It has become a part of research methods in text, music, sound recording, images, and multimodal communications studies with tools customized for these purposes. Text analysis is a specialized subset of data mining that focuses on analysis of language (Schmidt 2013). Keep in mind that data mining always processes digital files, which means that for analog originals, these are surrogates, representations whose properties might be quite different from those of their source. For instance, a scan of a sculpture that reduces it to two dimensions only preserves some of the original information about form. Data mining only takes place on the information literally in the file, so clarification about the process is essential.

One advantage to data mining is the analysis of cultural materials in their native format—as texts, images, and media. But much processing has to occur before text or image analysis can proceed. Concepts like "distant reading" and the more pejorative "not reading" have arisen to describe some of these approaches. As with any method or technique, the question of value is best posed by seeing what these methods add to existing and/or traditional approaches, rather than dismissing them out of hand—or embracing them with uncritical enthusiasm. One easy to use example of text analysis applied to the corpus of Google Books is the Ngram Viewer, which displays all of the many problems and some of the benefits of these approaches. [See: Exercise #1 Google Ngram Viewer.]

Beginnings of automated textual scholarship

A milestone frequently cited in early digital humanities projects is the work of Father Roberto Busa. He was engaged with text analysis in the form of a concordance—a list of all words in a work or body of work. This was a form of analysis with a long history within religious and classical scholarship, but Father Busa's project was ambitious intellectually as well as logistically. He was focused on the concept of "presence" in the Latin texts of the 13th century scholar Thomas Aquinas. This was a metaphysical concept, and thus had no simple literal meaning.

Busa had tracked the instances of the words *praesens* and *praesentia* to address their contexts in the 1940s (Busa 1980). He created thousands of index cards for individual instances and the phrases in which they were found. When he realized that the full corpus exceeded ten million words, he began to consider mechanical aids. This led him to a collaboration with IBM, thanks to the support of its CEO Thomas Watson. Many of the approaches Busa designed for his project, such as identifying text types (for example, the use of citations) have become part of standard markup and statistical analysis. Working with punch cards and a list of typological codes, Busa established a systematic approach to the analysis of natural language, including linking all forms and versions of a word to its root (Terras and Nyhan 2016). Busa was working in analog materials but developing formal methods compatible with automated processes.

A second area of early automated text analysis was in the area of stylometrics or stylometry (Hai-Jew 2015). Longstanding debates about whether or not William Shakespeare was the author of all of his plays, or whether some were actually composed by Christopher Marlowe, remained pressing matters through the 20th century (Fox, Ehmodea, and Charniak 2012). The idea that statistical approaches could be brought to bear on the problem motivated formal analysis of style. Sentence length, grammatical structure, vocabulary choices, and other features of the texts were used to make comparisons. Many of the features on which style was formally addressed had a history in analog scholarship. The task of making formal parameters on which to analyze style is a useful intellectual exercise, as is any other attempt to make explicit parameters on which to formalize traditional humanistic approaches for computational purposes.

Methods for statistical processing are now far more complex than the counting and sorting of words into lists that were central to Busa's early project or the techniques developed for analysis of style (Sculley and

Pasanek 2008). Current tools and platforms combine counting and sorting techniques with statistically driven capacities. The differences between these will be a recurring theme.

Fundamentals of data mining and text analysis

Data mining is not limited to texts. Many applications for extracting meaningful statistical information from humanities materials operate on quantitative data, by which is meant information that originates in numerical form. Quantitative research is central to the social sciences, and not surprisingly, historians frequently adopt such methods for analysis of numerical data from economic records (taxes, census records, demographic analyses of social groups, and so forth). Quantitative history is generally considered to have originated in the 1960s, in a shift to studies of broader populations and trends, and away from the conventional focus on political history and events among leaders and elites (Guldi 2018). This continues to be a motivation within digital humanities where patterns of events, not just individual accomplishments, are examined. Increased scope in collections allows once marginal figures and works to come into view and to be accessed. The availability of computational tools, desktop and then laptop computers, to create and process data made quantitative methods integral to the history field. More than half a century later, these methods have been codified into a set of highly flexible and useful tools that can be readily acquired and used across the humanities.

At the heart of data mining and text analysis are several processes that should be understood critically. These are the same processes noted earlier: parameterization and tokenization. These identify what can be counted and how the counting is done. Additional considerations come into play with data mining, which are the statistical analyses of frequency, proximity, and value of individual data points within the larger sets. Principles like *collocation* of words are judged relative to other usage—and in contrast to the sum of all other words in a sample. In other words, multiple factors go into determining how any individual word is valued, not just the number of times it appears. The difference between counting and statistical analysis is brought into focus by this contrast.

In text analysis, questions of proximity and frequency are gauged against the statistical probability of occurrence in relation to all of the words or instances in a work or corpus (depending on the boundaries set for the analysis). Thus, the frequency of occurrence of a word is not simply counted but calculated in relation to the frequency of other words—and to the likelihood or probability of its being used. These are considerations that are rarely part of humanities reading practices. Understanding the workings of automated processes is essential for engaging seriously with this work.

Sample size and testing procedures are central issues in statistics, and they are brought directly into humanities work in data mining and textual analysis (Delice 2010). Questions of generalizability and reliability determine the extent to which the results of an analysis can be applied outside the single sample (Sandelowski 1995). For example, in analyzing the relationship between class position and letter-writing in a particular period, are the records of a historical society in New England and another in the pre-Civil War South comparable? How many of the letters need to be analyzed? What if there are two hundred by one author and only two by another—do they carry the same weight? On what terms should the contrasts between the language of these authors be assessed? How representative are these letters—and representative of *what?* They can only give an idea of what has been preserved, but can the analysis be extrapolated to discuss education, literacy, and gendered language across a small segment of the population?

