frontiers in Digital Humanities

ORIGINAL RESEARCH
published: 12 January 2016
doi: 10.3389/fdigh.2015.00008

Discourses and Disciplines in the Enlightenment: Topic Modeling the French *Encyclopédie*

Glenn Roe[1]*, *Clovis Gladstone*[2] *and Robert Morrissey*[2]

[1] *Centre for Digital Humanities Research, Australian National University, Canberra, ACT, Australia,* [2] *The ARTFL Project (American and French Research on the Treasury of the French Language), University of Chicago, Chicago, IL, USA*

This article describes the use of latent Dirichlet allocation (LDA), or topic modeling, to explore the discursive makeup of the eighteenth century *Encyclopédie* of Denis Diderot and Jean le Rond d'Alembert (1751–1772). Expanding upon previous work modeling the *Encyclopédie*'s ontology, or classification scheme, we examine the abstractions used by its editors to visualize the various "systems" of knowledge that the work proposes, considered here as heuristic tools for navigating the complex information space of the *Encyclopédie*. Using these earlier experiments with supervised machine-learning models as a point of reference, we introduce the notion of topic modeling as a "discourse analysis tool" for Enlightenment studies. In so doing, we draw upon the tradition of post-structuralist French discourse analysis, one of the first fields to embrace computational approaches to discursive text analysis. Our particular use of LDA is thus aimed primarily at uncovering interdisciplinary "discourses" in the *Encyclopédie* that run alongside, under, above, and through the original classifications. By mapping these discourses and discursive practices, we can begin to move beyond the organizational (and physical) limitations of the print edition, suggesting several possible avenues of future research. These experiments thus attest once again to the enduring relevance of the *Encyclopédie* as an exemplary Enlightenment text. Its rich dialogical structure, whether studied using traditional methods of close reading or through the algorithmic processes described in this article, is perhaps only now coming fully to light thanks to recent developments in digital resources and methods.

Keywords: machine learning, topic modeling, discourse analysis, enlightenment, eighteenth century history, *Encyclopédie*

OPEN ACCESS

Edited by:
Jean-Gabriel Ganascia, University Pierre and Marie Curie, France

Reviewed by:
Claire Clivaz, Swiss Institute of Bioinformatics, Switzerland
Sara Tonelli, Fondazione Bruno Kessler, Italy

***Correspondence:**
Glenn Roe
glenn.roe@anu.edu.au

Specialty section:
This article was submitted to Digital Literary Studies, a section of the journal Frontiers in Digital Humanities

Received: *28 May 2015*
Accepted: *21 December 2015*
Published: *12 January 2016*

Citation:
Roe G, Gladstone C and Morrissey R (2016) Discourses and Disciplines in the Enlightenment: Topic Modeling the French Encyclopédie. Front. Digit. Humanit. 2:8. doi: 10.3389/fdigh.2015.00008

INTRODUCTION: GRAPHS, MAPS, AND TREES

In many ways, the eighteenth century French *Encyclopédie*, created under the direction of Denis Diderot and Jean le Rond d'Alembert between 1751 and 1772, seems to have almost been designed as a document classification exercise. For one, its structure, spread over 17 *in-folio* volumes of text and another 11 volumes of engravings, comes complete with a branching, hierarchical ontology, or classification scheme. Of the almost 75,000 articles contained in the *Encyclopédie*, some 62,000 were classified by the editors according to this ontology, while, for a variety of editorial reasons, 13,000 or so were left with no explicit classification. All references to the text, articles, and classes of the *Encyclopédie* are drawn from the digital edition made available by the ARTFL Project at the

University of Chicago.[1] This corpus of articles, both classified and unclassified, thus provides a readymade training and evaluation set on which to deploy a variety of machine-learning algorithms. For the purposes of this article, we propose the use of the latent Dirichlet allocation (LDA) topic-modeling algorithm as an exploratory "discourse analysis" tool. Our main contention is that LDA – used here primarily as a form of "exploratory data analysis" (EDA) – can help us move beyond the binary either/or logic of supervised classification and toward a better understanding of the rich, multivocal discursive structure of the *Encyclopédie*.

Any algorithmic approach to the *Encyclopédie*'s ontology is rendered particularly challenging, however, given that the editors at times applied classifications with polemical intent and/or varying degrees of rigor, thus rendering subject boundaries somewhat fuzzy [see Proust (1995) and Leca-Tsiomis (1999)]. Taken together, the 2,899 individual classes, or disciplines, that make up the *Encyclopédie*'s classificatory "system" (if it was, in fact, designed to be systematic) span the entire breadth of human knowledge. And, within the disciplines themselves, individual articles can range from brief cross-references or *renvois* containing only a few words (e.g., "ALCALI. Voyez ALKALI"), to protracted philosophical and historical treatises extending over many pages. In short, the *Encyclopédie*'s built-in ontology has everything to make a machine learner at once happy and miserable.

