DHQ: Digital Humanities Quarterly
2009
Volume 3 Number 2

Mining Eighteenth Century Ontologies: Machine Learning and Knowledge Classification in the *Encyclopédie*

Russell Horton <russ_at_diderot_dot_uchicago_dot_edu>, Digital Library Development Center, University of Chicago
Robert Morrissey <rmorriss_at_uchicago_dot_edu>, University of Chicago
Mark Olsen <markymaypo57_at_gmail_dot_com>, ARTFL Project, University of Chicago
Glenn Roe <glenn_at_diderot_dot_uchicago_dot_edu>, ARTFL Project, University of Chicago
Robert Voyer <rlvoyer_at_diderot_dot_uchicago_dot_edu>, Powerset

Abstract

The *Encyclopédie* of Denis Diderot and Jean le Rond d'Alembert was one of the most important and revolutionary intellectual products of the French Enlightenment. Mobilizing many of the great – and the not-so-great – *philosophes* of the 18th century, the *Encyclopédie* was a massive reference work for the arts and sciences, which sought to organize and transmit the totality of human knowledge while at the same time serving as a vehicle for critical thinking. In its digital form, it is a highly structured corpus; some 55,000 of its 77,000 articles were labeled with classes of knowledge by the editors making it a perfect sandbox for experiments with supervised learning algorithms. In this study, we train a Naive Bayesian classifier on the labeled articles and use this model to determine class membership for the remaining articles. This model is then used to make binary comparisons between labeled texts from different classes in an effort to extract the most important features in terms of class distinction. Re-applying the model onto the original classified articles leads us to question our previous assumptions about the consistency and coherency of the ontology developed by the Encyclopedists. Finally, by applying this model to another corpus from 18th century France, the *Journal de Trévoux, or Mémoires pour l'Histoire des Sciences & des Beaux-Arts*, new light is shed on the domain of Literature as it was understood and defined by 18th century writers.

Introduction

One of the crowning achievements of the 18th-century Enlightenment was the *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une société de gens de lettres*, edited by Diderot and d'Alembert. Published in Paris between 1751 and 1772, in 17 volumes of text and 11 volumes of plates, it contains some 77,000 articles written by more than 140 contributors. The *Encyclopédie* was a massive reference work for the arts and sciences, as well as a *machine de guerre* that served to propagate Enlightenment ideas. The impact of the *Encyclopédie* was enormous. Through its attempt to classify learning and to open all domains of human activity to its readers, the *Encyclopédie* gave expression to many of the most important intellectual and social developments of its time.[1] 1

The scale and ambition of the *Encyclopédie* inspired its editors to adopt three distinct modes of organization which, taken together, Diderot described as encyclopedic: dictionary, hierarchical classification, and the *renvois* (cross-references). The interaction of these three modes has led modern commentators to describe the *Encyclopédie* as an “ancestor of hypertext” and to depict Diderot as “ *l'internaute d'hier* ” (“websurfer *avant la lettre* ”) [Brian 1998]. D'Alembert underscores the importance of the organization of knowledge in the *Discours Préliminaire*: 2

> As an *Encyclopedia*, it is to set forth the *order* and *connection* of the parts of human knowledge. As a *Reasoned Dictionary of the Sciences, Arts, and Trades*, it is to contain the general principles that form the basis of each science and each art...and the most essential facts that make up the body and substance of each.[2]

Of the three modes of organization, the dictionary mode (organization of entries in alphabetical order) is certainly the simplest and the most arbitrary. The second mode of organization is classification, wherein each dictionary entry is assigned to a “class of knowledge,” placing it within the “order” of human understanding, as depicted in the *Système Figuré des connaissances humaines*. Modeled after Bacon's classification of knowledge and Enlightenment theories of epistemology, all understanding is founded upon memory, reason, or imagination, with numerous categories and sub-categories branching out from these three faculties.[3] However, simply placing an entry into this hierarchy of knowledge was insufficient to indicate the interconnections of knowledge. Thus, Diderot created an extensive system of *renvois* , or cross-references, the third mode of organization, providing a lattice of interconnections between individual leaves of the tree as well as between classes of knowledge.[4]

3 The central role of the classification system in the intellectual objectives of the *Encyclopédie* editors is indicated by the extent to which it has been discussed and debated by both contemporary scholars and later researchers. The editors were remarkably diligent in assigning classes of knowledge to each article and sub-article. Of the 77,085 main and sub articles, we have identified 55,248 as having classes of knowledge.[5] The editors were, however, somewhat less diligent in maintaining a precisely controlled list. Thus the classifications as found in the text are an amalgam of abbreviations, conflations, and singular categories that are not found on the *Système Figuré* at all. We have recently completed orthographic normalization of the classes of knowledge assigned to each article,[6] resulting in some 54,289 articles with 2,620 normalized classes of knowledge. The twenty most frequent classifications by number of articles are:

Géographie	5513
Géographie moderne	4794
Géographie ancienne	3084
Jurisprudence	2396
Grammaire	2304
Marine	1894
Commerce	1483
Histoire naturelle. Botanique	1277
Histoire moderne	1194
Mythologie	1115
Histoire naturelle	1069
Histoire ancienne	889
Medecine	796
Architecture	730
Jardinage	689
Littérature	682
Maréchallerie	627
Botanique	614
Histoire ecclésiastique	614
Théologie	517

Table 1. Original counts of top 20 most frequent classifications by number of articles

Like the *Système Figuré*, these classifications are a reflection of how knowledge was ordered and classified in the 18th century. This paper reports the results of experiments using machine learning and data mining techniques to understand and exploit this unique resource.

Preliminary Objectives

Our initial experimental objectives were threefold: 4

1. To train our classifier on all of the normalized classified articles of the *Encyclopédie* and then to apply the resulting model to the 22,796 unclassified articles, labeling each with a predicted class of knowledge.
2. Using the same model, we then sought to reclassify all of the classified articles, verifying the relative success of our classifier by examining the misclassified instances and then exploring the relationships of discrete classes of knowledge to one another through feature set analysis.
3. Finally, we wanted to apply the classification model generated from the *Encyclopédie* to other 18th century texts, allowing us to evaluate the applicability of this ontology to sources outside of the encyclopedic enterprise.

Methodology

Before we could begin any classification tasks we needed to preprocess the data into a usable format. To this end, we extracted the text from all articles and sub-articles in the *Encyclopédie* with a normalized class of knowledge. Explicit markers of class membership, present at the beginnings of these articles (e.g., *Comm.* for *Commerce*, *Phys.* for *Physics*, etc.) were removed to ensure that they would not provide facile criteria for classification. The texts were then tokenized and lemmatized[7] automatically, and frequencies of words and lemmas were computed both globally and for each article. Words and lemmas were then used as attributes, and vectors for each article were generated from the number of occurrences of each attribute in that article. 5

Text categorization is an extensively studied subfield of information science and accordingly, there exist many time-tested classification algorithms, each with their own strengths and weaknesses. We chose to implement a Multinomial Naive Bayesian classifier because of its simplicity and efficacy on large corpora.[8] Multinomial Naive Bayes (MNB) treats documents as bags of words, where word order is considered irrelevant. Essentially, we measure the likelihood of words occurring in a given class by calculating how many times they occur in all documents with known classes. These conditional probabilities allow us to guess the most “probable” class of an unclassified article based on the frequency of the words that occur within it. MNB assumes that the probability of a word occurring within a document is independent of the words that occur around it, an assumption that we know to be false. Yet, despite this, MNB is known to perform very well on large data sets and in fact, has been shown to produce more accurate results than many other classification algorithms.[9] 6

We designed a test bed using this implementation that allowed for easy modification of several parameters including the minimum word count required for an article to be included in our model, as well as the minimum and maximum number of occurrences across the entire corpus required for a word or lemma to be included in our vocabulary. Massaging the data in this way, we can avoid, as much as possible, skewed results caused by high frequency function words or highly weighted words occurring in a relatively small number of articles. After several rounds of experimentation we found that our classifier was most successful when excluding articles of 25 words or fewer and words (lemmas) occurring in less than 4 articles.[10] 7

Results

Classifying the Unclassified

Our first classification task was to train the classifier on the 54,289 articles that were assigned categories of knowledge by the editors of the *Encyclopédie* and to then apply this model to the 22,796 unclassified articles in an attempt to predict class membership for the articles in question. Once classified, the twenty most frequent labels for the newly 8 classified articles were:

Jurisprudence	4276
Art méchanique	1260
Géographie	828
Commerce	802
Anatomie	643
Marine	557
Histoire moderne	475
Architecture	435
Grammaire	402
Histoire naturelle. Ornithologie	367
Medecine	363
Géographie moderne	347
Art militaire	311
Histoire ecclésiastique	308
Géométrie	306
Géographie ancienne	306
Musique	295
Astronomie	274
Histoire naturelle. Botanique	266
Théologie	215

Table 2. Counts of top 20 most frequent classifications for previously unclassified articles.

While this distribution of classes bears some resemblance to the overall distribution cited above, we have no real way of verifying the accuracy of the classifier given the unknown content/classes of the unclassified articles.[11] While it is entirely plausible that 19% of the unclassified articles are concerned with Jurisprudence in a general sense, it is also likely that Jurisprudence (which represents only 5% of the classified articles) becomes of sort of “catch-all” category for the classifier into which articles concerned with any specific aspect of law (i.e. *droit romain*, *droit canonique*, *droit civil*, etc.) are grouped.