As we know, data in the cultural record is notoriously incomplete. The estimate is that of the 30,000 novels that were published in English in the 19th century, only 6,000 of these have survived, of which less than 300 are considered canonical. How are traits or features of these works to be assessed in relation to the wider reading experience of a past century?

Text analysis and distant reading

The term distant reading was invented in about 2000 by Franco Moretti, the scholar who was central to its development for literary study. (Serlen 2010) Distant reading is the idea of processing content—subjects, themes, persons, or places—or information about publication date, place, author, or title in a large number of textual items without engaging in the reading of the actual text. Could texts be "read" at a scale that exceeded human capacity? The "reading" is a form of data mining that allows information in the text or about the text to be processed and analyzed.

Debates about distant reading range from the suggestion that it is a misnomer to call it reading, since it is really statistical processing and/or data mining, to arguments that the analysis of the corpus of literary or historical (or other) works has a role to play in the humanities (Underwood 2017). Proponents of the method argue that text processing exposes aspects of texts at a scale that is not possible for human readers and which provides new points of departure for research. Patterns of changes in vocabulary, nomenclature, terminology, moods, themes, and a nearly inexhaustible number of other topics can be detected using distant reading techniques, and larger social and cultural questions can be asked about what has been included in and left out of traditional studies of literary and historical materials.

Moretti's earliest automated work was focused on questions of genre—the type of text produced in literary publications. The initial decisions about how to characterize genre—distinguishing mysteries from romance, didactic works from sentimental ones, and so on—had to be done with close reading techniques and conventional methods. Certain vocabulary words and terms, phrases, and other textual features were selected as markers of genre. This was done by human judgment, and these decisions play a major role in the way the automated processes unfold. Then a large corpus of digitized materials was analyzed using these terms in order to sort them by genre. Moretti described the process as one of reduction "to a few elements" that could be abstracted "from the narrative flow" (Serlen 2010). As in all such work, the outcome is only as good as the model on which the data are abstracted. Terminology shifts and changes. Usage also transforms the meaning of individual words over time. And characterization of vocabulary terms always involves some judgment.

One of the positive claims for distant reading was that it might shift focus away from the established canon, allowing for a broader insight into cultural patterns of reading. The recovery of a wide swath of ignored titles from the oblivion into which habits of teaching and critical writing had cast them promised to address class issues, and to some extent, those of race. But the works that are digitized in Hathi Trust, Google Books, and other major repositories cannot recover fully lost texts, only make available a wider array than the narrow canon that had formed a core of literary study. Problems in the relationship between distant reading and critical race studies have been brought up by Richard Jean So and Edwin Roland (So and Roland 2020). Distant reading requires explicit quantification, and the terms of racial identity, only sometimes explicit in authors, characters, and texts, have to be defined to expose the limits of the ways computational approaches have been formulated as well as to expand the possibilities for writing the histories of people of color as cultural figures.

In a large-scale project, for instance, to characterize sentimentality and its relation to bestsellers, Andrew Piper and Richard Jean So made use of a library of terms put together by a scholar, Bing Liu, who had produced lexicons of terms characterized according to their "positive" or "negative" sentimental value. They processed their corpus of digitized literature by looking for these words (negative terms were *abominable* or *shady* and positive ones were *admirable*, *courageous*, or *rapturous*, etc.) to see where different genres and also prize-winning works were situated relative to each other. Another example was a study of "mood words" in 20th century fiction by decade, to see if any correlation existed between large scale events like economic downturns, wars, or periods of prosperity and the tone of novels (Acerbi et al. 2013). The results are fascinating, but not altogether surprising. The "sadness" score was greatest during the years of the Second World War but, oddly, "joy" was highest during the period of the First World War.

Research results are generally shown in graphs that aggregate information. This raises questions about how outliers and anomalous individual works are treated. If a single highly popular work had a particular sentiment attached to it that was out of sync with the mainstream, it would be lost in the smoothing and averaging processes of data analysis. Also, something as basic as a publication date for a work can be complicated. Leo Tolstoy's mid-19th century novel, *War and Peace*, is still in print, the works of William Shakespeare were widely read in English throughout the 20th century, and so were *The Holy Bible* and the first Harry Potter book. Are these 20th century works? Are they part of the corpus because they were read? How is the volume of consumption measured? School texts commonly supplied the Shakespeare plays. Tolstoy's novel may well have been borrowed from a public library. The Bible texts were likely to have been available in the home, perhaps even in a family heirloom handed from generation to generation. What are the dates of publication for such works? How is their reading registered in relation to that of market records for commercial publishers of Harry Potter titles?

A critical framework for discussion of results of any data mining and text analysis is required to keep it from standing on its own, decontextualized. While patterns emerge, and large trends can be discerned, the question of what these are indicative of remains. Do they only show trends in the *data*? Or can they reveal trends in *phenomena* of the actual and lived world (Lee 2019)? What is the relation between these two? Abstraction, extraction, reduction, and simplification are frequently used terms in discussing large data sets. The value of the research needs to be measured against these considerations. As in so many aspects of digital work, the value is in the dialogue with traditional methods. Distant reading often points out areas where close reading will be of value.