The complex nature of the *Encyclopédie*'s classification scheme was famously laid out by d'Alembert in his introductory essay, the "Discours préliminaire" (1751). In this seminal philosophical text, d'Alembert describes the overall "system" of human understanding, which he illustrated through the image of a tree: "After reviewing the different parts of our knowledge and the characteristics that distinguish them, it remains for us only to make a genealogical or encyclopedic tree which will gather the various branches of knowledge together under a single point of view and will serve to indicate their origin and their relationships to one another" (D'Alembert, 1751, p. 14). D'Alembert's tree of knowledge, more diagrammatic than arboreal at this point, was subsequently published as the "Système figuré des connaissances humaines" at the very beginning of the first volume of the *Encyclopédie* (see **Figure 1**).

The use of the tree structure/metaphor as an organizational and genealogical visualization technique has a long history (Lima, 2014). Its appeal as an informational abstraction is made clear in the above image, as its various roots and branches allow readers to see – in the words of d'Alembert, "under a single point of view" – the inter-connectedness of the encyclopedic disciplines, from trunk to leaf. But, while the abstract nature of the "système figuré" was in many ways more illustrative than comprehensive, it was also somewhat reductive in its treatment of the disciplines as a whole. Nonetheless, the illustrative nature of d'Alembert's abstract system was also operative within the overall organizational structure of the text, which readers could navigate alphabetically, taxonomically (using the classes of knowledge), and dialogically (using the cross-references). In fact, as d'Alembert asserts later in the "discours préliminaire," the "système figuré" might be better understood as a map than a tree, and but one of many maps to be used to consult the *Encyclopédie*:

> It is a kind of world map which is to show the principal countries, their position and their mutual dependence, the road that leads directly from one to the other. This road is often cut by a thousand obstacles, which are known in each country only to the inhabitants or to travelers, and which cannot be represented except in individual, highly detailed maps. These individual maps will be the different articles of the Encyclopedia and the Tree or Systematic Chart will be its world map (D'Alembert, 1751, p. 15).

Thus, from a functional standpoint, the tree of knowledge works much like an XML document object model with its roots (parent elements) and various branches and leaves (child or sub-child elements): it allows readers to move from any given point in the general trunk of knowledge to the disciplinary branches that grow from it, or to the individual nodes or leaves – the articles. Whether treated as a graph, map, or tree, however, the system of knowledge on these abstractions are meant to represent can only ever be perspectival in nature, depending on the particular vantage point of the geographer (d'Alembert would say "philosopher") who assembles them. Maps of knowledge, d'Alembert continues, are thus necessarily as numerous as there are mapmakers:

> But as, in the case of the general maps of the globe we inhabit, objects will be near or far and will have different appearances according to the vantage point at which the eye is placed by the geographer constructing the map, likewise the form of the encyclopedic tree will depend on the vantage point one assumes in viewing the universe of letters. Thus one can create as many different systems of human knowledge as there are world maps having different projections, and each one of these systems might even have some particular advantage possessed by none of the others (D'Alembert, 1751, p. 15).

Here, the desire for "a single point of view" explodes into a rich multiplicity of innumerable different points of view. In the end, d'Alembert seems to be telling us that the ontology he and Diderot have generated for the *Encyclopédie* is largely subjective in nature, and must therefore be understood more as an heuristic tool – one of many possible such tools – rather than a comprehensive representation of the faculties of human understanding. From our perspective, d'Alembert's "system" can be thought of as pre-digital finding aid of sorts that, when combined with the other organizational strategies at play in the work, such as the cross-references, allowed users to navigate through the various dialogic layers of the encyclopedic text. On the *Encyclopédie*'s cross-referencing scheme, for example, see Brian (1998), Blanchard and Olsen (2002), Bianco (2002), Melançon (2004), and Guénard et al. (2006).

There is no single system of knowledge, then, that can fully represent the epistemological richness of human experience. Rather, single domains of knowledge, such as mathematics, must