9 The sample of results we examined reveal that the classifier performed reasonably well. By this, we mean that some classifications seemed right on; some made a good degree of sense, while others were perhaps a bit too general, failing to accurately represent the specificity of the subject matter. Naturally, the more than 22,000 newly generated classifications could not all be verified by hand, and so we focused on major articles and a selection of smaller ones. We were particularly encouraged by the assigned classifications for the 10 longest unclassified articles. The *Discours Préliminaire*, d'Alembert's famous preface to the *Encyclopédie* detailing the intellectual underpinnings of the enterprise, comes back as belonging to the class Philosophy.[12] Going down the list we see that the article “Anatomie” is assigned its own classification in Anatomy and “Chimie” is rightly placed into Chemistry, results we had originally hoped were easy enough for the algorithm to attain. Most of the classifications, however, don't fall into such clear categories. For example, “Venerie” — the art of hunting — was assigned to Natural History; the philosophical article “Eclectisme” to the History of Philosophy, etc. Indeed, while these and the better part of the predicted classifications can be justified on a

general level, we had to concede that the overall utility of this task was somewhat questionable. Quite simply, as we mention above, there was just too much data to sort through. The new labels were often interesting, but we were not able to study them easily or thoroughly enough to come to any deeper understanding about how the *philosophes* structured knowledge or indeed how the *Encyclopédie* itself fits together. Thus, trying to develop an experiment that could generate more legible results, we decided instead to leverage the information given us by the editors in exploring the known classifications and their relationship to each other and then later, to consider the classification scheme as a whole by examining the general distribution of classes over the entire work as opposed to individual instances.

Classified vs. Classified — Feature set evaluation

10 Having run a set of predictive classification experiments on the unclassified articles, our next task was an attempt at what we have named "comparative" classification, wherein we train the classifier on two particular classes, and then reclassify them in an effort to determine how separable they are and to ascertain which features best distinguish articles from the two sets. The goal here has less to do with the accuracy of the classifications than with the feature sets that are generated during the classification task. Any two classes can be compared. Feature weights are generated using the Naive Bayes Perl module written by Ken Williams. These weights represent the conditional probability of a feature f given a class c and are generated based on their frequency in the known articles.[13] To give an example of this sort of comparative classification, 889 articles classified as “Histoire ancienne” were compared with 1194 articles in the “Histoire moderne” class and the following features and their weights were extracted as the most relevant in determining class membership:

étoient	0.04356
avoit	0.03705
romains	0.02472
avoyer	0.02455
an	0.02341
peuple	0.02271
chez	0.02188
sous	0.02182
tems	0.02170
mot	0.02146
empereur	0.02146
g	0.02122
appelloit	0.01880
premier	0.01847
nous	0.01802
encore	0.01783
après	0.01763
homme	0.01756
dieu	0.01683
rome	0.01683

Table 3. Conditional probabilities of top 20 most common words in articles from *Histoire ancienne*

roi	0.03744
ordre	0.02870
prince	0.02390
sous	0.02314
nommer	0.02180
titre	0.02165
empire	0.02065
chevalier	0.02046
officier	0.02039
tems	0.02030
étoient	0.02003
premier	0.01959
empereur	0.01903
porter	0.01891
état	0.01871
mot	0.01851
avoit	0.01823
prendre	0.01814
maître	0.01814
sans	0.01780

Table 4. Conditional probabilities of top 20 most common words in articles from *Histoire moderne*

The overall performance of the classifier came in at 95.63%, which tells us that while both belong to the same "branch" of science, namely History, the ancient and the modern are nonetheless significantly distinguishable from one another. When considering the two lists of features, one immediately notices that the results make good sense, i.e., we find more verbs in the past tense (*avoient* , *étoient*, etc.) in the Ancient History articles as well as references to antiquity (*romains* , *empereur*, *rome*, etc.). The single-letter feature "g" is the signature of the Abbé Mallet who was the author of a significant portion of the articles on Ancient History. Interestingly, some of the features occur in both lists, which is more than likely a result of the close relationship and dialogue between the two classes. In the feature set for Modern History, we find references to forms of government that quite rightly belong to the modern period (*état*, *prince*, *roi*, *chevalier*, etc.) and the notable absence of the word " *dieu* " ("god").