Tools for text analysis

One platform for exploration of text analysis, Voyant, was developed by humanists, Geoffrey Rockwell and Stéfan Sinclair. Voyant is meant to be useful without technical expertise and is freely available for use online (no downloads and no learning curve). More advanced researchers will use Mallet and the Natural Language Toolkit, both built in the programming language Python. These are tools that can be trained on a text, with results modified by the user by eliminating some results and strengthening others, until the outcome conforms to the research goals. 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. [See Exercise #2: Voyant and Mallet.]

Voyant is a dashboard-based platform. This means its various functions and tools can be accessed simply by going to the site and entering text or URLs directly without any pre-processing or specialized knowledge. The tools automatically output the results into a set of screens. Voyant, as per its tagline, "Sees through your texts," and offers an array of visualizations through which to get results. The processing happens out of sight, and the various panels display a range of visualizations, such as WordClouds, TermsBerries (a cute name for a cloud of circles bouncing around next to each other), Bubblelines, columns, bar charts, area charts, and so on. The tools are particularly useful for comparisons, and filtering terms within works, narrowing the segments, and focusing on individual terms produces correlated visualizations. So, clicking on a single circle in a "berry cloud" changes what is shown in the trends window next to it. Correlation is one of the crucial features of data analysis, and Voyant is designed to facilitate these connections.

Voyant also offers an introduction to topic modeling in one of its panes. This feature clusters terms according to their connection (frequency and proximity) in the document(s) as a whole and exposes themes within a work or corpus. Keywords in context, or KWIC, provide another useful comparison tool for "reading" a corpus through crucial themes. Keywords that repeat are often clues to themes or topics in a body of work and seeing how the terms are resituated is a useful tool for seeing the facets of an argument to which the term is central. 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.

Topic modeling and advanced text analysis

Topic modeling is what it sounds like—a model of a topic that appears in a text. In more advanced work that uses machine learning tools, the program

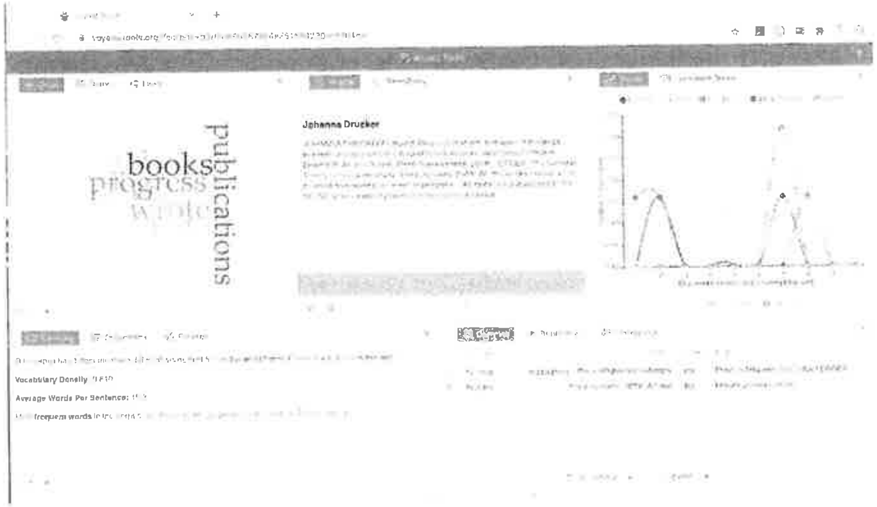

Figure 7.1 Voyant screen (Voyant Tools, Stéfan Sinclair and Geoffrey Rockwell, CC BY 4.0.) (CC License)

can be trained to produce more refined and precise topic models by eliminating certain terms and privileging others. The standard description of a topic modeling activity is that it "extracts" topics from the analysis of texts. This makes it seem as though topics are already present in a text, rather than being produced as an act of reading or interpretation. Like other computational techniques, topic modeling is often defined in opposition to human reading practices. This characterization mistakenly represents the processes as value free and neutral.

Topic modeling depends upon machine learning and natural language processing. Sophisticated tools for the analysis of parts of speech, and what are referred to as "named entities"—proper nouns with specific (but sometimes ambiguous) references, and other features of natural language have been developed by Stanford's Natural Language Processing Group. This means that it is "trained" on texts and then develops "knowledge" of how to analyze them. The training involves human interaction and reinforcement but is rooted in neural net processes (a machine learning technique that works from a "bottom up" approach) that depend on reinforcement. Because they work with natural language, which does not have the formal constraints that computational languages possess, these analyses also make use of dictionaries compiled to help disambiguate word usage.[1]

Text analysis brings the contrast between natural and formal languages into sharp relief. In formal languages, any expression has to be explicit and unambiguous. In natural language, the context and syntactic structure in which the word appears are crucial. So are definitions and probabilities of occurrence. What, for instance, tells a program to read the word "compound" as an adjective, a noun, or a verb? How is slang and dialect—frequently present in literary work—to be interpreted? What about words that might have widely variant meanings? For instance, a term like "sure" uttered in a response might mean anything from "yes," to "no," to anything in between. The other crucial category of words generally left out of topic modeling is what are termed "stop words." The term simply means they are not factored into the analysis, and tend to include the conjunctions (and, but, etc.) and articles (the, a, an). But think of the distinction between "the savior" and "a savior," "the god" and "a god." The definite and indefinite articles are crucial for meaning production, so leaving them out of an analysis can introduce distortions.

