CHAPTER 19

Distant Reading

Natalie M. Houston

A variety of technological shifts in the twenty-first century have transformed the questions, objects, tools, and methods of humanities scholarship. Today, even those who do not perceive the methods or objects of their research as having anything to do with digital technology are likely to consult digital library catalogs and reference databases, read secondary or primary source materials from digital surrogates, and manage some of their personal research materials as digital files. Although in recent years the idea of distant reading has largely become associated with computational analysis performed with large textual datasets, when Franco Moretti first coined the phrase "distant reading" in the year 2000, he was referring to comparative studies that synthesize data from a number of sources to enable analysis of literary history on a grander scale. His subsequent work with the Stanford Literary Lab, the increased availability of digitized texts, and the developing field of digital humanities has led to the current understanding of distant reading as "the computational processing of textual information in digital form" or "using digital tools to 'read' large swaths of text."[1]

In distant reading, Moretti suggests, "distance … *is a condition of knowledge*: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes – or genres and systems."[2] As the Google Library Project began scanning the contents of university library shelves in 2004, followed by the formation of the Hathi Trust Digital Library in 2008, the possibilities for large-scale literary

[1] Johanna Drucker, "Why Distant Reading Isn't," *PMLA* 132.3 (2017): 629; Matthew K. Gold and Lauren F. Klein, "Introduction. Forum: Text Analysis at Scale," in *Debates in the Digital Humanities*, ed. Matthew K. Gold and Lauren F. Klein (Minneapolis: University of Minnesota Press, 2016), open access edition: https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities-2016 (accessed April 22, 2022).

[2] Franco Moretti, *Distant Reading* (London: Verso, 2013), 48; emphasis in the original.

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

history evoked by Moretti's bold claims began to emerge.[3] At the same time, researchers and general internet users alike became accustomed to searching digital text, from web pages to scholarly databases. As Ted Underwood suggests, algorithmic search has been "invisibly naturalized" over the past two decades, obscuring the technological sophistication of search processes and the impact of relevance ranking of search results.[4] Full-text search and open access digital archives have expanded the kinds of research questions literary scholars could pursue, at both the very large and very small scales.

But even as these developments have made computers and digitization very familiar, the integration of computational analysis into humanities scholarship has been slow and sometimes controversial.[5] This chapter situates the methodological approach known as distant reading within the technological changes of the last twenty-five years by highlighting key examples of exploratory, generative, and predictive modeling applied to humanities research questions and showing how web-based tools now make such models available even to those without computer programming experience.

1 Defining Distant Reading

The rhetorical power of the phrase "distant reading" has persisted in part because of the way Moretti set it in opposition to the close reading methods enshrined in Anglo-American literary criticism after the New Criticism. Moretti argues that "the trouble with close reading (in all of its incarnations, from the new criticism to deconstruction) is that it necessarily depends on an extremely small canon … you invest so much in individual texts *only* if you think that very few of them really matter."[6] Moretti thus linked his argument for changing the scale of writing literary history to a critique of the literary canon as overly restrictive. He acknowledges that studying literary form and cultural processes at other scales means

[3] "About the Library Project," Google Books: https://support.google.com/websearch/answer/9690276 (accessed April 22, 2022); "Welcome to Hathi Trust!," Hathi Trust Digital Library: www.hathitrust.org/about (accessed April 22, 2022).

[4] Ted Underwood, "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago," *Representations* 127.1 (Summer 2014): 64.

[5] See, for example, Katie Trumpener, "Paratext and Genre System: A Response to Franco Moretti," *Critical Inquiry* 36 (Autumn 2009): 159–171 and Alexander R. Galloway, "Everything Is Computational," *Los Angeles Review of Books* (June 27, 2013): https://lareviewofbooks.org/article/franco-morettis-distant-reading-a-symposium/ (accessed April 22, 2022).

[6] Moretti, *Distant Reading*, 48; emphasis in the original.

