

A NEW COMPANION TO

DIGITAL HUMANITIES

EDITED BY
SUSAN SCHREIBMAN, RAY SIEMENS,
AND JOHN UNSWORTH

WILEY Blackwell

References and Further Reading

Clement, T. 2013. Text analysis, data mining, and visualizations in literary scholarship. In *Literary Studies in the Digital Age: A Methodological Primer*, ed. K. Price and R Siemens. New York: MLA Commons.

Cordell, R. 2014. On ignoring encoding. http://ryancordell.org/research/dh/on-ignoring-encoding (accessed June 20, 2015).

Drucker, J. 2011. Humanities approaches to graphical display. *DHQ: Digital Humanities Quarterly* 5 (1).

Fischer-Baum, R., Gordon, A., and Halsley, B. 2014. Which words are used to describe white and black NFL prospects? *Deadspin*. http://deadsp.in/1iNz1NY (accessed June 20, 2015).

Foster, D. 2000. *Author Unknown: On the Trail of Anonymous*. New York: Henry Holt and Company.

Harris, J. 2011. Word clouds considered harmful. *Nieman Journalism Lab*. http://bit.ly/QKNMdD (accessed June 20, 2015).

Hockey, S. 2000. *Electronic Texts in the Humanities*. Oxford: Oxford University Press.

Hockey, S. 2004. The history of humanities computing. In *A Companion to Digital Humanities*, ed. S. Schreibman, R. Siemens, and J. Unsworth. Oxford: Blackwell, 2004. http://www.digitalhumanities.org/companion (accessed June 20, 2015).

Jockers, M. 2013. *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.

Kenny, A. 1982. *The Computation of Style*. Oxford: Pergamon Press.

Kristeva, J. 1968. Le texte clos. *Langages* 3 (12).

Kristeva, J. 1980. *Desire in Language: A Semiotic Approach to Literature and Art*. Trans. T. Gora, A. Jardine, and L.S. Roudiez. New York: Columbia University Press.

McGann, J. 2001. *Radiant Textuality: Literature After the World Wide Web*. New York: Palgrave.

Michel, J.-B., Shen, Y.K., Aiden, A.P., *et al*. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331 (6014), 176–82.

Moretti, F. 2005. Graphs, Maps, Trees: Abstract Models for a Literary History. London: Verso.

Radicati Group. 2014. Email statistics report, 2014–2018. http://bit.ly/1o6GmQA (accessed June 20, 2015).

Ramsay, S. 2011. *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press.

Ramsay, S. 2014. The hermeneutics of screwing around; or what you do with a million books. In *Pastplay: Teaching and Learning History with Technology*, ed. K. Kee. Ann Arbor: University of Michigan Press.

Renear, A. 2004. Text encoding. In *A Companion to Digital Humanities*, ed. S. Schreibman, R. Siemens, and J. Unsworth. Oxford: Blackwell. http://www.digitalhumanities.org/companion (accessed June 20, 2015).

Ruecker, S., Radzikowska, M., and Sinclair, S. 2011. Visual Interface Design for Cultural Heritage: a Guide to Rich-Prospect Browsing. Burlington, VT: Ashgate.

Sinclair, S. 2003. Computer-assisted reading: reconceiving text analysis. *Literary and Linguistic Computing* 18 (2).

Sinclair, S., and Rockwell, G. 2012. Teaching computer-assisted text analysis: approaches to learning new methodologies. In *Digital Humanities Pedagogy: Practices, Principles and Politics*, ed. B.D. Hirsch. Cambridge: OpenBook. http://www.openbookpublishers.com/htmlreader/DHP/chap10.html#ch10 (accessed June 20, 2015).

Sinclair, S., and Rockwell, G. 2014. *Voyant Tools*. http://voyant-tools.org (accessed June 20, 2015).

Sinclair, S., Ruecker, S., and Radzikowska, M. 2013. Information visualization for humanities scholars. In *Literary Studies in the Digital Age: A Methodological Primer*, ed. K. Price and R Siemens. New York: MLA Commons.

Smith, C. 2011. Word cloud: how toy ad vocabulary reinforces gender stereotypes. *The Achilles Effect*. http://bit.ly/1osjjji (accessed June 20, 2015).

Tufte, E. 2001. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Wang, X., and Inaba, M. 2009. Analyzing structures and evolution of digital humanities based on correspondence analysis and co-word analysis. *Art Research* 9. http://bit.ly/1jLWnbX (accessed June 20, 2015).

20
Text-Mining the Humanities

Matthew L. Jockers and Ted Underwood

Why Mine?

In the humanities, more often than not, the focus of scholarly attention is on the details – often subtle and nuanced details that are revealed only through slow, thoughtful, close reading. The usual method of analysis is one largely driven by synthesis. Scholars read and make associations; they discover and reveal what is not obvious to the casual reader, and to the extent that computation is leveraged in this activity, it is usually at the level of simple keyword search: a scholar wonders about Melville's thoughts on God and then performs a search for *God* in the digital text in order to find passages that will be studied under the microscope of informed close reading. Computers are very good at this task, and for some types of questions computational keyword searching is all that is warranted. But what of other questions, questions of scale that have not been asked until quite recently?

In 1988, Rosanne Potter wrote that "until everything has been encoded, or until encoding is a trivial part of the work, the everyday critic will probably not consider computer treatments of texts" (93). She was right, and the same might be said for the tools as well. Until everything has been digitized, why bother building tools to analyze them? Potter's "everyday critic" could not imagine computer treatments of texts because the texts did not exist, and even if they had existed, the state of the tools was such that those critics would not have been likely to make huge discoveries. In fact, a number of scholars – including several who are sympathetic to the digital humanities – have in the past argued exactly this point: namely, that computer treatments of texts have had little impact on the mainstream humanities. Mark Olsen wrote in 1993 of how "computerized textual research has not had a significant influence on research in

A New Companion to Digital Humanities, First Edition. Edited by Susan Schreibman, Ray Siemens, and John Unsworth.
© 2016 John Wiley & Sons, Ltd. Published 2016 by John Wiley & Sons, Ltd.

the humanistic disciplines" (309), and Stephen Ramsay, in 2007, of how the "digital revolution, for all its wonders, has not penetrated the core activity of literary studies" (478). But already these comments feel rather antiquated, and they feel so exactly because the pace of change, even since 2007, has been so incredibly rapid. Not only do we now have massive digital archives,[1] but we also have new and sophisticated tools for studying them. And the importance of the tools should not be underestimated. Tim Lenoir has argued rhetorically, but convincingly, that quarks would not exist today were it not for the particle accelerators that were built to discover them.[2] Some of the most sophisticated and most promising new tools for text analysis and text mining have only recently come to the attention of scholars in the humanities, and embracing them, leveraging them, means that humanities scholars must learn from research in the seemingly unrelated fields of natural language processing and machine learning.