Like all computational activity that depends upon models, the processes can be criticized. Does text analysis find what it is trained to find, or engage in a discovery that is neutral? One area where this issue has been addressed directly is in the analysis of gendered identities and texts. In a critical text focused on this issue, Laura Mandell posed the question of whether particular approaches to cultural analytics "find" or "make" their stereotypes (Mandell 2019). The fundamental recognition is that gender is not a simple Male/Female binary equated with biological categories. Mandell's distinctions reinforce an approach in which gender is constructed—in literary work, language, lived experience, roles, and characterizations. But her critical eye focused on work that assumed gendered identity as a given and then tracked it in literary texts. She noted that this resulted in the reaffirmation of gender stereotypes, since the model on which the text analysis was carried out took gender difference as an a priori category.

Mandell's critical insights arose from work on gender, but they are also relevant to other topics or themes. Modeling is a powerful instrument, and it asserts its values in seeking confirmation of what it details. Numerous scholars thought distant reading would avoid the biases of close reading. But "macro-analysis," to use a term identified with Matt Jockers, has often replicated the very value judgments it was believed computation would avoid.

This section has dealt with methods of data and text analysis. In addition, it is useful to address ways of extracting data from online sites. For this purpose, a combination of features is useful: one part, known as APIs, provides access to data on a site. The other consists of tools for "scraping" data from online resources. Both are powerful assets for research.

APIs: application programming interfaces and web-scraping

One crucial source of data for digital research is in projects and sites that offer a well-structured option for export and re-use of their files. The main method by which this is accomplished is in the use of APIs, applied programming interfaces, which are designed to package data to make it portable and usable. APIs are built. They are not automatically present. Scholars who see their work as useful for others will add this feature to their online repositories or sites. Other methods of data mining can be used to extract information from such a site if an API is not present, in a process known as "web-scraping," which consists of tools used to capture data from social media sites and other platforms.

The ethics of data capture and re-use from online sites are subject to the same considerations as repurposing of any data. Attention to intellectual property and privacy concerns forms one set of ethical issues, and authentication and verification of data form another. Here, as in other practices, documentation is essential to preserve the trail of scholarship, even if the data themselves are no longer recoverable from an original source.

An API is actually a program designed to let another program access or manipulate data through an interface. We usually think of interface as a human-to-computer function, but programs can also have this capacity. Of course, it will be a human user who takes advantage of the protocols built into the API—to query the data on a site and search and extract information from its repositories. But the code is designed to make it easy for human users to get one machine to talk to another for purposes of information exchange and export.

When a site is enormous—like a national library—then it can be a rich resource for many kinds of research. The Australian National Library, for instance, has a platform called "Trove" that supports use of its API through "query syntax" and a collection of case studies that demonstrate what can be done with its tools. The Australian National Library uses a console format to guide users in constructing searches that can be incredibly broad (everything) or highly focused (newspaper weather reports in a particular place on a particular day). You will find that many APIs return their results in JSON, which we discussed earlier as a popular data exchange format.[2] [See Exercise #3: Trove: Reading an API.]

Web-scraping tools are designed to acquire data from existing sites without the use of APIs or custom coding. These are also frequently built in the already frequently mentioned language, Python. Off-the-shelf tools exist, most of which were developed to serve marketing and commercial purposes, but they can be used for humanities research as well. They extract data in real time and many allow for anonymous data collection, the ethics of which you will need to consider not just as a user of these tools, but as someone whose work and information might be scraped and put to purposes you did not envision or authorize.[3] A look at the description of these tools will acquaint you with vocabulary relevant to the automated operation of web-scrapers and their ability to bypass bots, to crawl sites automatically, to convert data to a usable format, and to collect and store cookies from other sites.

Some of the simplest scrapers can be installed as browser extensions, and in fact, Zotero, a bibliographical tool designed by humanists, can be used to "scrape" data and store it for research purposes. Zotero stores metadata from sites and is particularly focused on the information relevant to scholarly bibliographical formats. It works at a very small scale and only when initiated by a user, while many web-scraping tools used commercially operate automatically to generate massive data sets in all sectors of private and public activity.

Takeaway

Fundamental issues of digital humanities are present in distant reading: the basic decisions about what can be measured (parameterized), counted, sorted, and displayed are interpretative acts that shape the outcomes of the research projects. Distant reading is a combination of text analysis and other data mining performed on metadata or other available information. Natural language processing applications can summarize the contents of a large corpus of texts. Data mining techniques can show other patterns at a scale that is beyond the capacity of human processing (e.g. How many times does the word "prejudice" appear in 200,000 hours of newscasts?).

The term distant reading is created in opposition to the notion of "close reading," the careful attention to the composition and meaning of texts, images, musical works, or other cultural artifacts that is at the heart of humanistic interpretation. Automated web scraping and data export through APIs allow data to be captured and repurposed at multiple scales for research purposes.

Exercises

Exercise #1: Google Ngram Viewer

Open the Google Ngram Viewer and select a date range. Enter several terms or names and see what has changed over time. How much can you trust your results? What are they based on?