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

that "between the very small and the very large, the text itself disappears": Distant reading is not at all the same thing as reading.[7]

Working with a very large corpus of texts requires new kinds of analysis, as he suggests in a later study: "we are not doing the same thing, 1,000 times bigger; we are doing a different thing. The new scale changes our relationship to our object, and in fact it changes the object itself."[8] Although the rationale for selection is not always made explicit in traditional literary studies, scholars inevitably focus on a selection of texts: texts chosen because they exemplify certain patterns, genres, or themes; because they were produced under certain historical conditions; or because they were written by certain authors. Selecting and preparing texts for computational analysis requires explicit justification for numerous decisions made throughout the process. As Moretti points out, "no one writes corpora; they are not 'communicative events' – they are not events at all, they are artificial objects created by a researcher."[9] Even for those with good institutional support and resources, researchers must decide whether they will work within the selection biases and technological constraints of already-digitized materials, or instead put significant time, money, and labor into digitizing other texts.[10] By changing the scale at which texts are analyzed, whether breaking a text down to individual parts of speech or examining the developments of genres over many decades, distant reading transforms the object of study and thereby the kinds of questions that can be explored.

Ted Underwood traces the intellectual origins of distant reading in fields like "book history, sociology, and linguistics," which "pose broad historical questions about literature, and answer them by studying samples of social or textual evidence."[11] Adopting an experimental approach requires distinguishing among the stages of knowledge production: forming a research question, developing a hypothesis, selecting appropriate data, conducting analysis, and formulating an argument for its significance. Importantly, this experimental methodology "separates the construction of the sample

[7] Ibid., 49.

[8] Franco Moretti, "Patterns and Interpretation," *Pamphlets of the Stanford Literary Lab* 15 (September 2017), 1: https://litlab.stanford.edu/LiteraryLabPamphlet15.pdf (accessed April 22, 2022).

[9] Ibid., 2.

[10] Mark Algee-Hewitt and Mark McGurl, "Between Canon and Corpus: Six Perspectives on 20th-Century Novels," *Pamphlets of the Stanford Literary Lab* 8 (January 2015), 1–8: https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf (accessed April 22, 2022).

[11] Ted Underwood, "A Genealogy of Distant Reading," *DHQ: Digital Humanities Quarterly* 11.2 (2017), para. 3, para 10: www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html (accessed April 22, 2022).

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

from the process of historical inference."[12] The distinctions among these steps are often only implied in literary criticism. As Matthew Jockers points out, experimentation "offers a method through which competing observations and conclusions may be tested and ruled out," which runs counter to the traditional emphasis on debate and subjective interpretation in literary studies.[13]

2 Models

In statistics, a model is a "high-level, global description of a data set" that identifies its important features, and uses them either to summarize and describe the data (known as descriptive modeling); to hypothesize about the distributional processes that generated the data (in generative modeling); or as the basis for learning about the data in order to predict aspects about other data points (in predictive modeling).[14] A model can be understood as "a general scheme, or template" that represents both "components of the information (features of texts, images, numbers, etc) and their relations to each other."[15] The purpose of the analysis would determine the choice of model, which would then guide the selection of specific features of the data.

A library's catalog, for example, can be understood as a model for representing the books that sit on its shelves. Most library catalogs for general circulating collections represent each book with its bibliographical information, such as author name, book title, publication year, and publisher name and location. A catalog can thus generate different summary descriptions of the library's holdings: The same collection could be described as containing 500 books; books by 200 different authors; books published from 1705 to 1999; or books published in London, New York, and Paris. A slightly more complex descriptive model would analyze how those different features relate to each other, by counting the number of books held by each author or looking at the distribution of publisher location in different decades.

Models are always constructed for a specific purpose and involve numerous decisions about which aspects of the object of study will be represented. For example, circulating library catalogs rarely represent the

[12] Ibid., para. 11.
[13] Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana: University of Illinois Press, 2013), 6.
[14] David Hand, Heikki Manilla, and Padhraic Smyth, *Principles of Data Mining* (Cambridge: MIT Press, 2001), 165; David Foster, *Generative Deep Learning: Teaching Machines to Paint, Write, Compose and Play* (Sebastopol: O'Reilly, 2019), 1–2.
[15] Drucker, "Why Distant Reading Isn't," 629.

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

material aspects of their holdings in detail, such as the paper composition or binding materials. But rare book libraries may choose to represent those features in their catalog records, because they are considered more relevant to understanding or classifying rare books. As Andrew Piper suggests, even though the language of modeling may seem alien to the domain of literary studies, literary critics nonetheless represent texts in a kind of descriptive model, by quoting passages from the text and producing historical, biographical, or formal descriptions of it.[16]

In order to prepare texts for statistical modeling, words and other features of the text must be represented by numeric values. This can be either a binary representation (1 if the word is present, 0 if it is not) or a count of the word's frequency in the text. Nonword features of a text can also be represented in the model, through categorical variables describing features of the text such as genre, form, or authorship, or through discrete or continuous numeric variables recording information such as dates, line counts, or formal elements of the text.

