DHQ: Digital Humanities Quarterly
2009
Volume 3 Number 2

Words, Patterns and Documents: Experiments in Machine Learning and Text Analysis

Shlomo Argamon <argamon_at_iit_dot_edu>, Linguistic Cognition Lab, Dept. of Computer Science, Illinois Institute of Technology
Mark Olsen <markymaypo57_at_gmail_dot_com>, ARTFL Project, University of Chicago

Abstract

This introduces the set of papers reflecting initial collaborative work between the ARTFL Project at the University of Chicago and the Linguistic Cognition Laboratory at the Illinois Institute of Technology on the intersection of machine learning, text mining and text analysis.

One of the emerging grand challenges for digital humanities in the next decade is to address rapidly expanding repositories of electronic text. A number of efforts, such as Google Book Search and the Bibliothèque numérique européenne, are digitizing the holdings of many of the world's great research libraries. The resulting collections will contain nothing less than, in Gregory Crane's view, the “stored record of humanity” [Crane 2006]. This expansion beyond existing digital collections will be one of at least a couple of orders of magnitude, and will introduce a variety of new problems beyond simply scale, including heterogeneity of content and granularity of objects. The problems posed by the emerging global digital library offer opportunities for collaborative work between scholars in the humanities and computer scientists in many domains, from optical character and page recognition to historical ontologies [Argamon and Olsen 2006]. The papers presented here reflect initial collaborative work between the ARTFL Project at the University of Chicago and the Linguistic Cognition Laboratory at the Illinois Institute of Technology on one subset of the technologies required for a future global digital library: the intersection of machine learning, text mining and text analysis. 1

Traditional models of text analysis in digital humanities have concentrated on searching for a relatively small number of words and reporting results in formats long familiar to humanities scholars, most notably concordances, collocation tables, and word frequency breakdowns.[1] While effective for many types of questions, this approach will not scale effectively beyond collections of a relatively modest size, as result sets for even uncommon groups of words will balloon to a size not readily digestible by humans. Furthermore, this approach does not lend itself to abstract discussions of entire works, the oeuvre of an author or period, or issues related to the language of gender, genre, or ethnicity. It places the onus on the user to construct queries and assimilate results, without leveraging the capacity of machines to identify patterns in massive amount of data. 2

Machine learning and text mining approaches appear to offer a compelling complement to traditional text analysis, by having the computer sift through massive amounts of text looking for “suggestive patterns.” The power of modern machine learning systems to uncover patterns in large amounts of data has led to their widespread use in many applications, from spam filters to analyzing genetic sequences. And the potential for using these sophisticated algorithms to find meaningful patterns in humanistic texts has been recently observed. Drawing a link between Ian Witten's general description of data mining and the practice of literary criticism, Stephen Ramsay states that “[f]inding interesting patterns and regularities in data is generally held to be of the deepest significance.” Any such findings must be approached with critical prudence, he warns, as they will contain “the spurious, the contingent, the inexact, the imperfect, and the accidental in a state of almost guaranteed incompleteness” [Ramsay 2005, 186]. Ramsay is quite correct to point out both the potential power and pitfalls of applying text mining to questions in the humanities. Our current work is to design sets of relatively constrained experiments using text mining systems on specific problems in 3 order to examine what works, what does not work, and just what such results might mean.

To this end, the ARTFL Project has developed a set of machine learning extensions to PhiloLogic, our full-text search and analysis system.[2] PhiloMine replaces the notion of “searching” a database for one or more words with “task” submission. We currently view three broad classes of “tasks”: predictive text mining, comparative text mining, and clustering/similarity analysis. Predictive mining approaches are widely used in applications such as spam e-mail filters, which are trained on samples of spam and non-spam messages and used to identify incoming junk mail. This supervised learning technique can be applied to a wide variety of tasks, such as learning on topically classified articles in the *Encyclopédie* and assigning these classes to unclassified articles or parts of other documents. It is common in digital humanities to work with corpora where many classifications, such as gender or nationality of author, are already known. In this case, machine learning algorithms may be used to compare texts based on different attributes. For example, one may compare works by American and non-American Black playwrights, returning measures of how well the classification task was performed, identifying incorrectly classified documents, and the features (often words) most characteristic of the distinction. Finally, document similarity and clustering is an unsupervised form of machine learning, designed to identify groups of documents statistically that share common features. We are using nearest neighbor document similarity, for example, to identify passages in one text that may have been copied from an earlier document. 4

The three papers which follow use all three approaches to attempt to shed light on specific research questions in the humanities. “Gender, Race, and Nationality...” examines how well machine learning tools can isolate stylistic or content features of authors and characters by gender, race, and nationality in a large collection of works by Black playwrights. In general, the classification results on a range of mining tasks were quite good, suggesting that these techniques can effectively distinguish, for example, the writing of male and female or American and non-American authors. In some cases, the results provide insight into the texts as literary works, but in others we found the intellectual value of the feature sets to be less interesting. We also found that, while classifying texts under binary oppositions is generally effective for the machine learning algorithms employed, doing so tends to reduce complex works and corpora to very limited sets of common features. 5