11 We can also run comparative classification tasks on seemingly dissimilar classes of knowledge such as the 682 Literature articles and the 200 articles dealing with Physics. Not surprisingly, we obtain a very high rate of success for this sort of classification, in this case, 99.29%:

nous	0.03330
mot	0.02902
avoit	0.02607
étoient	0.02294
livre	0.02283
ancien	0.02149
je	0.02117
tems	0.02030
bien	0.01997
encore	0.01963
sans	0.01956
vers	0.01907
dieu	0.01835
auteur	0.01820
latin	0.01785
usage	0.01742
devoir	0.01738
notre	0.01695
homme	0.01691
chose	0.01671

Table 5. Conditional probabilities of top 20 most common words in articles from *Littérature*.

corps	0.05394
air	0.04216
eau	0.04185
nous	0.03507
froid	0.02597
chaleur	0.02584
monsieur	0.02580
degré	0.02296
moins	0.02251
fort	0.02242
glace	0.02196
couleur	0.02169
feu	0.02160
lorsque	0.02109
effet	0.02048
peu	0.01991
rayon	0.01986
fluide	0.01962
mouvement	0.01952
trouver	0.01948

Table 6. Conditional probabilities of top 20 most common words in articles from *Physique*

12 The feature scores from this model are what we would expect. The Literature class' most significant features are those words pertaining to language and grammar (*mot*, *livre*, *vers*, *auteur*, *latin*, *usage*, etc.) whereas the Physics class is dominated by materialistic terminology (*corps*, *air*, *eau*, *degré*, *fluide*, *mouvement*, etc.) consistent with the scientific writings of the period.

13 Evaluation of these feature sets can be invaluable when testing certain hypotheses, such as word usage differences across similar disciplines or between authors. In this particular case, the features provide an intuitive illustration of the differences between these two distinct classes of knowledge. While features are simply terms that the algorithm finds statistically representative of a particular class, the feature sets as a whole can also give a snapshot of the make-up of the individual classes or indeed of larger concepts more generally. From the list for literature, for example, we get a sense of the importance that classical Roman authors still had for the *philosophes* with the terms “ancien” (“ancient”) and *latin.* “Vers” (“verse”) perhaps reflects the fact that versification was a predominant aspect of literary style, whether in poetic, dramatic, and other writing at that time. Terms such as “mot” (“word”) and “usage” (“use”) might point to the 18th century's expansive, belle-lettristic sense of literature that we discuss below. In a more general manner, feature sets provide us with an expanded thesaurus for any given classification task -- leads for further investigation and study -- that can then be exploited by a more traditional full text analysis system.[14]

Reclassifying the Classified — the Ontology of the Encyclopédie

14 Finally, we applied the model assembled for our first experiment — trained on all of the known classifications — onto all

of the already classified articles. By this, we mean that we effectively ignored any given classes of knowledge, treating each article as if it were unclassified, and then assigned class membership using the algorithm described above. Here our goal in the results analysis was twofold: first, we were curious as to the overall performance of our classification algorithm, i.e., how well it correctly labeled the known articles; and secondly, we wanted to use these new classifications to examine the outliers or misclassified articles in an attempt to understand better the presumed coherency and consistency of the editors' original classification scheme.

We achieved a 71.4% success rate in the re-categorization of the 54,289 classified articles, a performance that could perhaps be improved with a more accurate morphological stemmer and the inclusion of n-grams as features, fucnctions we intend to implement in the future. Nonetheless, developing a model to reliably guess an article's given class of knowledge is ultimately not our primary concern as even a perfect model, while impressive in terms of performance, could only yield that which we already know, namely the assigned classes of knowledge. The sheer size and complexity of the *Encyclopédie*, drawing its contents from hundreds of distinct writers, all but guarantees a lower rate of performance for any classification algorithm. This fact need not be discouraging however, as we are more interested in exploring the use of these text mining techniques as knowledge discovery tools, uncovering previously unnoticed connections and classifications, such as the particular use of the class “Literature” outlined below, rather than simply using these approaches as a statistical platform for hypothesis testing. 15

The twenty most frequent classes after re-classification: 16

Géographie	3926
Géographie ancienne	3492
Géographie moderne	3273
Jurisprudence	2552
Commerce	2104
Art méchanique	1662
Histoire naturelle. Botanique	1615
Marine	1575
Histoire moderne	1514
Mythologie	1334
Architecture	1213
Grammaire	1111
Histoire ancienne	1061
Histoire ecclésiastique	781
Medecine	746
Histoire naturelle	727
Littérature	646
Maréchallerie	592
Morale	573
Jardinage	566

Table 7. The 20 most frequent classes of knowledge by number of articles after re-classification.

17 When comparing the results to the original classifications we note that the class “Grammar” falls out of the top ten while

“Art méchanique,” which is not included in the original top twenty, ranks as the sixth most frequent class. The Grammar class is known to be problematic as Diderot frequently used this seemingly innocuous label to hide more polemical entries.[15] As for the “Art méchanique” category, we suspect that many of the overly specific classes dealing with the mechanical arts were subsumed into this larger, more inclusive set. By and large the rest of the classes are consistent with the overall distribution in the *Encyclopédie* although the rankings differ slightly.