3 Descriptive Modeling

Descriptive models offer new understanding of a dataset by changing the scale at which its observations are understood. For example, transforming a literary text into a list of all of its words with their frequency counts shifts the meaningful unit of interpretation from the book, paragraph, or sentence down to the level of the word. Bypassing theme, plot, or other narrative conventions to examine patterns at the level of the sentence, vocabulary, or genre system changes what counts as evidence, as well as what kinds of questions can be pursued.

Even very simple measures of word frequency can reveal interesting patterns when examined in a large corpus of texts. Type-token ratio (TTR), for example, is a standard measure of lexical variety: the ratio of word types (distinct words) in a text to the number of its tokens (all words, including repeated ones). A 100-word text containing 100 distinct words would have a TTR of 1, and very high lexical variety; a 100-word text containing 25 distinct words, would have a TTR of 0.25, indicating more word repetition. Examining TTR in a large corpus of over 1,000 canonical and noncanonical nineteenth-century novels, Algee-Hewitt et al. found that the canonical texts tended to have lower TTRs, corresponding to their emphasis on realistic depictions of oral speech. Although the less-canonical novels also

[16] Andrew Piper, "Think Small: On Literary Modeling," *PMLA* 132.3 (2017): 652–653.

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

included passages of dialogue, "their 'spoken' passages have a decidedly 'written' quality … formal tirades that sound closer to a written disquisition than to an oral exchange."[17] They found that the noncanonical novels use more nouns and adjectives than do the canonical ones, which contributes to this discrepancy in register. Exploratory analysis thus provides a picture of the larger context of novelistic language against which specific examples can be compared.

Some experiments in distant reading transform works by a single author, or even a single work, by changing the scale of observation. In her research, Sarah Allison created a descriptive model of George Eliot's novel *Middlemarch* by selecting every third sentence, totaling 4,400.[18] This practice of selection strategically disrupts typical narrative reading practices, and her analysis of their syntactical structure reveals important patterns in Eliot's writing. Allison identifies a frequent grammatical construction that she calls the "commentative clause": a sentence that links the narrative's past tense with a generalizing present tense by means of the relative pronoun "which."[19] Allison demonstrates how the novel uses this clause to multiply and diffuse critical judgments in its description of characters and events, pointing out that "a major philosophical problem in *Middlemarch* is how to see others clearly but without contempt: the many facets of this problem and the partial solutions to it that *Middlemarch* proposes are most clearly seen not in its plot but at the level of the sentence."[20]

Andrew Piper similarly uses descriptive modeling to resist the force of plot in narrative theory, focusing on "words in the distributional sense, as a larger set of patterns."[21] Working with German, French, and English novels, Piper calculates the vocabulary distribution for each chapter within a given novel and then measures their similarity to one another. Using these measurements, he identifies a set of novels in which chapters in the second half of the text are more similar to each other than those in the first half. Examining the vocabularies used in the first and second halves of these "contractive" novels points to shifts in the use of characters' proper names and the resolution of the marriage plot as connections between the social world reflected in modernist fiction and the vocabularies used

[17] Algee-Hewitt, Mark, et al., "Canon/Archive: Large-Scale Dynamics in the Literary Field," *Pamphlets of the Stanford Literary Lab* 11 (January 2016), 10: https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf (accessed April 22, 2022).

[18] Sarah Allison, *Reductive Reading: A Syntax of Victorian Moralizing* (Baltimore, MD: Johns Hopkins University Press, 2018), 63.

[19] Ibid., 61–63.

[20] Ibid., 80.

[21] Andrew Piper, *Enumerations: Data and Literary Study* (Chicago: University of Chicago Press, 2018), 42.

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

in its writing: "There is an implicit argument being made here about the relationship between marriage, lexical constraint, and semantic surplus. As we move into the interior space of marriage, the linguistic world collapses even as the semantic resonance of that world increases."[22] His descriptive model adds new insight to the far-reaching effects of marriage, not only at the level of plot, but also of language in the novel.