Background

Quantitative approaches have a long history in the humanities, but contemporary text mining is also a deeply interdisciplinary project with affinities to computer science, statistics, linguistics, sociology, and other social sciences. In the space available here, we can only sketch a few important lines of development.

As John Unsworth (2013) has pointed out, quantitative analysis of text has a history stretching back to the nineteenth century. Quantification was often understood as a way of getting at something called "style," either in order to understand the history of style writ large, as in L.A. Sherman's *Analytics of Literature* (1893), or in order to identify works by a particular author, as in the research of T.C. Mendenhall (1887, 1901). In the twentieth century, the project of authorship attribution came to be closely associated with the more general and varied practice of "stylometry," and it remains an important aspect of text mining today. Twentieth-century linguists approached style as a social phenomenon, particularly in a subfield called "stylistics." Stylistics, in turn, overlaps with quantitative approaches to language that don't necessarily characterize their object of inquiry as "style" – for instance, with corpus linguistics, which uses collections of samples (corpora) to describe real-world linguistic variation.[3]

The phrase *text mining* itself is modeled on *data mining*, an informal name for a subfield of computer science also known as knowledge discovery in databases (KDD). Coalescing in the late 1980s, this field emerged from the broader project of artificial intelligence, and especially from efforts to model and automate learning processes. The terms *KDD*, *data mining*, and *machine learning* are bound together in a complex topology, and it is not easy to separate intellectual history from prescriptive definition.[4] Today, *data mining* often implies unsupervised learning, whereas *machine learning* is more commonly applied to supervised learning processes (see below). But this boundary can be drawn in several different ways: sometimes *data mining* names the practice that corresponds to *machine learning*'s theory.

Textbooks on data mining often include a chapter on text mining, which is seen by computer scientists as a subfield devoted to the extraction of knowledge from unstructured text.[5] But in humanistic practice, text mining is an interdisciplinary endeavor that also borrows freely from corpus linguistics and computational linguistics, as well

as social-scientific traditions like social network analysis. Perhaps most importantly, humanistic text mining seeks to frame questions that contribute meaningfully to existing traditions of humanistic inquiry. Given this complex confluence of disciplines, it is not surprising that controversies about text mining commonly involve differences of opinion about the relative weighting of different disciplinary methodologies.

Methods: Machine Learning and Text Mining

The terms *text* (or *data*) *mining* and *machine learning* are frequently conflated and sometimes confused but do represent two different practices. Generally speaking *mining* is applied to techniques focused on exploration and discovery whereas *machine learning* refers to techniques or methods that are designed for prediction. The former is generally referred to as *unsupervised* learning and the latter as *supervised* learning. At a deeper level of specificity, these kindred practices may be called machine *clustering* and machine *classification*. The simplest way of differentiating between them is to consider the role of the researcher and whether or not that researcher has advanced and specific knowledge of the structure and composition of the data.

In machine *clustering*, for example, we do not have a preconceived notion of how the data is or might be organized and do not pre-label the individual data points as belonging to one group or another; the objective is to discover hidden structure in data by machine grouping, or clustering, the data objects based on the similarity of their features. If we were clustering shapes, for example, we might have a feature called "number of sides." Given this data about the features of these shapes, an unsupervised algorithm might cluster three-sided objects into one pile and four-sided objects into another. The machine would not, however, be given information about these *classes* of shapes in advance. The machine is only given the features and attempts to group the objects into categories or classes based on analysis of the features.

In text mining, we frequently wish to group documents together according to their similarity. Similarity is often based on, or measured by, some finite set of textual features, such as the relative frequency of the most frequently occurring words. If we are interested in clustering texts according to the similarity of their style, for example, we know from years of authorship-attribution research that the most effective features for distinguishing one author's style from another's are high-frequency features such as the words "the," "of," "him," "her," and "and," as well as common marks of punctuation.[6] But, of course, machine clustering can be used for much more than authorship analysis. Say we are interested in exploring the extent to which Irish authors have a distinct literary-linguistic style.[7] For the sake of illustration, we constructed a random sample of 300 nineteenth-century novels: 100 by British authors, 100 by Irish authors, and 100 by American authors.[8] For each novel we calculated the total number of instances (the *tokens*) of each unique word and mark of punctuation (the *types*). We then divided the raw count of each feature in a given text by the total number of words in the text in order to calculate the relative frequency of each word type.

This information can be represented as a data matrix in which each row is a text (in the nomenclature of machine learning, each text is an *observation*) and each column a different word type (each word is a *feature*). In this example, the resulting matrix was 300 rows by

154,312 columns. Since we are interested in computing stylistic similarity based on the use of high-frequency features, this matrix was then reduced by keeping only those features with a mean relative frequency across the corpus of at least 0.1. This *thresholding* resulted in a new matrix of 107 features.[9] With the matrix reduced in this manner, the machine was then configured to cluster the texts using this set of 107 features and a measure of similarity called Euclidean distance.[10] The result of such clustering can be visualized as a tree dendrogram, and since the dendrogram is hierarchical, it is possible to identify groups by "cutting" the tree at specific points. Figure 20.1 shows a representation of the full tree and – through the use of shading and dotted branch lines – shows three distinct clusters. The cluster in black (Cluster Two) contains 36% of the Irish texts and exactly 0% of the British and 0% of the Americans (Table 20.1).

In a world where there were perfectly distinct stylistic differences between the three nations, splitting the dendrogram at three branches would have resulted in a perfect separation of the three nationalities: one cluster of 100 Irish-authored novels, another containing the 100 British texts, and a third of 100 Americans. Instead of perfect separation, we observe some mixing of the texts, but the absence of any British and American texts from Cluster Two is suggestive; it suggests that there is indeed something distinct about at least 36% of the Irish novels in the corpus: the aggregate *signal* – expressed through the 107 features in these 36 books – is not at all like the *signal* typical to American and/or British novels.