The most interesting results here come from the examination of misclassified articles, which belie vocabularies that do not belong probabilistically to their assigned categories. Upon analyzing a random subset of the misclassified articles, we identified three distinct types of misclassifications. First, there are articles whose original classification was too infrequent; for example, the article “Accrues” (metal rings used to knit together fishing net) is the sole member of the class “Marchands de Filets” (net merchants) and was placed into the more general class of “Pêche” (fishing). There are also articles whose vocabularies mislead the classifier. One such case is the article “Achées” (a type of worm used in bait-fishing), originally classified as “Pêche,” it was later assigned to the class “Jardinage” (gardening). The article is in fact less a description of anything to do with fishing, but rather contains instructions on how to find and cultivate bait worms in a garden. Finally, there are entries whose predicted class, while incorrect, seems more logical than the original. The article “Tepidarium,” which describes an ancient Roman bathhouse, would appear to have more in common with its predicted class, Architecture, than the one assigned by the editors, namely Literature. Certainly our judgment that the predicted class is more appropriate than the original class of knowledge is biased by our modern epistemological paradigm, but this does not necessarily mean that the original system of classification was entirely consistent and coherent. Naturally then, applying our model onto other 18th century French texts should provide further insight into the power of the classifier and more importantly, into the ontology originally laid out by the *philosophes*. 18

Classification outside of the Encyclopédie

The *Journal de Trévoux*, or *Mémoires pour l'Histoire des Sciences & des Beaux-Arts*, was one of the most influential 18th century French periodicals. A sort of literary/scholarly journal reviewing and commenting on a wide variety of contemporary publications, the *Journal de Trévoux* dealt with almost every discipline of knowledge. Given the great variety of subject matter contained in this collection we felt it would be a natural choice for us to begin studying the relationship of the Encyclopédie ontology to other 18th century texts. Of course, the 18th century in France was a time of intellectual ferment and, as in most historical moments, there existed more than one approach to the classification of the known world. And, indeed, the *philosophes*' intellectual and political bent made their way of organizing ideas different from that of other thinkers, particularly the Jesuits who were behind the *Journal de Trévoux*. Knowing this, we wanted to test the degree of overlap between the structure of knowledge in the *Encyclopédie* and the *Journal de Trévoux*, discovering in the process the commonalities, differences, and unique aspects of each. We hoped this experiment would give us a "slice of life" look at the intellectual milieu of the day, or at least provide an insight into the presumed differences in discourse between the two camps. To this end, we processed the ARTFL Project's 109 volumes of the *Journal de Trévoux*, splitting them into 1,027 separate articles. Extending from 1751 to 1757, this collection covers the years during which the editors of the *Journal* engaged in a fierce polemic with the encyclopédistes concerning the publication of the *Encyclopédie*.[16] Our previous model, trained on all of the classified *Encyclopédie* articles, was thus applied to the Trévoux articles, assigning each with a predicted class of knowledge. The twenty most frequent assigned classes are listed below: 19

Littérature	317
Morale	86
Géographie moderne	61
Théologie	54
Philosophie	50
Histoire moderne	46
Belles lettres	45
Astronomie	35
Métaphysique	30
Histoire ecclésiastique	26
Physique	22
Art militaire	18
Economie politique	18
Géographie	16
Medecine	16
Histoire romaine	14
Peinture	14
Histoire	14
Histoire naturelle	13
Chimie	12

Table 8. The 20 most frequent classes of knowledge assigned to the Trévoux articles using the Encyclopédie model.

A cursory glance at these results gives us a general idea about the most significant themes found in the *Journal de Trévoux*; themes that correspond nicely to our preconceived notions concerning the *Journal*, its writers and subject matter. Along these lines, it is not surprising to find in a Jesuit publication such as this a greater emphasis on articles about Literature, Morality, Theology, and Philosophy. However, the surprising fact that more than 1/3 of the 1000 articles were assigned the label of *Littérature* caused us to question somewhat the performance of the classifier and ultimately, to reconsider our modern notion of Literature when applied to the specific instances of this classification. 20

In the first edition (1694) of the *Dictionnaire de l'Académie française* the entry for “Littérature” reads thus: “ *Litterature. s. f. Erudition, doctrine. Grande litterature. profonde litterature. il est homme de grande litterature. il n'a point de litterature. il a beaucoup de litterature.* ” and indeed the definition changes little by the fourth edition of 1762: “ *LITTÉRATURE. s.f. Érudition, doctrine. Grande littérature. Profonde littérature. Il est homme de grande littérature. N'avoir point de littérature. Avoir beaucoup de littérature. Un ouvrage plein de littérature. Ce mot regarde proprement les Belles-Lettres.* ” The addition of the last sentence, “This word is properly used in regard to Belles Lettres,” in the 1762 edition seemingly restricts this particular form of erudition to the more traditionally literary realm of the “Belles-Lettres,” or Poetry and Rhetoric. The definition offered by the Jesuit editors of the *Dictionnaire de Trévoux* (1742) differs only slightly from that of the Academy: “LITTÉRATURE, s. f. Doctrine, connoissance profonde des Lettres. *Doctrina, litteratura, eruditio*.” While these definitions shed little light as to why the *Encyclopédie* literature class should be so prevalent in the classification of such a diverse collection of articles, many of which deal with the Sciences and Natural History, the ambiguity of this erudite possession of “ *littérature* ” and “ *lettres* ” should nonetheless cause us to broaden our 21 understanding of these terms as they were used in the mid-18th century.