Descriptive modeling builds on the pioneering work of John Tukey, who offered methods for "looking at data to see what it seems to say."[23] These methods included using quartile and mean measures to describe the spread of values in a dataset (now enshrined in most statistical software packages as a "five number summary" or "Tukey summary" function) and using simple visualizations to reveal patterns within the data. Tukey's book contains instructions on creating stem and leaf plots and box plots by hand; today, data visualization techniques are widely available in spreadsheet software and online tools as well as in different programming languages. Exploratory analysis focuses on identifying trends and patterns in the data, which can often be better understood through graphical representation than in data tables alone.

For example, Moretti compiles bibliographic data and performs an exploratory analysis of changes in novel titles over the course of a century.[24] He presents twenty-one scatter plots and bar graphs that allow the reader to examine changes in title length, and how they relate to the numbers of novels published per year; the use of proper names in short titles; and a comparison of the use of definite and indefinite articles in novels from different genres. These data visualizations provide the overall view of the corpus, juxtaposed with Moretti's discussion of illustrative examples. Because most of the titles he mentions come from what Margaret Cohen called "the great unread," it is the visualizations that allow readers to compare his findings to what they already know about literature from the period and to make sense of the patterns in the data for themselves.[25] In exploratory analysis, data visualizations help construct the argument, rather than merely illustrating it.

Another example of exploratory modeling in literary studies is Matthew Wilkens's map of geographical references found in a collection of 100 nineteenth-century American novels that were digitized as part of the

[22] Ibid., 56.
[23] John W. Tukey, *Exploratory Data Analysis* (Reading, MA: Addison-Wesley, 1977), 2.
[24] Moretti, "Style, Inc.: Reflections on Seven Thousand Titles (British Novels, 1740–1850)," *Distant Reading*, 179–210.
[25] Margaret Cohen, *The Sentimental Education of the Novel* (Princeton: Princeton University Press, 1999), 23; Moretti, *Distant Reading*, 180.

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

Wright American Fiction Project.[26] Although one could simply analyze the number of references to each major city, Wilkens also chose to visualize the data on a world map. Exploring this data by mapping those locations reveals that "the imaginative landscape of American fiction in the mid-nineteenth century appears to be pretty diversely outward looking" when "our standard model of the period is that its fiction is strongly introspective … concerned largely with American identity and belonging."[27] In other words, when visualized, the data tells a different story than what literary critics have assumed to be true.

4 Generative Modeling

Topic modeling, a kind of generative modeling, is also widely used in distant reading to explore the semantic contents of large corpora. Generative models create a probabilistic model of how a set of data came to be distributed, and then use that model to describe the structure of the dataset or to generate new data instances. As David M. Blei explains, topic models "reverse the imaginary generative process to answer the question 'What is the likely hidden topical structure that generated my observed documents?'"[28] The most commonly used form of topic modeling in the humanities is Latent Dirichlet Allocation (LDA) topic modeling, which assumes that documents in a corpus are made up of topics, each of which are present in every document in varying proportions.[29] The model also assumes that topics are made up of words, which are also described by a probability distribution. It is as if "first, the writer picks a 'mixture' of topics to write about; then the writer constructs the text word-by-word by first randomly choosing a topic from the mixture and then picking a word based on the probability table for that topic."[30] Even though this does not sound like the way authors actually write texts, this model enables the algorithm to discover sets of co-occurring terms.

[26] Matthew Wilkens, "Canons, Close Reading, and Evolution," in *Debates in the Digital Humanities*, ed. Matthew K. Gold (University of Minnesota Press, 2012), open access edition: http://dhdebates.gc.cuny.edu/debates/text/17 (accessed April 22, 2022).

[27] Ibid.

[28] David M. Blei, "Topic Modeling and Digital Humanities," *Journal of Digital Humanities* 2.1 (2012): http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/#topic-modeling-and-digital-humanities-by-david-m-blei-n-3 (accessed April 22, 2022).

[29] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (2003): 993–1022.