Exercise #2: Voyant and Mallet

Go to https://voyant-tools.org/ and enter a block of text with which you are familiar. Look at the first display and figure out what each pane is showing. Play with the other tools and display modes. Do they match? Why is there a range of results in the displays? Then, see if you can understand the workings of Mallet: https://programminghistorian.org/en/lessons/topic-modeling-and-mallet

Exercise #3: Trove: reading an api

Look at the Australian National Library Trove API console. What can you learn from reading the documentation? Construct a query and assess the results. What ideas does this give you for designing an API? Compare the results of a search in the library's catalog and the format of its results, with the XML output generated by Trove which is useful for data analysis. http://troveconsole.herokuapp.com/#otherzones Also useful: https://studentwork.prattsi.org/dh/2019/05/13/getting-data-for-digital-humanities-with-apis/ tutorial for DH data extraction with APIs.

Recommended readings

Janicke, S., G. Franzini, M. F. Cheema, and G. Scheuermann. 2015. "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges." Euro-Graphics Conferenceon Visualization. https://pdfs.semanticscholar.org/20cd/40f3f17dc7d8f49d368c2efbc2e27b0f2b33.pdf.

Piper, Andrew and Richard Jean So. 2015. "Quantifying the Weepy Bestseller." *The New Republic*. https://newrepublic.com/article/126123/quantifying-weepy-bestseller.

Schulz, Kathryn. 2011. "What is Distant Reading?" *New York Times Book Review*. www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html.

7b Cultural analytics, multi-modal communication, media, and audio mining

Text analysis, topic modeling, and other computationally enabled processes of producing meaning from language-based documents play a significant role in digital humanities. So does the analysis of quantitative data. Texts and numbers are readily remediated into computationally tractable forms. Numbers have a discrete identity and texts can be processed at the level of the word, or phrase, with relative efficiency, even if their meanings may be ambiguous or complex. Though it sounds reductive and mechanistic, numbers and texts can both be correlated to keystrokes and this gives them a structure within digital file formats. They are essentially alphanumeric code which translates easily into machine code and binary files.

But what about other forms of cultural expression—images, audio recordings, video and film, and the many representations in photographs and other documents? Every medium poses its own set of challenges for extracting information in a meaningful way. But each process has in common the same set of requirements—to translate analog or digital materials into a form in which a feature set can be identified, parameterized, tokenized, and processed computationally. This generally involves making discrete features from continuous phenomena. As in all such processes, the conceptual work and the technological developments have to coordinate. The ways we think about music or images will structure some of the ways digital representations are created—and the purposes to which they are put. In whose interest is it to do data mining of images or social media? Art historians, or police surveillance units? And music? Artists looking for inspiration, scholars studying historical materials, or industry sleuths tracking piracy—or those indulging in it (Kennedy and Moss 2015)?

Cultural analytics

Images pose particular challenges for data mining. They are not expressed in a structured notation (like letters, punctuation, and spaces) in the same way as language. In spite of these apparent impediments, experts in image recognition and analysis have developed useful tools and platforms in the humanities (alongside other developments in science and applied research). One of these is the Cultural Analytics, a term coined by Lev Manovich to describe work that uses digital capacities to analyze, organize, sort, and computationally process large numbers of images. Images have different properties than texts. As we noted in the section on digitization, the act of remediating an image into a digital file involves choices that determine what kind of information it contains. If the human visual apparatus can glean more from visual information in certain areas of the spectrum or even color wheel, then what is the point of digital files that contain other visual content? Machine processing may not correspond to human perception, and in many advanced imaging technologies, this is an advantage to discovery and exploration (Jofre et al. 2020).

Manovich's project, however, pioneered the processing of visual image data and its analysis in visualizations. Manovich's research was motivated by the question of how to analyze images at scale, just as Moretti's was prompted by "reading" massive amounts of text. What features of digital surrogates could be quantified and used for comparison? Metadata on images also plays an important role in data mining, particularly for those features like size, medium, artist, collection, provenance, date, and other information that is not visible. But cultural analytics focuses on those features of a visual surrogate that can be extracted automatically. If an image is stored as a pixel-based file, then color values supply crucial information. If it is stored as a vector-based-graphic, then shape and proportion lead. If the image is black and white, then tonal range, values, and contrast are particularly conspicuous. Image recognition at the level of iconography relies on combinations of object and scene recognition, and links to image libraries that are working with deep learning algorithms.

Nuance is still lacking. At what point is an image of a mother holding a young child actually a Madonna? Can the Mona Lisa be distinguished from other portraits by Leonardo da Vinci—and can her emotional state be discerned automatically (Dunne 2015)? Can a religious allegory of a sheep be differentiated from a literal painting of animals in a pasture? All discussions of automated learning and processing come up against these limits with regard to issues of context and learned conventions within human experience. Many machine learning projects use human labor, such as Mechanical Turk, to provide sample data sets to train the algorithms (Anderson 2017). [See Exercise #1: Cultural Analytics.]

Multi-modal analysis

Recent work in digital humanities and computer vision have developed a "Distant Viewing" approach to moving image analysis (Tilton et al. 2018). These focus on color and lighting, time codes for shots and breaks, means of establishing boundaries for identifying faces and other objects, sound analysis, automatic voice to text transcripts for speech, and so on. Each of these dimensions requires its own computational processes and functions with different degrees of accuracy. However, consider that an archive of oral histories might contain a hundred thousand hours of recordings. For an individual to watch all of these, working ten hours a day, would take ten thousand days . . . about thirty years. Automated techniques for searching and sorting clearly have a purpose. The development of parsers—programs that can automatically detect features and structural elements to a degree that starts to resemble detecting the content of images—has been rapid and has produced workable results. These combine the "generative" or "top down" analysis (using classification systems and categories) with "discriminative" or "bottom up" processes (Kuhn 2018).