We thus began a more thorough investigation of the Literature category by examining five randomly selected articles belonging to the assigned class “Littérature” in the *Journal de Trévoux*. While some categorizations make sense as literature — e.g., the article “Nouvelles Litteraires,” a sort of literary “news of the day”; and, less convincingly, a commentary on Rousseau's first discourse — others have ostensibly nothing to do with our modern idea of Literature — e.g., articles commenting on a history of jurisprudence, a treatise on diseases, and a compilation of treatises on Physics and Natural History[17]. This apparent anomaly necessarily leads us back to the *Encyclopédie* and the articles belonging to the class of knowledge Literature, which serve as the basis for these class assignments. 22

As we mentioned above concerning the reclassification of the article “Tepidarium,” there are a great many articles whose original classifications seem inappropriate. This phenomenon is all the more evident when examining the reclassification of the Literature articles, the majority of which deal more with Ancient History, Mythology, and Architecture than with accepted literary issues. Of the 682 Literature articles, 460 were written by the Chevalier de Jaucourt, author of more than 17,000 *Encyclopédie* entries. Jaucourt is known to have borrowed extensively from other sources and thus, we attributed these inconsistencies to intellectual laziness, given the enormous number of articles for which he was responsible.[18] Upon closer examination of the Literature class of knowledge however, this characterization proves unjust. 23

Indeed, the article titled “Littérature” belongs not to its own class of knowledge, but rather to three seemingly unrelated and disparate classes: Sciences, *Belles-Lettres*, and Antiquity. The text of the article, written by Jaucourt, is in fact a polemic advocating a universal erudition and an expanded definition of what it means to possess a great literature — in a word to be literate. Jaucourt includes a *renvoi* to the article “Lettres” in an effort to define better this notion of Literature. Following the cross-reference we find that the article in question, “Lettres,” an article that normally falls innocuously amongst numerous similarly titled entries, is the sole member of the class “Encyclopédie,” suggesting that the idea of literacy is essential to the entire encyclopedic enterprise. Here, Jaucourt's understanding of “Lettres” as a much larger category of knowledge than “belles-lettres” or even the Humanities as a whole (*les lettres humaines*), harkens back to the Classical acceptation of the term which encompassed all areas of human understanding from Epic Poetry to Physics. The inter-connectedness of knowledge, both literary and scientific, is thus the essence of Jaucourt's idea of encyclopedic literacy, wherein “ *il en résulte que les lettres & les sciences proprement dites, ont entr'elles l'enchaînement, les liaisons, & les rapports les plus étroits; c'est dans l'Encyclopédie qu'il importe de le démontrer* ” (“the result is that Letters and Sciences, properly speaking, have between each other a strong and direct network of links and relationships; it is in the Encyclopaedia that the demonstration of this network becomes important”).[19] 24

Conclusions and Future Work

This discovery — that for the writers of the *Encyclopédie*, Literature as a branch of human understanding included not only what we today consider Literature (*les Belles-Lettres*) but also Natural History, the Natural Sciences, the study of Antiquity, etc. — was made possible through the machine learning techniques outlined in this paper. We now understand precisely why the classifier, using the ontology of the *Encyclopédie*, labels so many Trévoux entries as “Littérature.” Furthermore, the presence of this more inclusive view of lettered erudition should come as no surprise given that, as is expressed in the very title of the *Encyclopédie*, this *Dictionnaire raisonné* was the work of a society of “Gens de Lettres.” 25

It would seem, however, that this notion of Literature as a sort of universal erudition did not survive the turmoil of the late 18th century, and by 1798 the fifth edition of the *Dictionnaire de l'Académie française* had already begun restricting Literature to a purely literary domain, defined as the “ *Connoissance des ouvrages, des matières, des règles, et des exemples littéraires* ” (our emphasis).[20] 19th century lexicography would move further in this direction, eliminating any mention of “doctrine” or “erudition” in its definition of Literature as “ *La science qui comprend la grammaire, l'éloquence et la poésie, et qu'on appelle autrement Belles-lettres* ” (*Dictionnaire de l'Académie française*, 6th edition 1832). By 1872, perhaps as a reflection of the disciplinary codification which took place during the first part of the 19th century, the positivist lexicographer Émile Littré simply defines Literature as the “ *Connaissance des belles-lettres* ” which is by and 26 large its accepted meaning today.