As noted previously, unsupervised clustering is often viewed as an exploratory method in which we seek to uncover some hidden structure in the data. A researcher

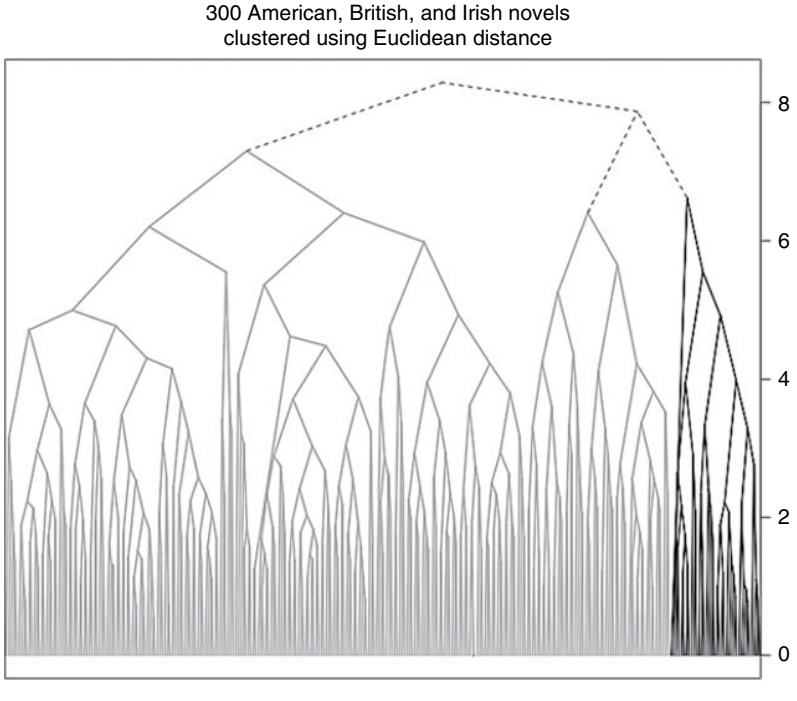

Figure 20.1 Cluster dendrogram, with Irish cluster shaded black.

Table 20.1 Three-cluster test.

	Cluster One	Cluster Two	Cluster Three
American	15	0	85
British	24	0	76
Irish	18	36	46
Total in cluster	57	36	207

observing this result might now go back to the data and examine how the features in these 36 Irish books differ from those Irish-authored books found in the other clusters. Upon deeper inspection, it may be discovered that some other factor such as religion, or class, or gender is responsible for the separation. The results of this test might also be considered good enough to warrant further testing using supervised methods of classification.

In *supervised document classification*, a researcher establishes, in advance, a set of known text classes and then writes a program to classify unseen documents based on the similarity or difference between the unseen text and the known classes of documents. The classic example of this is the authorship attribution problem. A document of unknown or uncertain authorship is processed and classified according to its statistical similarity to a known author within a closed set of candidates. As with unsupervised clustering, the greatest difficulty often comes in having to determine the features that will be compared. Say, for example, you have a closed set of two-dimensional objects: squares, triangles, and circles. *Closed* here means that these are the only types of objects that will be considered. Each of these classes is defined by some set of features. For this example, let's assume there are only three features: "shape size" (measured in area), "shape color," and "number of sides." We begin by gathering many examples from each class, and for each we extract the feature information for these three features. Assume that we have 100 different triangles, 100 different squares, and 100 different circles. For each object, we extract three data points: the area of the object, the color of the object, and the number of sides. This data are then fed into a classifier: a computer algorithm designed to identify statistical differences between the classes. In this case, the classifier is told in advance (supervised) that these shapes are representative of three distinct classes of two-dimensional objects. The machine "knows" which shapes are which and is able to use that information to organize the data. In this example, the classifier examines all of the data and identifies, or figures out, that the *number of sides* feature is incredibly useful in differentiating between the classes.[11] The classifier identifies that the *area* and *color* features seem to fluctuate randomly between classes and therefore finds no correlation between shape *size* and shape class or between shape *color* and shape class. In this way the machine is *trained* to recognize which types of object are members of which predefined classes. Once trained in this manner, the machine can be given a new object and asked to classify it according to its similarity to the known classes. If the new object happens to be an oval, this hypothetical classifier would guess that (i.e., *classify*) the oval is most likely a circle: like the circle, the oval only has one side.[12] If we gave this classifier a rectangle, it would be classified as a square: though not exactly the same as a square (i.e., with sides of equal length), the

rectangle is most like the square in having four sides. Naturally, in text classification, the problems are a lot harder because the feature sets are much larger than just three features. An important point worth emphasizing here, however, is that regardless of whether the problem is simple or hard, we can perform a test of the model in order to determine how well it is performing. This is not something we can do in unsupervised clustering where we have not already established a predefined set of classes. In classification, we execute this performance test by first training a model using a randomly selected subset of the total observations and then testing the model by seeing how well it classifies the remaining held-out samples.[13]

Consider the example used above in which we clustered novels and then examined whether author nationality could be seen as an explanatory factor in how the books clustered. Using this same corpus, a classification experiment can be constructed in which a classifier is trained to identify American, British, and Irish novels based on the same 107 features used in the clustering. The process begins by having the computer randomly select a subset of the novels from each nation for training. For this example, we used 66% of the American-authored books, 66% of the British and 66% of the Irish; the remaining 34% were held out for testing the trained model. In supervised classification, unlike the unsupervised clustering done above, we (and the machine) already know the possible groups or classes (e.g., triangles, circles, and squares, or American, British, and Irish novels), and we want the computer to take a particular object that it has never seen before and classify it into one of these known groups. In order to do this, we provide the machine with information about the typical features of the squares, circles, and triangles (or the national literatures), by giving the computer lots of examples of these objects. We say, "here are some typical American books and here are some typical British and Irish books; take these objects and build a model that understands what features are most common or most typical to each class." From these examples the computer builds a model of what constitutes an American signal, a British signal, and so on. Then when the machine is given a new object, it looks at all of its features and figures out which of the known classes is most similar to the new object.