[30] Jeffrey M. Binder, "Alien Reading: Text Mining, Language Standardization, and the Humanities," in *Debates in the Digital Humanities*, ed. Matthew K. Gold and Lauren F. Klein

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

The algorithm uses an unsupervised iterative process by which it assigns each word in a document to a topic, gradually refining the assignments to correspond to the co-occurrence patterns it finds in the corpus. This iterative process gradually uncovers distinctions between different meanings of the same word through their collocations. For example, imagine a corpus in which the word "bats" appears 100 times. On a first pass, all of the "bats" might be grouped together in one topic; after several iterations through the documents, some instances of the word "bats" will be sorted into a topic about baseball and others into one about mammals.

Although these lists of co-occurring words output by the LDA algorithm are called "topics" in the scientific literature, they do not exactly resemble the phrases or sentences that human readers and writers would call themes or topics. They are simply lists of co-occurring words, which reveal different kinds of textual relationships. For example, in their examination of several different topic models of articles from over eighty years of the journal *PMLA*, Andrew Goldstone and Ted Underwood demonstrate how the algorithm identifies the presence of different national languages within the corpus, broad subject categories, and discursive patterns that reflect both critical methodologies and also more subtle linguistic "habits": "A topic model doesn't just show you what people are writing about (a list of 'topics' in our ordinary sense of the word). It can also show you how they're writing … topic modeling can identify discourses as well as subject categories and embedded languages."[31] As they suggest, words in LDA topics do not necessarily mean what we might think they mean, "forcing us to attend to concrete linguistic practice" rather than assuming that texts fit with our preconceived ideas about their contents.[32]

The benefits of topic modeling are perhaps best seen in its applications to very large corpora that could not feasibly be interpreted by a single human reader. In an influential early study, Sharon Block modeled 82,000 newspaper articles and advertisements from the eighteenth-century *Pennsylvania Gazette*, exploring "how print discourses operated to marginalize particular groups in early America" and demonstrating the utility of topic modeling for examining discursive trends over time.[33] Two years

(Minneapolis: University of Minnesota Press, 2016), open access edition: https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities-2016 (accessed April 22, 2022).

[31] Andrew Goldstone and Ted Underwood, "What Can Topic Models of *PMLA* Teach Us about the History of Literary Scholarship?," *Journal of Digital Humanities* 2.1 (2012): http://journalofdigitalhumanities.org/2-1/ (accessed April 22, 2022).

[32] Ibid.

[33] Sharon Block, "Doing More with Digitization," *Common-place* 6.2 (2006): http://commonplace.online/article/doing-more-with-digitization/ (accessed April 22, 2022).

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

after their study of *PMLA*, Goldstone and Underwood used topic modeling to explore long historical trends in the field of literary studies, represented by 21,367 full-length articles published in seven scholarly journals over 120 years.[34] Lauren Klein constructs a topic model of 224,160 North American newspaper articles published between 1827 and 1865 in order to analyze the impact of particular editors in shaping abolitionist discourse.[35]

The LDA algorithm creates two kinds of outputs that can be used to explore and assess the model's results: a list of topics in each document, ranked by probability, and a list of the words in each topic, ranked by probability. Researchers tend to focus on the top-ranked topics and top-ranked words, as that is where meaningful semantic distinctions can be observed. In the 2003 paper that introduced the LDA algorithm, Blei et al. displayed four lists of top-ranked words from a topic model of 16,000 articles from the Associated Press, labeling the topics as "Arts," "Budgets," "Children," and "Education."[36] This practice of labeling topics with simple one- or two-word names based on domain knowledge has become widespread in subsequent topic modeling research, in part because it helps make broad semantic distinctions clear. As Block notes, "Readers may quibble over the exact topic label, but should be able to see the relationship between the words listed in each topic."[37] Yet this practice is not without controversy, particularly when applied to literary, rather than journalistic or critical writing.

Matthew Jockers presents selected results from his 500-topic model of a corpus of 3,346 nineteenth-century novels by displaying the top-ranking words for each topic as word clouds, with the words sized by their probability rank, and a companion website to the book allows viewers to explore data visualizations for all 500 topics.[38] The word clouds allow the viewer to situate any particular word among "the much larger company of other words that the model returns," and to explore how each word functions differently in the context of each topic.[39] But because Jockers also applies a label to each topic, such as "Native Americans" and "Ireland," the reader is guided to understand the results as perhaps more cohesive than any topic

[34] Andrew Goldstone and Ted Underwood, "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us," *New Literary History* 45.3 (Summer 2014): 359–384.
[35] Lauren F. Klein, "Dimensions of Scale: Invisible Labor, Editorial Work, and the Future of Quantitative Literary Studies," *PMLA* 135.1 (2020): 37.
[36] Blei et al., "Latent Dirichlet Allocation," 1009.
[37] Block, "Doing More with Digitization."
[38] Jockers, *Macroanalysis*, 126. The companion website is: www.matthewjockers.net/macroanalysisbook/macro-themes/ (accessed April 22, 2022).
[39] Ibid., 127.