These modest conclusions lend further weight to our initial view that traditional humanistic inquiry can be enhanced and broadened through the judicious application of machine learning and data mining techniques. As large-scale textual resources such as the *Encyclopédie* become more readily available to scholars in a digitized format, new search and analysis tools will be needed. It is our opinion that approaches similar to those outlined above can successfully leverage the power of data mining tools for use in the Humanities. And while these techniques can certainly aid in a variety of hypothesis testing and classification tasks, it is our hope that they will also lead to the discovery of new knowledge through the uncovering of previously unnoticed textual connections. 27

Moving forward, we plan to continue improving the performance of our classifiers through better morphological text extraction, allowing for a greater freedom in the selection of features. Possible features would include n-grams (bi- and tri-grams of surface forms, bi- and tri-grams of lemmas) as well as part of speech information. We are also planning to investigate several different unsupervised machine learning techniques such as vector space analysis, latent semantic indexing (LSI), and several other clustering models. When applied to the *Encyclopédie* and other 18th century works, these tools will propose connections based on a measure of lexical similarity between arbitrary chunks of texts, whether paragraphs, articles, chapters, or entire works. It is our hope that within this new system, researchers will be able to explore and, more importantly, to evaluate the proposed connections between these articles and texts. Although the connections will be brought to light with the help of computers, it will nonetheless be necessary for scholars to provide the system with the critical element of human scrutiny that is essential to Humanities research. 28

Notes

[1] For an exhaustive treatment of the Encyclopédie and its authors, see [Schwab et al. 1971-1984]. For a more general discussion of the work, see [Proust 1995]. The ARTFL implementation of the *Encyclopédie* is discussed in [Morrissey et al. 2001], and [Andreev et al. 1999].

[2] English translation cited in [Hoyt and Cassirer 1965, xxiii] (our emphasis).

[3] For various representations of the *Système Figuré* and the Editors' descriptions, see http://www.lib.uchicago.edu/efts/ARTFL/projects/encyc/texts/ and http://artfl.uchicago.edu/cactus/.

[4] Blanchard and Olsen examined the structure of the *renvois* generating a “mappemonde” of the cross-references and node level classes of knowledge. See [Blanchard and Olsen 2002].

[5] Classes of knowledge were originally extracted automatically using a simple rule-based identifier, written in Perl, based on typographic conventions in the text. A small number of classified articles remain either misclassified or unclassified altogether.

[6] The normalized classifications are the result of a collaborative project with Professor Dena Goodman, Kevin Hawkins, and Benjamin Heller at the University of Michigan.

[7] Lemmatization was accomplished using TreeTagger, a probabilistic part-of-speech tagger and lemmatization utility developed by Helmudt Schmidt under the auspices of the TC Project at the University of Stuttgart. It is freely available for download from the project's official homepage at http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.

[8] See [Hand and Yu 2001].

[9] See [Witten and Frank 2005, 94–96].

[10] Success can be gauged in various ways. The accuracy achieved when testing our classifier against the same set of classified articles on which it was trained is generally indicative of the Naive Bayesian algorithm's power. However, this sort of validation often leads to over-fitting. Given that our accuracy never exceeded 75%, we can only assume that there are enough anomalies in the data to have compensated for any over-fitting. Additionally, we believe that it is these anomalies that often end up being the most revelatory.

[11] The accuracy of a classifier refers to its ability to correctly predict the label of an unseen instance. Typically, classifier accuracy is estimated using a test set that is independent of the training set, one that is often generated by hand-labeling a set of unseen documents and measuring how often the result of the classifier corresponds to the class given by the human tagger. However, asking modern scholars to classify a large

number of articles by hand would necessarily involve not only a large amount of effort, but also a fair amount of guesswork, particularly given the fact that we are dealing with an ontology that is more than 200 years old.

[12] This classification coincides nicely with the conclusions drawn by Martine Groult, one of our collaborators at the CNRS, whose work on the philosophy of *Discours préliminaire* is summarized here: http://encyclopedie.uchicago.edu/?q=node/162.