In this example classification test, we used the nearest shrunken centroid (NSC) classification algorithm, and we were able to accurately identify author nationality in the held-out data with an average accuracy of 71%.[14] Considering that chance in this experiment is 33%, the observed result of 71% is considerably better than what could be expected from a machine simply guessing at random. To really understand how the machine performed, however, it is useful to examine the confusion matrix and understand what is meant by *precision* and *recall*.

Table 20.2 shows the confusion matrix for this classification test. The confusion matrix is produced as part of the cross-validation routine that iteratively samples from the data in order to train and test a series of models. The first column in the table indicates the true class of each sample. So we begin with the row of data labeled with the class *American*. In this test, when the actual class was *American*, the machine guessed *American* 52 out of 66 times; these 52 we call the *true positives*. Five times it guessed *British* and nine times it guessed *Irish*; these we call the *false negatives*. We represent the *recall* (shown in the fifth column of the table) of this result by dividing the number of correct guesses by the combined number of correct guesses and incorrect guesses, in this

Table 20.2 Confusion matrix from classification test.

	American	**British**	**Irish**	**Recall**
American	52	5	9	0.79
British	0	44	22	0.67
Irish	1	20	45	0.68
Precision	0.98	0.64	0.59	

case 52/66. The resulting figure (0.79) tells us how good the classifier is at recognizing books that are written by American authors.

If we look at the second row, we notice that the machine had no problem guessing that British-authored books were not American. The model never guessed that the author of a British book was an American. About a third of the time (22 cases), however, the machine misrecognized a British book as Irish. So even though the machine assigned no British-authored books to American authors, it still only managed to correctly recognize 67% of British books. The third row shows the results for the Irish-authored books. One book was thought to be of American authorship and 20 were thought to be British. Here again, the model had little trouble differentiating Irish and American writing, and more difficulty differentiating Irish and British writing.

This difficulty will not come as a huge surprise: readers of these national literatures would expect to find greater affinity between British and Irish texts. But *recall* only tells part of the story. If recall tells us how often a given category was correctly recognized, *precision*, shown in the last row, tells us how often a given prediction was correct. Precision expresses the number of true positives (correct classifications) divided by the number of true positives *and* false positives. So, in the case of the *column* labeled *American*, we take the total number of true positives and divide by the sum of true positives and false positives. In this case, the precision is excellent: 52/(52 + 1) or 0.98. The model only guessed *American* incorrectly once. Put another way, when the machine guesses that a work is by an American author, it guesses correctly 98% of the time. Alternatively, when the machine guesses that a book is Irish-authored, it only guesses correctly 59% of the time.

In terms of overall classification accuracy, the results of this experiment, especially when viewed in the context of mere chance, were not too bad. Sometimes the overall accuracy of a classification experiment such as this will be generalized into what is called an *F-measure* or *F-score*, a measure that attempts to provide a weighted average of the precision and recall. In this example, the *F*-score is 0.72. Other times, researchers may simply describe the total number of correct classifications divided by the total number of cases: in this example, 0.71. Though we might consider these classification results fairly good, a lot about how we interpret these results depends upon the domain in which we are working and upon the stakes. If these were text samples from suspects in a murder case, we might not be satisfied with these cross-validation results. What is clear from this example, however, is that there is a fairly marked difference between American texts and those of Irish and British authorship. Given such a result, it can be productive to explore the actual features that the model found to best distinguish between the groups. Of the three groups in this test, for example, the Americans showed

a greater fondness for the *comma* and for the word *the*, the Irish were next, and then the British. British authors, on the other hand, showed a substantially higher preference for the *period*, followed in turn by the Americans, and then the Irish.[15]

Adapting Algorithms to Concrete Problems

In both examples above, the learning algorithms work on a list of features produced simply by selecting the most common words (or punctuation marks) in the corpus. Other feature-selection strategies are possible. A researcher can use all the words in a corpus as features, or use measures like *information gain* to identify the features that most effectively discriminate a particular pair of classes.[16] Feature-selection strategies may depend on the problem being explored: sets of common words distinguish authors effectively, but might be less effective at distinguishing subject categories. Finally, there are many things aside from words that can be treated as features: for instance, one can use two-word phrases (bigrams).

Expanding the frame of reference to include phrases raises a larger set of issues that we can only glance at briefly here. Newcomers to text analysis are often surprised that so many algorithms rely on a "bag-of-words" model, discarding word order in order to focus simply on the frequencies of individual words.[17] Our everyday experience of text is sequential, and intuitions based on that experience would suggest that the meaning of a text depends entirely on sequence. For some purposes this is true, but for many purposes word order is surprisingly dispensable. Classifiers using a bag-of-words model often perform just as well as algorithms that consider multiword phrases (Bekkerman and Allan, 2003). On the other hand, there are certainly kinds of analysis where word order becomes central. Natural language processing toolkits, which rely on supervised models, can tag each word with a part of speech, parse sentences to identify grammatical dependencies, and extract named entities (people, places, organizations, and so on).[18]

Introductory textbooks on data mining list a range of algorithms that are known to work well on text. Naïve Bayes, regularized logistic regression, and support vectors are often used for supervised classification; clustering may involve *k*-means, mixed-model, or hierarchical methods. Computer scientists and statisticians are constantly refining these algorithms to improve performance, and performance may also improve when multiple models are combined in an *ensemble*. But even specialists in machine learning recommend trying relatively simple methods first, because "small gains in laboratory performance … may be swamped by other factors when machine learning is applied to a practical data mining problem" (Witten *et al.*, 2011:377–8). The time a researcher might spend optimizing an algorithm is often better spent improving the data.