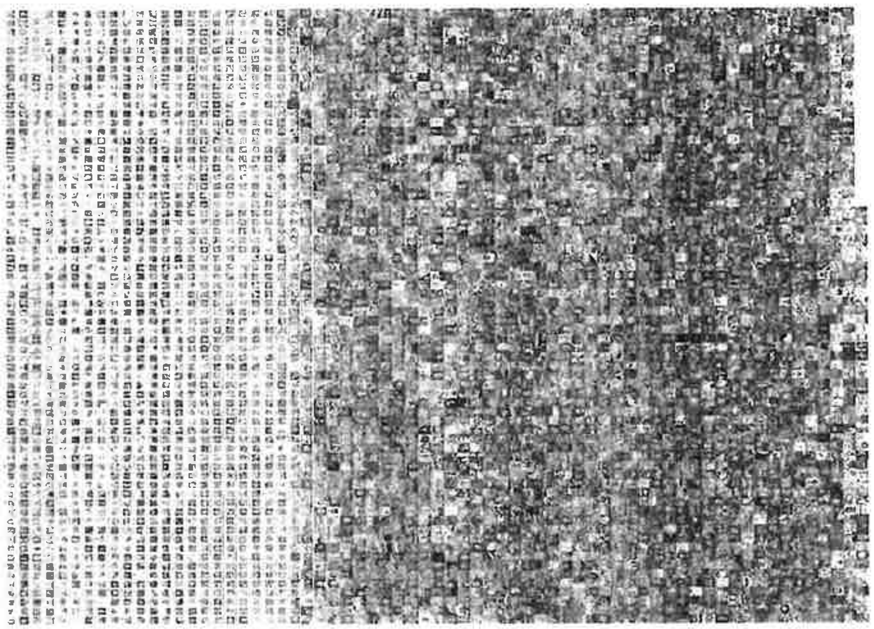

Figure 7.2 Cultural Analytics analysis of features of *Time* magazine covers

Source: (Image courtesy of Lev Manovich and Jeremy Douglass, Cultural Analytics Lab.)

Biases are built in. Multiple studies confirm that facial recognition and emotion detection work differently on images of people of color. Voices with accents or unusual timbre will not translate to text. Experimental and avant-garde films, even independent and documentary footage, will be processed differently than commercial ones. Above all, the multiple layers of meaning production identified by art historians like Erwin Panofsky, that require situating an image within a field of others to understand how it communicates in a particular cultural moment, are not likely to be grasped by an algorithm. Nor, sad to say, is humor. Jokes, references, parodies—all of the features of communication in which images play a role, are difficult to codify. A dog in glasses? A cat in a dress? A drunk frog on a cocktail napkin? We read these easily as commentary on human activities—but can a computer?

Automated processing of sound and audio-visual media

The issues of feature recognition and segmentation in file formats plague data mining, as does the quality of digital surrogates. Only some features of an analog experience are represented in a sound recording. Formats matter in the analog world as well as in the digital. Analog sounds are continuous waves so translating them into discrete units of binary information involves translation. At a small enough granularity, human ears will not detect this, any more than we see the "pixels" on a screen image. But since sound is a time-based medium, sampling rates, the number of channels, and the bits-per-sample have to be taken into account. Audio analysis research and applications exist in a generative tension between machine capabilities and the simulation of human perceptual experience. In many ways, the advantages of digital processing are best appreciated for doing what humans cannot do rather than for trying to emulate our capacities (Giannakopoulos and Pikrakis 2014).

Audio classifications of acoustic data, environmental sound, and musical classification have been applied to historical materials as well as contemporary ones. Sentiment analysis in voice, as in texts and facial recognition, depends on classification of sound, but also, assessment of value across a range and at a rate of change. Voice recognition has profound gender and race biases, as do the facial recognition processes already mentioned. But tools for editing, for "pumping up" signals, and for recovering lost or damaged sound information has many benefits in the work of historical preservation.

Attempts to classify massive numbers of musical recordings that are not identified, but are accessible in online environments, have prompted development of musical data mining techniques (Patchet, Westermann, and Laigre n.d.). Industry applications have led professional groups, like the IEEE (Institute of Electronic and Electrical Engineers) to turn their attention to multimedia analysis (Ogihara and Tzanetakis 2014). Cross-mediation, or the rendering sound data in visual format for analytic purposes has become a common feature of audio editing. Images are easier to work with on a screen, and rendering audio files as visualizations has become common practice (Hartquist 2018).

Sound analysis has also been applied to the study of poetry performance and other voice recordings. Humanities scholars, such as Charles Bernstein and Al Filreis, who have been the directors of Penn Sound, an archive of recordings of poetry readings, offer challenges for research. How does inflection, or variation in the oral production of a work, vary over time—and how is the significance of this to be identified and analyzed. As sound studies have grown, so have digital techniques for engaging with a wide range of ambient, human, musical, industrial, and natural sounds. Historical archives of sound recordings, digitized and made available, provide another important research corpus for humanists (Clement 2012). [See Exercise #2: Sound files.]

Time-based media pose their own specific problems for both digitization and analysis. Recording times and running times may not be the same, and time-stamped files may play at different rates on different platforms. Emulation stations for recovering the correct display of audio-visual materials (and other file formats) are designed to wrestle with these technical issues. Emulation has been particularly important for preservation of games but is relevant across genres and formats.