[13] It is worth noting that Williams' implementation of the Naive Bayesian classifier uses the log-likelihood ratio of features *fi* given each class c in question. It also uses a smoothing term to account for probabilities of zero. Because these probabilities are in general very small, their logarithmic weights are very negative. We find it easier to deal with positive numbers, so for display purposes, we make each of these scores positive by raising 2 to the power of that weight.

$$\text{classify}(f_1,\ldots,f_n) = \arg\max_c p(C=c)\prod_{i=1}^{n} p(F_i = f_i \mid C = c)$$

$$= \log(p(c)) + \sum_{i=1}^{n} \log(p(f_i \mid c))$$

$$\text{With smoothing term } \log(p(f_i \mid C)) =$$

$$\log(\mid f_i \mid +1) - \log(\mid F_C \mid + \mid F \mid)$$

Williams' Perl module can be downloaded at http://search.cpan.org/~kwilliams/Algorithm-NaiveBayes-0.04/lib/Algorithm/NaiveBayes.pm.

[14] The importance of word usage and the evolution of language in the *Encyclopédie*, key concepts when considering feature set analysis, are discussed in [Anderson 1984].

[15] The use and abuse of the category “Grammar” by Diderot is one of the subjects treated in [Leca-Tsiomis 1999].

[16] The ARTFL database of the *Journal de Trévoux* can be found at http://www.lib.uchicago.edu/efts/ARTFL/projects/trevoux.

[17] The full titles of the “Literature” articles we uncovered are as follows: *1) ARTICLE XII. NOUVELLES LITTERAIRES. 2) ARTICLE XXIX. DISCOURS QUI A REMPORTE' le prix à l'Académie de Dijon en l'année 1750, sur cette question proposée par la même Académie: Si le rétablissement des Sciences & des Arts a contribué à épurer les moeurs. Par un Citoyen de Genève. 3) ARTICLE XXXV. HISTOIRE DE LA JURISPRUDENCE Romaine. 4) ARTICLE LXI. TRAITÉ DES MALADIES qu'il est dangereux de guérir. 5) ARTICLE XIV. RECUEIL DE DIFFERENS Traités de Physique & d'Histoire Naturelle.*

[18] For a thorough discussion of Jaucourt's contributions to the *Encyclopédie*, see [Lough 1973].

[19] From the article “Lettres” in the *Encyclopédie*.

[20] Note that we are focusing exclusively on the definitions of Literature as a discipline, rather than a collection of literary and/or other documents (as in “I have consulted all the available literature on cancer and have found nothing”). This modern usage of literature is expressed somewhat in the dictionaries we mention, i.e. “ *L'ensemble des productions littéraires d'une nation, d'un pays, d'une époque. La littérature française. La littérature du moyen âge* ” (Émile Littré, *Dictionnaire de la langue française* (1872-77).

Works Cited

Anderson 1984 Anderson, Wilda. *Between the Library and the Laboratory: The Language of Chemistry in Eighteenth-Century France*. Baltimore: Johns Hopkins University Press, 1984.

Andreev et al. 1999 Andreev, Leonid, Jack Iverson and Mark Olsen. “Re-Engineering A War Machine: ARTFL's Encyclopédie”. *Literary and Linguistic Computing* 14: 1 (1999), pp. 11-28.

Blanchard and Olsen 2002 Blanchard, Gilles, and Mark Olsen. “Le système de renvois dans l'Encyclopédie: une cartographie de la structure des connaissances au XVIIIème siècle”. *Recherches sur Diderot et sur l'Encyclopédie* 31-32 (2002), pp. 45-70.

Brian 1998 Brian, Eric. “L'ancêtre de l'hypertexte”. *Les Cahiers de Science et Vie* 47 (1998), pp. 28-38.

Hand and Yu 2001 Hand, David J., and Keming Yu. “Idiot's Bayes -- Not So Stupid after All?”. *International Statistical Review* 69: 3 (December 2001), pp. 385-398.

Hoyt and Cassirer 1965 Hoyt, Nelly, and Thomas Cassirer. *Encyclopedia: Selections by Diderot, D'Alembert, and a Society of Men of Letters*. Indianapolis: Bobbs-Merrill, 1965.

Leca-Tsiomis 1999 Leca-Tsiomis, Marie. *Ecrire L'Encyclopédie: Diderot: de l'usage des dictionnaires à la grammaire philosophique*. Oxford: Voltaire Foundation, 1999.

Lough 1973 Lough, John. *The Contributors to the Encyclopédie*. London: Grant and Cutler, 1973.

Morrissey et al. 2001 Morrissey, Robert, Jack Iverson and Mark Olsen. “Présentation: L'Encyclopédie Electronique”. In Robert Morrissey and Philippe Roger, eds., *L'Encyclopédie du réseau au livre et du livre au réseau*. Paris: Champion, 2001. pp. 17-27.

Proust 1995 Proust, Jacques. *Diderot et L'Encyclopédie*. Paris: Albin Michel, 1995.

Schwab et al. 1971-1984 Schwab, Richard, Walter Rex and John Lough. *Inventory of Diderot's Encyclopédie*. Oxford: Studies on Voltaire and the Eighteenth Century, 1971-1984.

Witten and Frank 2005 Witten, Ian, and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann, 2005.