On the other hand, it is useful to understand how simple methods can be adapted to address a range of different problems. Classification algorithms do not necessarily have to characterize documents by predicting membership in a single class; in *multilabel classification*, for example, they can assign a document to multiple classes (perhaps characterizing each prediction with a different degree of confidence) (Tsoumakas and Katakis, 2007). Fuzzy and mixed-model clustering provide an unsupervised equivalent. When training data is relatively scarce, approaches like *active learning* can conserve

human effort by identifying selected cases where human guidance would help the algorithm improve its performance most significantly (Han *et al*., 2012:433–4). When a problem is imperfectly defined (often the case in the humanities), *semi-supervised learning* may be appropriate, since it allows a researcher to provide initial guidance without fully determining the range of possible classes in a model (Han *et al*., 2012:432). Finally, there are cases where humanists may need to collaborate with computer scientists in order to develop new methods appropriate to a particular domain. *Probabilistic graphical models* support this kind of innovation with a flexible language for representing human assumptions and translating them into algorithms.[19]

Challenges

There are more potential applications of text mining in the humanities than existing research projects, because projects unfortunately confront a number of significant barriers in the start-up phase. The main obstacle is the interdisciplinary character of the enterprise itself. The parts of the data-mining process that are easy to standardize generally have been standardized: implementations of popular algorithms are readily available in toolkits like *Weka* (Hall *et al*., 2009) and MALLET (McCallum, 2002). But tools are never a complete solution. Since every research question is different (almost by definition), each entails some problems that resist standardization. Idiosyncratic types of metadata need to be gathered, special-purpose analyses need to be performed, and results need to be translated into visualizations that address a specific question. Many of these steps are likely to require familiarity with programming and with the humanistic discipline being explored; some steps may also require knowledge of statistics. As a result, humanistic text-mining problems often call for interdisciplinary teams, or researchers with an unusual breadth of experience, or both.

It is a proverbial truth that data preparation often consumes more time than analysis. This can be true even in projects that begin with relatively structured data, since names and dates come in a variety of formats that may need to be standardized before meaningful comparison is possible. Preparation of unstructured text can be even more challenging, and researchers in the humanities confront special difficulties associated with historical change. As a reader travels back across the centuries, for instance, the rules of capitalization, word division, and spelling change. These changes could be viewed as subjects of linguistic inquiry in their own right. But a scholar who is studying the history of medicine rather than English spelling may want to ensure that "physic" and "physick" are treated as a single word. More debatably, a researcher might decide to treat occurrences of "any body" in the eighteenth century as equivalent to twenty-first-century "anybody." Normalization can be taken even further through processes of *stemming* or *lemmatization* – flattening the distinctions between possessives, plurals, and verb tenses to associate them with a single root.

Although we often speak casually about data "cleaning," decisions to standardize different aspects of text involve trade-offs that are far from straightforward. Details that count as *noise* in one context might become *signal* in another. Lemmatization, for instance, can improve the efficiency of search engines, but discards grammatical inflections that might be useful as clues about authorship or genre. For this reason, there is

no single agreed-upon process of data preparation, although resources do exist, for instance, to support normalization of spelling variants when researchers want to normalize them.[20]

In an ideal world, all texts would be available in accurate copies, marked up with TEI, or some similar standard, to distinguish footnotes and running headers from body text.[21] In reality, the texts in large digital libraries are usually transcribed by optical character recognition (OCR). The OCR process produces errors (particularly on older texts), and does not provide many explicit clues to distinguish body text from paratext. Confronted with these challenges, humanists sometimes despair. Our disciplines have taught us that identifying an accurate edition is the first step in responsible research; here it seems impossible. Alternatively, we may look for a fixed accuracy cutoff that would guarantee our results are "good enough."

A more useful approach to this problem might begin with the nature of the questions being posed, and their relation to specific kinds of error. On a macroscopic scale, truly random errors are not necessarily a big obstacle for analytical methods that consider words individually. If every word in a given language had a constant 10% chance of random mistranscription, volume-level classification and methods such as topic modeling might proceed almost undisturbed. Volume-level word counts are redundant, and since the random strings produced by mistranscription will be individually rare, they are likely to fall out of the analysis. Some errors are close to being this random: coffee spots, ripped corners. But problems arise because other errors are distributed unequally across a corpus. Paper quality or worn metal type make some volumes more prone to mistranscription than others. The worst problems are those that preferentially affect specific words in specific periods – for instance, in some eighteenth-century books using the notorious "long s," "ship" will almost always be mistranscribed as "fhip" or "flip." Left uncorrected, these systematic errors could distort analysis. Fortunately, there are algorithms we can use to correct OCR even in tricky cases like the confusion between "ship" and "flip" (Tong and Evans, 1996).

Since different kinds of error have radically different effects, there is no single accuracy percentage that proves a text is good enough to support analysis. But some general principles are clear. The small category of errors that occur often and systematically, because of ambiguous typefaces or ligatures, are more problematic than the much larger category of uncommon errors. (Collectively, uncommon errors become numerous, but collectively they also approximate randomness.) It follows that partial OCR correction, addressing a limited number of predictable errors, may be adequate for many research purposes. On the other hand, different kinds of research have different degrees of sensitivity. Bag-of-words methods are more robust to random error than natural language processing, where a single misspelling can make a whole clause hard to parse. More research is needed to establish the relative robustness of different methods. Research is also under way to support automatic separation of text from paratext (Underwood *et al*., 2013).

So far, we've focused on challenges specific to text mining. But some of the most important challenges confronting this work are those it shares with other fields of the humanities. By allowing a researcher to survey a larger set of documents, text mining promises to give a picture of the print record that is "more representative" than an

account based on a few hand-selected examples. But representative of what? Digital libraries, largely based on university libraries, do not include every book ever published. Moreover, even if we had a copy of every book, the print record itself would not reflect the demographic reality of the past, since access to print has been shaped by class, gender, and race. One could argue for a stratified corpus, rebalanced to redress these inequalities. More commonly, critics of text mining seize the other horn of the dilemma, suggesting that it would impose a misleading equality to count every title once. Perhaps popular and widely reprinted titles (or "important" titles) should carry more weight in the corpus (Rosen, 2011)?

None of these questions are new. They are versions of a debate that humanists have long pursued under the rubrics of "culture" and "canonicity" – a debate that is not likely to be resolved soon. Fortunately, it does not have to be resolved before we attempt macroscopic research. Text mining is not bound to any particular model of representation, and needn't presuppose consensus on the topic. Nothing stops us, for instance, from creating a demographically stratified corpus, or a corpus where frequently reprinted titles do count more than once. As long as researchers are clear about the criteria of selection they have used, these are all valid avenues of inquiry. We might even learn more this way than by debating abstract definitions of "representativeness." By posing the same question to corpora constructed differently, researchers can discover what difference criteria of selection make for specific questions. In some cases, selection criteria have turned out to make less difference than one might suppose – e.g., because the history of genre has broadly similar outlines in popular and less popular works (Underwood *et al*., 2013). In short, the questions about representativeness that are often presented as obstacles to text mining would be better construed as opportunities.