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

model actually can be. Jeffrey Binder cautions against trying to interpret the words output in a topic model as if they hold meaning in the same ways that words do in an individual text, because each is an "abstract composite" representing numerous different, though semantically related, uses of the word.[40]

Jockers discusses in detail the steps he took in preparing his corpus of novels for topic modeling, including the "controversial step" of modeling only the nouns in the novels, in order to produce "highly interpretable and thematically coherent topics."[41] The clarity and elegance of Jockers's results are usefully balanced by his thorough technical discussion, which explains how each step in selecting and preparing the data, and configuring the parameters of the model affects the coherence of the results. As he points out, literary critics discuss "the specific elements of a text that we consider to be the interpretable elements"; the advantage of topic modeling is that it always also shows us what does not fit into our assumed interpretive framework.[42]

In her groundbreaking work on topic modeling poetic language, Lisa M. Rhody points out that the fundamental relationship between word and meaning is configured differently in poetry than in informational texts, because "the use of highly figurative speech actually increases the scope of the language one might expect to see in a document."[43] The expectation that co-occurring words relate to the same semantic topic could be confounded by metaphor and other devices that expand the signification of poetic texts.

In keeping with this interpretive caution, Rhody consistently refers to topics in her model by numbers, rather than by assigning her own interpretive labels to them. In her sixty-topic model of a corpus of 4,500 poems, she finds four types of topics: those relating to OCR errors or other linguistic variations, such as the presence of foreign languages or dialect; topics corresponding to very long, repetitive poems; "semantically evident" topics; and "semantically opaque" topics. Even in those that appear "semantically evident," she warns that "the illusion of thematic comprehensibility obscures what is actually being captured by the topic model," arguing that rather than being simply "about night," poems such as Anne Sexton's "The Starry Night" "participate in discourses surrounding 'night' and 'natural

[40] Binder, "Alien Reading."
[41] Jockers, *Macroanalysis*, 131.
[42] Ibid., 130.
[43] Lisa Rhody, "Topic Modeling and Figurative Language," *Journal of Digital Humanities* 2.1 (2012): http://journalofdigitalhumanities.org/2-1/ (accessed April 22, 2022).

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

landscapes.'"[44] Examining the poems associated with more opaque topics reveal the subtle discourses of poetry's history: "To say that Topic 2 is about 'death, loss, and internal turmoil' is overly simplistic … to say that Topic 2 draws from the language of elegy would be more accurate."[45] As Rhody demonstrates, topic modeling can be valuable for exploring literary corpora only if used with an awareness of the methodological and linguistic assumptions that underlie the method.

5 Predictive Modeling

In supervised machine learning, one of the most widely used techniques in predictive modeling, a machine learning algorithm is trained with a portion of the dataset in which each instance has been labeled according to two or more predefined categories. The model is then tested against the remaining portion of the dataset to see how accurately it can predict the category labels of those instances. A successful model may then be used to predict the labels for additional, similar data. Predictive models are commonly used in situations where the prediction itself is of value, such as an algorithm that examines the words used in an email to guess the likelihood of its being spam.

Although the scientific literature refers to disagreements between the model's label assignment and the actual label as misclassifications, in computational humanities scholarship they are often seen as a valuable source of insight, rather than treating them as errors. In distant reading, the predictions made by such models are used to better understand critical categories such as gender, genre, and historical period. As Andrew Piper suggests, "Predictive models allow us to say with how much certainty we can identify texts that belong to a specific group and under what criteria."[46] Because the model estimates the likelihood of each label it assigns, it offers a way to compare different categorizations of the same data.

An influential early example was Matthew Jockers's use of predictive modeling with a dataset of 106 nineteenth-century novels from twelve different fiction subgenres, such as historical, religious, silver-fork, Gothic, and industrial novels. Jockers demonstrates that the model could accurately predict the genre of a text 67 percent of the time using the relative frequency counts of only the highest-frequency word and punctuation

[44] Ibid.
[45] Ibid.
[46] Piper, *Enumerations*, 101.