Multi-modal communication, the phrase used to describe the ways human cognition processes image, sound, inflection, body language, and other features in a dimensionally rich experience, has also been an object of research in digital analysis and processing. Multiple features need to be identified for such work—and translated into computable elements. In a 1998 paper, Tsuhan Chen listed these: lip reading, facial animation, speaker verification, joint audio-visual coding, and so on. But when these many facets of expression are at odds with each other—when a speaker's tone deliberately contradicts a message, or when gestures undermine a statement—how are these coded and processed? What can be learned from the attempts to train automated systems in these matters? Defining the research questions from which humanists can benefit requires strategic thinking. [See Exercise #3: Media processing.]

What is clear is that most computational processes produce benefits at scale—making it possible to search, query, and make use of large corpora for specific purposes. But some valuable results have come from automated processing of small data sets, not just large ones. The imaging work of the Western Semitic Epigraphy project has pioneered techniques that make use of aggregated image data. The photographic and imaging techniques they combine to extract information from ancient inscriptions combine digitization and computation. One of these is Reflectance Transformation Imaging. This depends on making multiple photographs from different directions and with different lighting angles. In combination with multispectral imaging, which makes use of different wavelengths of light, it reveals information that is simply not visible to the human eye under any circumstances. Ultraviolet or infrared light can reveal traces of chemicals that have vanished. By aggregating the features that each imaging technique can record, the researchers are able to create a "picture" of an artifact that is legible in digital formats. This is work that focuses intensely on a very small corpus of materials but extends its value in ways that would not be possible without computational processing.

Takeaway

Computational tools to analyze big data have to balance the production of patterns, summaries at a large scale, with the capacity to drill down into the data at a small scale. Automated analysis of materials in all media—images, sound, video, and even material objects—has produced new insights into the dialogue between digital tools and humanistic research goals like recovery of historical materials, study of vast corpora, and trends in current cultural practices. Challenges for the claims to objectivity and scientific method remain—and one of the crucial questions is that of repeatable results. In many instances the repeated applications of tools of data mining and cultural analytics return shifted and changed outcomes. The corpus is unstable, for one thing, but the probabilistic character of processing contributes to this as well. Perhaps that is the most human aspect of what is otherwise a technological operation.

Exercises

Exercise #1: cultural analytics

Look at Lev Manovich's http://lab.softwarestudies.com/2008/09/cultural-analytics.html and design a project for which cultural analytics would be useful. Think in terms of the large scale of comparison and stay within humanities disciplines.

Exercise #2: sound files

Examine the project by Tanya Clement, Hipstas "John A. Lomax and Folklore Data." What are the ways in which these folklore files from the early 20th century become more useful as a result of the digital interventions? What other kinds of materials do you think would benefit from such research? https://hipstas.org/2015/05/11/john-a-lomax-and-folklore-data/

Exercise #3: media processing

If you were given the task of teaching an automated system to distinguish between news stories and advertisements in a television broadcast, what features would you identify for digital processing? Keep in mind that the task is to identify features that can be distinguished on the basis of their formal properties.

Notes

[1] A look at the work of the Stanford Natural Processing Group, including analysis of their topics and tools, provides useful insight into this aspect of computational work.

[2] See the Programming Historian for lessons on developing an API with Python. The lessons include information on downloading and installing Python and Flask to do this work.

[3] A few useful tools for getting data from online sites include: Conifer, Beautiful Soup, and Webscraper.io.

Recommended readings

Bajorek, Joan Palmiter. 2019. "Voice Recognition Still Has Considerable Race and Gender Biases." *Harvard Business Review*. https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases.

Manovich, Lev. 2011. "How to Compare One Million Images." *Cultural Analytics*. http://softwarestudies.com/cultural_analytics/2011.How_To_Compare_One_Million_Images.pdf.

Hartquist, John. 2018. "Audio Classification Using FastAI and On-the-Fly Frequency Transforms." *Towards Data Science*. https://towardsdatascience.com/audio-classification-using-fastai-and-on-the-fly-frequency-transforms-4dbe1b540f89.

Harwell, Drew. 2019. "Federal Study Confirms Racial Bias of Many Facial Recognition Systems." *Washington Post*. www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/.

References cited

Acerbi, Alberto, Vasileios Lampos, Philip Garnett, and R. Alexander Bentley. 2013. "The Expression of Emotions in 20th Century Books." *PLoS One* 8 (3). www.ncbi.nlm.nih.gov/pmc/articles/PMC3604170/.

Anderson, Steve. 2017. *Technologies of Vision*. Cambridge, MA: MIT Press.

Busa, Robert. 1980. "The Annals of Humanities Computing: The Index Thomisticus." www.alice.id.tue.nl/references/busa-1980.pdf.

Clement, Tanya. 2012. "Announcing High Performance Sound Technologies for Access." http://tanyaclement.org/2012/08/09/hipstas/ and https://hipstas.org/2015/05/11/john-a-lomax-and-folklore-data/.

Delice, Ali. 2010. "The Sampling Issues in Quantitative Research." *Educational Sciences: Theory and Practice* 10 (4). https://files.eric.ed.gov/fulltext/EJ919871.pdf.

Dunne, Carey. 2015. "Microsoft's New Emotion-Detecting App Deems the Mona Lisa 43% Happy." *Hyperallergic*. https://hyperallergic.com/261508/microsofts-new-emotion-detecting-app-deems-the-mona-lisa-43-happy/.