One obstacle that does still loom ominously on the horizon involves the question of whether or not mining texts and producing derivative data about those texts is in fact a violation of copyright. The simple fact of the matter is that text miners need digital texts to mine, and "modern copyright law," as Loyola law professor Matthew Sag (2012) puts it, "ensures that this process of scanning and digitization is ensnared in a host of thorny issues" (2). Because of a lack of clarity around what exactly constitutes fair use, many researchers have thus far limited themselves to the study of texts produced before 1923.[22] The legal troubles began in the USA in 2005, shortly after Google announced that it was scanning and digitizing the collections of a number of private and public academic libraries in order to make their collections searchable. The Authors Guild, an advocacy group that represents member authors, sued Google, claiming that these scanning efforts were a violation of copyright.[23] After more than eight years of back and forth litigation, the case was settled on November 14, 2013 with a summary judgment in which Judge Denny Chin ruled in favor of Google. In that judgment, Chin wrote that *Google Books*:

> advances the progress of the arts and sciences, while maintaining respectful consideration for the rights of authors and other creative individuals, and without adversely impacting the rights of copyright holders. It [*Google Books*] has become an invaluable research tool that permits students, teachers, librarians, and others to more efficiently identify and

> locate books. It has given scholars the ability, for the first time, to conduct full-text searches of tens of millions of books. (Authors Guild v. Google, p. 26)

Chin goes on to write specifically about the opportunities for humanities scholars interested in text mining. He cites an amicus brief that was submitted to the court on behalf of Digital Humanities and Law Scholars:

> in addition to being an important reference tool, Google Books greatly promotes a type of research referred to as "data mining" or "text mining." (Br. of Digital Humanities and Law Scholars as Amici Curiae at 1 (Doc. No. 1052)). Google Books permits humanities scholars to analyze massive amounts of data – the literary record created by a collection of tens of millions of books. Researchers can examine word frequencies, syntactic patterns, and thematic markers to consider how literary style has changed over time. (Authors Guild v. Google, p. 9–10)

Though this phase of the case was resolved with Judge Chin's ruling, at the time of writing the Authors Guild has vowed to appeal. If they make good on this promise, then the case will go before the Second Circuit and possibly then to the Supreme Court. Sag believes the possibility one of these higher courts would overrule Chin is very unlikely because, among other reasons, the Authors Guild failed to convince the judge that Google's scanning efforts have been in any way harmful to the copyright holders.

Regardless of what happens in the case between Google and the Authors Guild, humanities researchers can take some comfort in knowing that the HathiTrust Research Center has now received permission from the HathiTrust Board, with the approval of its host, the University of Michigan, and with the backing of Indiana University and the University of Illinois, to begin providing computational access to the copyrighted material held in HathiTrust's repository. This decision came after the HathiTrust won a similar lawsuit that was also brought by the Authors Guild. The ruling in that case made clear that the HathiTrust's use of books scanned as part of Google's book scanning project was fair use under US law.[24] At the time of this writing, HathiTrust has indicated that access will begin at the end of 2014 or early in 2015.

Despite these challenges, the future of text mining in the humanities is very promising. At no time in history have we ever had such access to the written record, and though that record is imperfect in many ways, it now marks a moment of great promise and great progress. And though this chapter may have overemphasized applications of text mining to literary studies (and unashamedly revealed the biases of the authors), text mining has already also been applied to a wide range of problems in other areas of the humanities. While many projects explore digital libraries, which contain mostly printed books, historians are actively working on newspapers, legal scholars on court cases, and scholars in media studies have done a great deal of text mining on social media. When scholars work with contemporary material, the boundary between "humanistic text mining" and "computational social science" can become porous. On the other hand, it is also possible for text mining to be applied to individual works, in service of interpretive projects that are quite distinct from social science. Below we offer a short list of exemplary projects.

Exemplary Projects and Examples of Text Mining in the Humanities

- Cameron Blevins. *Topic Modeling Martha Ballard's Diary* (http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary). In this blog post, Blevins uses topic modeling to better understand the 27-year diary of Martha Ballard, an American midwife born in 1735.
- Dan Cohen. *With Criminal Intent* (http://criminalintent.org). This project uses computational models to explore and visualize the history of crime as it is expressed in the court records of the Old Bailey.
- David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. *Gender identity and lexical variation in social media* (http://arxiv.org/abs/1210.4567). A study of the relationship between gender, linguistic style, and social networks, using a corpus of 14,000 users of Twitter.
- Jean-Baptiste Michel *et al*. *Culturomics* (http://www.culturomics.org). A project undertaken in conjunction with Google to study "culture" as it gets expressed in the Google Books Corpus.
- Matt Wilkens. *The Geographic Imagination of Civil War-Era American Fiction* (http://mattwilkens.com/2013/12/02/new-article-in-alh). Explores representations of place in a corpus of over 1000 novels by American authors published in the USA between 1851 and 1875.
- Robert K. Nelson. *Mining the* Dispatch (http://dsl.richmond.edu/dispatch). Uses topic modeling to explore the social and political life of Civil War Richmond as it is expressed in the pages of the *Daily Dispatch* from 1860 to 1865.
- Ryan Cordell, Elizabeth Maddock Dillon, and David Smith. *Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines* (http://www.viraltexts.org). This project explores the reuse of text in 19th century newspaper reportage in order to analyze the culture of reprinting in the United States before the Civil War.
- Sarah Allison, Ryan Heuser, Matthew Jockers, Franco Moretti, and Michael Witmore. *Quantitative Formalism* (http://litlab.stanford.edu/LiteraryLabPamphlet1.pdf). A report on a study designed to establish whether computer-generated algorithms could "recognize" literary genres in the nineteenth-century British novel.
- SEASR (http://www.seasr.org). A multi-year, Mellon-funded project that seeks to create a text-mining platform for scholarly research. A number of text-mining projects, including several by the authors of this chapter, have been completed in collaboration with SEASR.
- The HathiTrust Research Center (http://www.hathitrust.org/htrc). A virtual research center for large-scale, high-performance, and secure computation with the materials in the HathiTrust collection (as of 2014, about 4 billion pages of text in many languages, from many periods).