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

features from the novels, such as articles and pronouns, rather than words that are directly tied to theme or setting.[47] Jockers argues that when dealing with classes such as genre, which "are human inventions that are not always clearly delineated," researchers ought "to consider both the first- and the second-place assignments when assessing the overall performance and value of the model," noting that in some cases these two probabilities may be very close.[48] If two novelistic genres are frequently labeled as each other, for example, he suggests there might be a "stylistic affinity" between those genres that further research could uncover.[49] A frequent misclassification might also indicate that there are confounding factors in the experiment. Texts are not only produced as examples of genres; they are written by particular authors, in particular historical moments, in particular places. Each of these factors might produce a stronger textual signal than that of genre.

In a second experiment with the same dataset of 106 novels, Jockers compared how successful predictive models were in predicting the authorship, author gender, publication decade, genre, or specific novel title for segments of the texts. Jockers found that the models for author signal were significantly more accurate than those for other categories, suggesting that "classifications based on loosely defined genres produce comparatively poor results when viewed next to the results for a more objectively defined category."[50] In a dataset of only 106 novels spread over twelve genres, authorship would significantly overlap with the genre categories, making them difficult to distinguish.

In their study of English-language haiku, Hoyt Long and Richard So use predictive modeling to examine "whether the machine recognizes a haiku over and against a text that is nonhaiku and if so, what statistical evidence it uses to make that decision."[51] To do this, they compiled a corpus of 400 English haiku, both translations and adaptations, and 1,900 short poems published in periodicals during the early twentieth century, and ran a series of experiments that trained the classification model with different language features from the poems. A model that included syllable counts and excluded function words was able to correctly identify haiku translations 91 percent of the time and haiku adaptations 86 percent of

[47] Jockers, *Macroanalysis*, 72.
[48] Ibid., 73.
[49] Ibid., 73.
[50] Ibid., 92.
[51] Hoyt Long and Richard So, "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning," *Critical Inquiry* 42.2 (Winter 2016): 252.

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

the time. But like Jockers, Long and So are more interested in examining the misclassifications. Another model that ignored syllable counts and included function words increased the number of nonhaiku poems that the model classified as haiku. Long and So examine the 202 poems that were most frequently misclassified by the model as haiku, organizing them into three groups: "haiku in waiting," "potential haiku," and "machine haiku," suggesting that rather than simply defining these misclassifications as errors, critics can productively use them to "reconsider our own interpretive biases."[52]

Ted Underwood suggests that "a mode of inquiry founded on models can study relationships rather than isolated facts," allowing researchers to connect sociocultural phenomena with textual change.[53] Genre is a key focus of his study, in part because its defining features change over time. What readers in the 1890s recognized as detective fiction is different from what readers in the 1990s expected in detective fiction. Among the many examples discussed in his book, two exemplify Underwood's thoughtful deployment of predictive models and experimental thinking.

He shows that a useful way to test the validity of specific genre models is to test them on different corpora. For example, he tests a model trained on detective fiction to see if it can identify works of science fiction. If the model is truly identifying something distinctive in the texts it identifies as detective fiction, one would expect that experiment to fail, and indeed it does. He calls this "the method of mutual recognition," as it can be run in both directions to show how distinctive different subgenres of fiction are.[54]

To test critical assumptions about genre and literary history, Underwood develops a method he calls "perspectival modeling": a "strategy that gets leverage on cultural history by comparing models trained on different subsets of the evidence."[55] Because a supervised model "describes one set of texts through their relation to another," he examines changes within a specific genre such as science fiction by asking a model trained on early texts to classify texts from later decades, and vice versa.[56] Underwood's experiments suggest that although traditional critical accounts of fictional genres such as detective fiction, Gothic fiction, and science fiction emphasize change and discontinuity, in part because they focus on the influence

[52] Ibid., 262–265.
[53] Ted Underwood, *Distant Horizons: Digital Evidence and Literary Change* (Chicago: University of Chicago Press, 2019), xii.
[54] Ibid., 46–47.
[55] Ibid., 36.
[56] Ibid., 61–65.

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

of individual authors, a distant reading approach using predictive models with large corpora shows more continuity within genres over the nineteenth and twentieth centuries.