Fox, Neal, Omran Ehmodea, and Eugene Charniak. 2012. "Statistical Stylometrics and the Marlowe-Shakespeare Authorship Debate." https://cs.brown.edu/research/pubs/theses/masters/2012/ehmoda.pdf.

Giannakopoulos, Theodoros, and Aggelos Pikrakis. 2014. *Introduction to Audio Analysis*. Academic Press. www.sciencedirect.com/book/9780080993881/introduction-to-audio-analysis.

Guldi, Jo. 2018. "Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora." *Journal of Cultural Analytics*. https://culturalanalytics.org/article/11028.

Hai-Jew, Shalin. 2015. "A Light Stroll through Computational Stylometry and its Early Potential." *C2C Digital Magazine*. https://scalar.usc.edu/works/c2c-digital-magazine-fall-winter-2016/a-light-stroll-through-computational-stylometry-and-its-early-potential.

Jofre, Ana, Josh Cole, Vincent Berardi, Carl Bennett, and Michael Reale. 2020. "What in a Face? Gender Representations of Faces in Time, 1940s-1990s." *Journal of Cultural Analytics*. https://doi.org/10.22148/001c.12266.

Kennedy, Helen, and Giles Moss.2015. "Known or Knowing Publics? Social Media Data Mining and the Question of Public Agency." *Big Data & Society* 2 (2), SAGE Publications Ltd. DOI: 10.1177/2053951715611145.

Kirschenbaum, Matthew G. n.d. "The Remaking of Reading: Data Mining and the Digital Humanities." www.academia.edu/35646247/The_remaking_of_reading_Data_mining_and_the_digital_humanities.

Kuhn, Virginia. 2018. "Images on the Move." In *The Routledge Companion to Media Studies and Digital Humanities Routledge*. New York: Routledge.

Lamarche, Stephen. 2012. *LARB*. Los Angeles, October. https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/.

Lee, Changsoo. 2019. "How Are 'Immigrant Workers' Represented in Korean News Reporting?—A Text Mining Approach to Critical Discourse Analysis." *Digital Scholarship in the Humanities* 34 (1): 82–99. DOI.org: 10.1093/llc/fqy017.

Mandell, Laura. 2019. "Gender and Cultural Analytics: Finding or Making Stereotypes?" In *Debates in Digital Humanities*. Minneapolis, MN: University of Minnesota Press. https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities-2019.

Ogihara, Mitsunori, and George Tzanetakis. 2014. "Special Section on Music Data Mining." *IEEE Transactions on Multimedia* 16 (5): 1185–187. https://ieeexplore.ieee.org/document/6856270.

Patchet, François, Gert Westermann, and Damien Laigre. n.d. "Musical Data Mining for Electronic Music Distribution." www.music.mcgill.ca/~ich/classes/mumt621_09/Query%20Retrieval/Pachetwedelmusic.pdf.

Sandelowski, Margarete. 1995. "Sample Size in Qualitative Research." *Research in Nursing and Health*. https://onlinelibrary.wiley.com/doi/abs/10.1002/nur.4770180211.

Schmidt, Benjamin M. 2013. "Words Alone: Dismantling Topic Models in the Humanities." *Journal of Digital Humanities*. http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/.

Serlen, Rachel. 2010. "The Distant Future? Reading Franco Moretti." *Literature Compass* 7. https://warwick.ac.uk/fac/arts/english/currentstudents/undergraduate/modules/fulllist/special/en264/serlen_reading_franco_moretti.pdf.

So, Richard Jean, and Edwin Roland. 2020. "Race and Distant Reading." *PMLA* 135 (1): 59–73. www.mlajournals.org/doi/abs/10.1632/pmla.2020.135.1.59?journalCode=pmla.

Sculley, D., and B. M. Pasanek. 2008. "Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities." *Literary and Linguistic Computing* 23 (4): 409–24. DOI: 10.1093/llc/fqn019.

Terras, Melissa, and Julianne Nyhan. 2016. "Father Busa's Female Punch Card Operatives." In *Debates in the Digital Humanities*. Minneapolis, MN: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/57.

Tilton, Lauren, Taylor Arnold, Thomas Smits, Mark Williams, Lorenzo Torresani, Maksim Bolonkin, John Bell, and Dimitrios Latsis. 2018. "Computer Vision in DH." In *Digital Humanities 2018*. Mexico City. https://dh2018.adho.org/computer-vision-in-dh/.

Underwood, Ted. 2017. "A Genealogy of Distant Reading." *Digital Humanities Quarterly* 11 (2). http://digitalhumanities.org:8081/dhq/vol/11/2/000317/000317.html.

Resources

API data creation https://programminghistorian.org/en/lessons/introduction-to-populating-a-website-with-api-data. and https://programminghistorian.org/en/lessons/creating-apis-with-python-and-flask.

Computer Vision (Heidelberg University) https://hci.iwr.uni-heidelberg.de/compvis/projects/digihum.

Cultural Analytics http://lab.culturalanalytics.info/p/projects.html.

Emulation https://libguides.bodleian.ox.ac.uk/digitalpreservation/emulation.

Image-Net www.image-net.org/.

Inscriptifact. www.inscriptifact.com/aboutus/index.shtml.

Natural Language Processing https://nlp.stanford.edu/software/.

Python (an introduction) https://wiki.python.org/moin/SimplePrograms.

Quantitative history http://historymatters.gmu.edu/mse/numbers/what.html.

R (an introduction) www.r-project.org/about.html.

Voyant. https://voyant-tools.org/.

Zotero www.zotero.org/.