Notes

1 For example: Project Gutenberg, Google Books, HathiTrust.

2 Lenoir has made this argument on multiple occasions, primarily in lectures on pragmatic realism and social construction. He has written about this extensively in his book *Instituting Science* (1997), particularly the chapter on Haber-Bosch, in which he

discusses this idea at length. See also Hacking (1983).

3 See, for example, Biber (1998). Corpus linguistics is particularly valuable for identifying features that are over-represented in one group of sources relative to another: see Kilgarriff (2001).

4 For instance, practitioners have influentially claimed that data mining is properly understood as part of a "KDD process," but that's not necessarily how the phrase originated, or how it is used in the wild today (Fayyad *et al*., 1996).

5 See, for example, Witten *et al*. (2011:386–9). Note further that there is no consensus on what exactly constitutes "structured" versus "unstructured" data or text. In the linguistics community there tends to be a preference for thinking of text as highly structured, whereas in computer science and related fields, text is often considered *unstructured* and the term *structured* is typically reserved for discussions of databases and tables that impose a meta-structure onto the objects contained within that structure. This was the topic of a lively debate on the Corpora-List in December 2013 under the subject heading "Quotable Statistics on Unstructured Data on the WWW" (http://mailman.uib.no/public/corpora/2013-December/019362.html).

6 For a useful study of how feature set composition impacts attribution accuracy, see Grieve (2007).

7 This is not an arbitrary example. In *Representative Irish Tales*, W.B. Yeats identified two basic categories of Irish fiction characterized by what he called "the accent of the gentry and the less polished accent of the peasantry" (Yeats, 1979). Other scholars including Thomas MacDonagh (1916), Thomas Flanagan (1959), John Cronin (1980), and most recently Charles Fanning (2000) have all commented upon the distinct and specific use of language that appears to characterize Irish narrative and, moreover, the extent to which this use of language reflects, or does not, the unique position of Irish and Anglo-Irish writers in a country where the use of English evolved in a rather dramatic fashion. Though Mark Hawthorne has written that the "Irish were not accustomed to the English language and were unaware of its subtleties and detonations" (Hawthorne, 1975), Fanning and Cronin have separately argued that the Irish became masters of the English language and employed, in Fanning's words, a mode of "linguistic subversion" (Fanning, 2000).

8 This random sample was derived from the larger collection of 3500 novels that Jockers collected for his work in *Macroanalysis*. You can download the sample data and the necessary R code for repeating this experiment at http://www.wiley.com/go/schreibman/digitalhumanities. *British* here encompasses authors of the British Isles excluding the island of Ireland.

9 Features here include both words and marks of punctuation. Column headers for marks of punctuation in the sample data begin with a "p": e.g., the column for the *comma* is headed *pcomma*.

10 Euclidean distance is a fairly standard mathematical formula for calculating the "distance" between points in a multidimensional dataset. That said, readers wishing to employ distance metrics should be aware of potential problems with such measures. The so-called *curse of dimensionality* describes a situation in which the number of dimensions becomes so large that the data become sparse and all observations seem to be very dissimilar. In this example, we have reduced the feature space to 107 dimensions, and we have 300 observations.

11 It is difficult to avoid anthropomorphic representations of what the machine is doing in these examples. Obviously the machine doesn't actually "know" or "figure out" what objects are which. The machine merely calculates based on the rules that we set as programmers and then gives the appearance of "learning" something about the data. The analogy to human learning is useful, especially for a short chapter such as this, but readers are advised not to take the analogy too far.

12 Mathematically, of course, a circle doesn't really have any sides, at least if we define *side* as something unique to polygons. Some might argue that a circle has infinite sides. In either case, in this example, the oval is most like the circle.

13 There are various techniques for this type of testing, and *k*-fold cross validation is probably the most common, with *k* most frequently set to 10. Two other typical cross-validation tests include *repeated subsampling* and *leave-one-out* methods. A full description of each of these is beyond the scope of this chapter, but all of these methods share the common goal of training a model on a randomly generated subsample of the larger corpus and then testing the model's accuracy using the held-out data not selected in the randomization process.

14 NSC has been used effectively in several authorship attribution studies, including a benchmarking study (Jockers and Witten 2010) which demonstrated its efficacy in this type of problem. See also Tibshirani *et al*. (2002).

15 In a larger and more nuanced study of author nationality, Jockers discusses how nineteenth-century British authors tended to favor words indicative of absolutes and determinacy: words such as "always, should, never, sure, not, must, do, don't, no, always, nothing, certain, therefore, because, can, cannot, knew, know, last, once, only, right" are popular indicators of British prose. The classifier found that the Irish novels were best distinguished by words that can be thought of as characteristic of imprecision and indeterminacy, words such as "near, soon, some, most, still, less, more," and "much." Taken together, the former suggest confidence, whereas the later suggest uncertainty or caution. Interested readers should consult the chapter titled "Nation" in Jockers's *Macroanalysis* (2013).

16 For an introduction to the feature-selection problem see Yang and Pederson (1997). But note also that for some contemporary algorithms (e.g., support vector machines) feature selection becomes less critical than it was in 1997.

17 Topic modeling, which has become very popular in recent years, is an excellent example of a powerful technique that relies entirely upon a bag-of-words representation of text.

18 See, for example, http://nlp.stanford.edu/software/corenlp.shtml and/or http://nltk.org/

19 For an example of this process, see Bamman *et al*. (2013).

20 For normalization of early-modern spelling, see Baron (2013). For later texts see Underwood (2013).

21 TEI is an XML standard developed by the Text Encoding Initiative. See http://www.tei-c.org/index.xml.

22 Copyright has expired for all works published in the United States before 1923.

23 An excellent overview of the case can be found on Matthew Sag's blog (http://matthewsag.com).

24 As with the Google case, the Authors Guild is appealing this decision to the Second Circuit.

References and Further Reading

Authors Guild v. Google Inc., 770 F. Supp. 2d 666 – Dist. Court, SD New York 2011.

Bamman, D., O'Connor, B., and Smith, N.A. 2013. Learning latent personas of film characters. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 352–61.