In “Vive la différence...”, we examine a single binary classification, on gender of author in French literature predominantly from the 17th to the early 20th centuries. Using balanced male and female corpora, we found substantial agreement with Olsen's previous studies of gendered writing in published works, with our results supporting his observation of a more personal and emotional sphere of female authorship. Our results also comport with Argamon's previous work (with Koppel) on the British National Corpus, where female writing was found to be characterized by more frequent use of personal pronouns, with male writing characterized by more frequent use of determiners and numerical quantifiers. Additionally, a number of strong thematic groups of content words were found for both genders that were consistently useful in classification across the time period represented in the corpus, suggesting some enduring differences between male and female writing in the corpus. 6

The third paper, “Mining Eighteenth Century Ontologies...”, uses predictive classification to examine the ontology of Diderot and d'Alembert's *Encyclopédie*. Our initial experiments attempting to classify the unclassified articles of the Encyclopédie led us to reconsider the coherency of the editors' classification scheme and overall distribution of classes in the entire work. Lastly, applying this ontology of the classes of knowledge to the *Journal de Trévoux*, an 18th century scholarly journal, we were able to make several new connections between the two corpora that went previously unnoticed. 7

The power of machine learning and text mining applications to detect patterns is clearly demonstrated in these papers, yet several issues arose during this work which we believe should be raised. The first is the surprisingly small size of the patterns detected. In all of the experiments, the systems dutifully created models to fit classes, but these were often based on quite tiny fractions of all of the available features -- a mere 60 surface words can adequately distinguish hundreds of American and non-American plays by black authors. Similarly, we find that for both predictive classification and clustering tasks, the number of features for most tasks used is a tiny fraction of all possible features. Resulting features may well reflect a “lowest common denominator” which, while perfectly adequate for specific mining tasks, may not be as useful in characterizing works in an intellectually satisfying fashion. The fact that our studies examining the 8 issue of gendered writing arrived at similar conclusions regarding the differences between male and female writing and characterizations may thus in part be an artifact of the way learners and classifiers function. Finally, our classification tasks are generally considered to have produced a significant result when we achieve an accuracy of 70% or more, although the most successful tasks can surpass 95%. When examining the features most useful to the model, we must not assume that their importance holds for the documents whose class could not be predicted; indeed, their incorrect classification suggests that these documents may have quite different patterns of word usage.

9 The “lowest common denominator” problem would also appear to be related to a second concern which may be specific to machine learning on humanities texts. By treating relatively small numbers of documents with very large numbers of possible features, classifiers are thus given a wide range of features to accomplish any particular task. While we used different techniques to validate results, including n-fold cross validation and random falsification, there would appear to be some danger of obtaining results based on the construction of the task itself. Even if significant results are found, showing, e.g. that classification by a particular binary opposition can be performed reliably at 80% accuracy, in itself this says little about the underlying phenomenon under investigation. A binary opposition that is thus “empirically supported” may well be an epiphenomenon that is merely correlated with another underlying complex of causes, which remain to be teased out. So finding such “statistical patterns” is, ultimately, merely the first step in what must be a critically well-grounded argument, supported also by evidence external to the classification results themselves.

10 To help us argue for the general efficacy of machine learning approaches and address the concerns set forth above, we include a reaction piece by Sean Meehan, who writes about the anxieties of doing criticism by algorithm. Meehan raises the issue of distance in any critical endeavor, pointing out that interpretive analysis is always “a dynamic between tools and texts.” In the end, he sounds the theme of scholarly circumspection and care that we try to bring out in all of the articles. Using machine learning tools on humanities texts requires the same understanding of the texts and degree of self-awareness that are necessary for any literary critical study.

11 As we hope these small scale experiments have demonstrated, text mining and machine learning algorithms offer novel ways to approach problems of text analysis and interpretation. One can pose questions of many hundreds or thousands of documents and obtain results that are interesting and sometimes even striking. It further seems clear that text mining will be a powerful technology deployed in order to make the emerging global digital library manageable and meaningful.

Notes

[1] The PhiloLogic text search and analysis package, developed at ARTFL, is one of many examples of such traditionally oriented systems. Documentation, downloads and samples are available at http://philologic.uchicago.edu/.

[2] See http://philologic.uchicago.edu/philomine/ for samples, documentation, and downloads. Not all machine learning and text mining tasks in the following papers used Philomine.

Works Cited

Argamon and Olsen 2006 ArgamonShlomo, and Olsen Mark. “Toward meaningful computing”. *Communications of the Association for Computing Machinery* 49: 4 (2006), pp. 33-35.

Crane 2006 Crane, Gregory. “What Do You Do with A Million Books?”. *D-Lib Magazine* 12: 3 (2006). http://www.dlib.org/dlib/march06/crane/03crane.html.

Ramsay 2005 Ramsay, Stephen. “In Praise of Pattern”. *TEXT Technology* 14: 2 (2005), pp. 177-190.