6 Tools for Distant Reading

In the early to mid-2000s, AJAX (Asynchronous JavaScript and XML) programming fostered the development of what became known as Web 2.0, which allowed for user interactivity, including user-generated content on blogs and wikis, social media, and the creation of web-based software applications.[57] These technologies emphasized participation and leveled the technological playing field: "Users could 'write' the Web in the browser itself, rather than using separate tools that required additional technical knowledge."[58] As these technologies have continued to develop, the ability to combine clientside data with serverside processing has allowed users to explore, analyze, and remix text, audio, and visual data. A large number of digital humanities tools, including both installed software packages and web-based applications, have been developed for text analysis, geospatial analysis, network analysis, and data visualization.[59] Although most of the scholarship cited earlier was performed in programming languages such as R or Python, today there are several robust toolsets that allow scholars and students without programming backgrounds to begin experimenting with distant reading. Toolsets such as Voyant Tools, Lexos, and Orange can be used as a low-stakes way to learn about its methods and insights and contribute to new knowledge. All three of these toolsets can be used for analyzing large text corpora and for data visualization (exploratory modeling); Voyant and

[57] Susannah Fox and Mary Madden, "Riding the Waves of 'Web 2.0.'" Pew Internet Project, October 5, 2006: www.pewresearch.org/internet/2006/10/05/riding-the-waves-of-web-2-0/ (accessed April 22, 2022).

[58] Doug Belshaw, "A Unified 20-Year History of the Radically Changing Way We Relate to the Web," *Washington Post*, December 3, 2014: www.washingtonpost.com/news/the-intersect/wp/2014/12/03/a-unified-20-year-history-of-the-radically-changing-way-we-relate-to-the-web/ (accessed April 22, 2022).

[59] Unfortunately, a number of these tools have since suffered the effects of terminated funding or have simply become technologically obsolete. Even some of the major directories of tools, including DiRT and Project Bamboo, have met the same fate; a current directory as of this writing is TERESAH, the Tools E-Registry for E-Social science, Arts and Humanities. See Quinn Dombrowski, "What Ever Happened to Project Bamboo?" *Literary and Linguistic Computing*, 29.3 (September 2014): 326–339: https://doi.org/10.1093/llc/fqu026 (accessed April 22, 2022); "DiRT: Digital Research Tools Wiki," last modified 1 September 2012: https://digitalresearchtools.pbworks.com/w/page/17801672/FrontPage (accessed April 22, 2022); and "TERESAH: Tools E-Registry for E-Social science, Arts and Humanities," Data Service Infrastructure for the Social Sciences and Humanities (DASISH): http://teresah.dariah.eu/ (accessed April 22, 2022).

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press

Orange offer workflows for LDA topic modeling (generative modeling); and Orange offers workflows for text classification (predictive modeling).[60]

Finally, an extensive array of resources exists for users interested in learning enough programming in R or Python to analyze texts in the ways surveyed in this chapter. In addition to general-audience books, courses, and online tutorials, some are specifically targeted at humanities scholars and students.[61]

Before the 1980s, humanities computing was largely restricted to those with access to university mainframe computing resources, and data storage limits and the necessity of hand-keying textual data imposed significant constraints on the creation of textual corpora. Today, open access archives of digital texts combine with powerful personal computers and online tools to put distant reading within reach of anyone with an experimental mindset.

[60] Stéfan Sinclair and Geoffrey Rockwell, "Voyant Tools," version 2.5.4, February 11, 2022: https://voyant-tools.org/ (accessed April 22, 2022); "Lexos," Wheaton Lexomics, 2019: http://lexos.wheatoncollege.edu/upload (accessed April 22, 2022); "Orange," University of Ljubljana: https://orangedatamining.com/ (accessed April 22, 2022).

[61] Taylor Arnold and Lauren Tilton, *Humanities Data in R: Networks, Geospatial Data, Images, and Text* (Cham: Springer, 2015); Matthew L. Jockers, *Text Analysis with R for Students of Literature*, 2nd ed. (Cham: Springer, 2020); and Folgert Karsdorp, Mike Kestemont, and Allen Riddell, *Humanities Data Analysis: Case Studies with Python* (Princeton: Princeton University Press, 2021).

https://doi.org/10.1017/9781108560740.023 Published online by Cambridge University Press