Baron, A. 2013. *Variant Detector: (VARD2)*. http://ucrel.lancs.ac.uk/vard/about (accessed June 20, 2015).

Bekkerman, R., and Allan, J. 2003. Using Bigrams in Text Categorization. CIIR Technical Report. http://people.cs.umass.edu/~ronb/papers/bigrams.pdf (accessed June 20, 2015).

Biber, D. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bird, S., Klein, E., and Loper, E. 2009. *Natural Language Processing with Python*. O'Reilly Media. http://nltk.org/book (accessed June 20, 2015).

Clement, T. 2008. "A thing not beginning or ending": using digital tools to distant-read Gertrude Stein's *The Making of Americans*. *Literary and Linguistic Computing* 23 (3), 361–82.

Cronin, J. 1980. *The Anglo-Irish Novel*. Totowa, NJ: Barnes & Noble.

Fanning, C. 2000. *The Irish Voice in America: 250 Years of Irish-American Fiction*, 2nd edition. Lexington: University Press of Kentucky.

Fayyad, U., Piatetsky-Shapiro, G., and Smythe, P. 1996. From data mining to knowledge discovery in databases. *AI Magazine* 17, 37–54.

Flanagan, T. 1959. *The Irish Novelists, 1800–1850*. New York: Columbia University Press.

Grieve, J. 2007. Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing* 22 (3), 251–70.

Hacking, I. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

Hall, M., Frank, E., Holmes, G., *et al.* 2009. The WEKA data mining software: an update. *SIGKDD Explorations* 11 (1).

Han, J., Kamber, M., and Pei, J. 2012. *Data Mining: Concepts and Techniques*. Burlington, MA: Morgan Kaufmann.

Hawthorne, M.D. 1975. *John and Michael Banim (the "O'Hara Brothers"): A Study in the Early Development of the Anglo-Irish Novel*. Salzburg Studies in Romantic Reassessment, vol. 50. Salzburg: Institut für Englische Sprache und Literatur, Universität Salzburg.

James, G., Witten, D., Hastie, T., and Tibshirani, R. 2013. *An Introduction to Statistical Learning: with Applications in R*. New York: Springer.

Jockers, M. 2014. *Text Analysis with R for Students of Literature*. New York: Springer.

Jockers, M. Text-mining. http://www.matthewjockers.net/category/tm (accessed June 20, 2015).

Jockers, M.L. and Witten, D.M. 2010. A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing* 25 (2), 215–24.

Jockers, M.L. 2013. *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.

Jones, K.S. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28 (1), 11–21.

Kilgarriff, A. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6 (1), 97–133.

Lenoir, T. 1997. *Instituting Science: The Cultural Production of Scientific Disciplines*. Writing Science. Stanford, CA: Stanford University Press.

Macdonagh, T. 1916. *Literature in Ireland: Studies Irish and Anglo-Irish*. London: T.F. Unwin.

Manning, C.D., Raghavan, P., and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.

McCallum, A.K. 2002. MALLET: a Machine Learning for Language Toolkit. http://mallet.cs.umass.edu (accessed June 20, 2015).

Mendenhall, T.C. 1887. The characteristic curves of composition. *Science* n.s. 9 (214), 237–46.

Mendenhall, T.C. 1901. A mechanical solution of a literary problem. *Popular Science Monthly* 60, 97–105.

Muralidharan, A. Text mining and the digital humanities. http://mininghumanities.com (accessed June 20, 2015).

Olsen, M. 1993. Signs, symbols, and discourses: a new direction for computer-aided literature studies. *Computers and the Humanities* 27 (5–6), 309–14.

Potter, R. 1988. Literary criticism and literary computing. *Computers in the Humanities* 22 (2), 93.

Ramsay, S. 2007. Algorithmic criticism. In *A Companion to Digital Literary Studies*, ed. R.G. Siemens and S. Schreibman. Oxford: Blackwell.

Rosen, J. 2011. Combining close and distant, or the utility of genre analysis: a response to Matthew Wilkens's "Contemporary Fiction by the Numbers". *Post* 45. http://post45.research.yale.edu/2011/12/combining-close-and-distant-or-the-utility-of-genre-analysis-a-response-to-matthew-wilkenss-contemporary-fiction-by-the-numbers (accessed December 3, 2011).

Sag, M. 2012. Orphan works as grist for the data mill. *Berkeley Technology Law Journal* 27 (4).

Shaw, R. 2012. Text-mining as a research tool in the humanities and social sciences. http://aeshin.org/textmining (accessed June 20, 2015).

Sherman, L.A. 1893. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry*. Boston, MA: Ginn.

Tibshirani, R., Hastie, T., Narasimham, B., and Chu, G. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Science* 99 (10), 6567–72.

Tong, X., and Evans, D.A. 1996. A statistical approach to automatic OCR error correction in context. In *Proceedings of the Fourth Workshop on Very Large Corpora*, 88–100.

Tsoumakas, G., and Katakis, I. 2007. Multilabel classification: an overview. *International Journal of Data Warehousing & Mining* 3 (3), 1–13.

Underwood, T. 2013. A half-decent OCR normalizer for English texts after 1700. *The Stone and the Shell*. http://tedunderwood.com/2013/12/10/a-half-decent-ocr-normalizer-for-english-texts-after-1700 (accessed June 20, 2015).

Underwood, T. 2015. Where to start with text mining. http://tedunderwood.com/2012/08/14/where-to-start-with-text-mining (accessed June 20, 2015).

Underwood, T., Black, M.L., Auvil, L., and Capitanu, B. 2013. Mapping mutable genres in structurally complex volumes. *arXiv preprint:1309.3323*. http://arxiv.org/abs/1309.3323 (accessed June 20, 2015).

Unsworth, J. 2013. Digital humanities: from 1851? Brandeis University Library and Technology Services. http://blogs.brandeis.edu/lts/2013/05/17/digital-humanities-from-1851 (accessed June 20, 2015).

Witten, I.H., Frank, E., and Hall, M.A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann.

Yang, Y. and Pederson, J.O. 1997. A comparative study on feature selection in text categorization. In *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*.

Yeats, W.B. 1979. *Representative Irish Tales*. Atlantic Highlands, NJ: Humanities Press.