[1] http://encyclopedie.uchicago.edu/

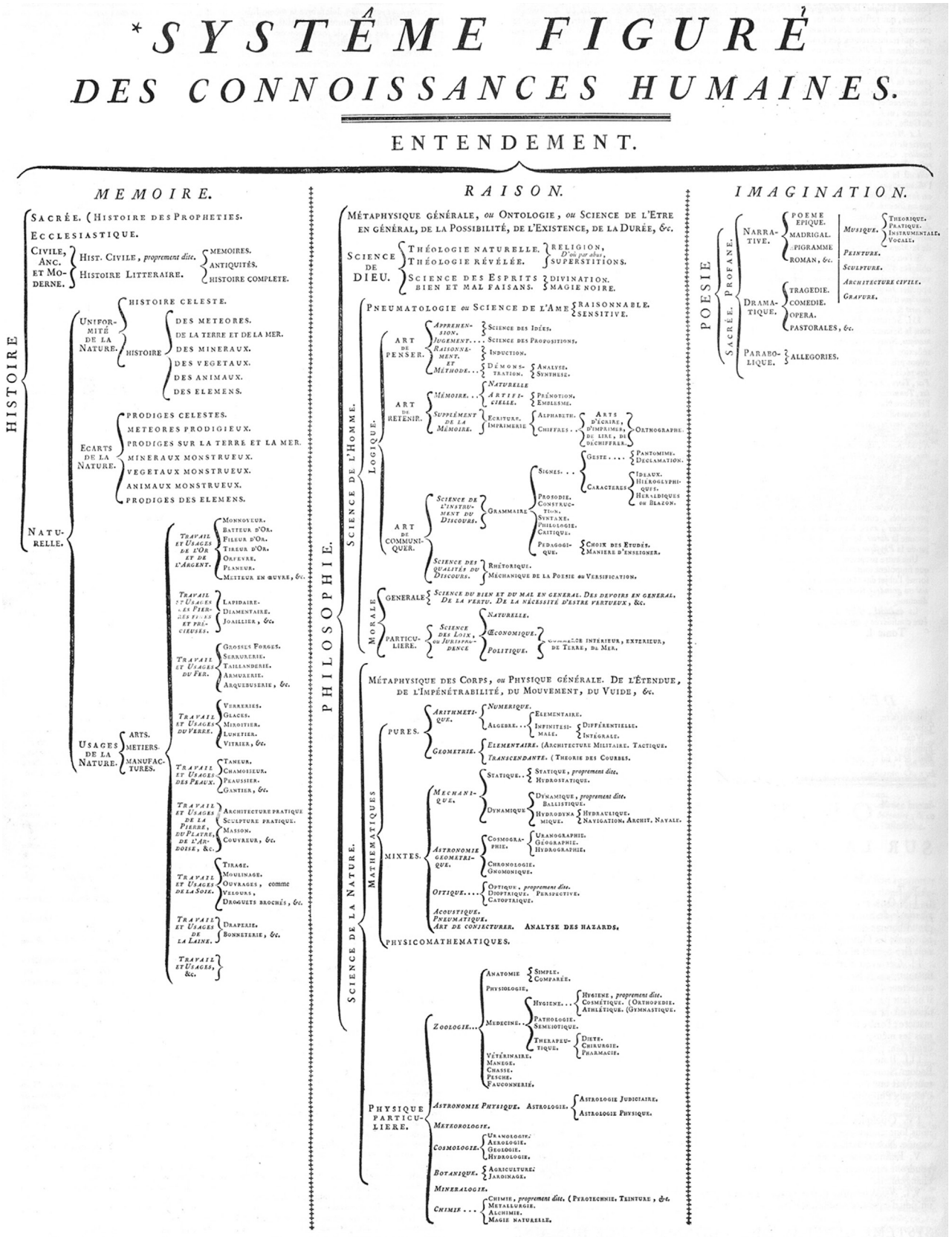

FIGURE 1 | "Système figuré des connaissances humaines," ***Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers*****, vol. 1, 1751**.

be reduced down to their most basic principles, in order to situate them with other similarly reduced branches on the tree of knowledge. According to d'Alembert, "this reduction which, moreover, makes them easier to understand, constitutes the true 'systematic spirit' [*esprit systématique*]. One must be very careful not to mistake this for the 'spirit of system' [*esprit de système*], with which it does not always agree" (D'Alembert, 1751, p. 6). D'Alembert is here drawing a clear distinction between the "esprit de système" of seventeenth century – embodied in the totalizing philosophical systems of thinkers, such as Descartes, Spinoza, and Leibniz, among others – and the more open-ended, and experimental, "esprit systématique" of the eighteenth century *philosophers*. According to Cassirer (2009), "The whole theory of knowledge of the eighteenth century strives to confirm this distinction," i.e., between the inductive and rational "esprit systématique" and the deductive, rationalistic "esprit de système." The systematic spirit of the *philosophers* was thus the guiding force behind the *Encyclopédie*'s multiple organizational structures; local systems, or maps, which were meant to guide engaged readers outward, toward the multiplicity of global maps found in empirical reality, and finally reach the limits of human understanding.

The combinatorial nature, and almost infinite number, of heuristic approaches that d'Alembert imagines in the "discours préliminaire" – "one can create as many different systems of human knowledge as there are world maps" – were nonetheless constrained by the material limitations of the print medium, a fact that led, for instance, to many spurious cross-references. In the digital medium, however, no such limitations exist. And, as far as the *Encyclopédie* is concerned, the ARTFL Project (American and French Research on the Treasury of the French Language) at the University of Chicago has been exploring the new terrain of the digital edition since the late 1990s (Andreev et al., 1999; Morrissey et al., 2001). More recently, given the advent of "big data" approaches to humanities collections in the last decade or so, the complex ontology of the *Encyclopédie* outlined above has proven well suited for exploration using contemporary information retrieval techniques and machine-learning algorithms.

By allowing us to classify and reclassify articles, to cluster and re-cluster disciplines, to construct, deconstruct, and reconstruct the various systems or maps of knowledge at work in the *Encyclopédie*, machine-learning techniques can move us toward a better understanding of the truly radical epistemological underpinnings of Diderot and d'Alembert's *magnum opus*. What is more, this process of experimentation – echoing the "systematic spirit" of the *philosophers* themselves – allows us to bring our deep knowledge as humanists and domain specialists to bear on these various "black-box" algorithmic approaches. This is particularly relevant for an approach, such as LDA/topic modeling, which relies on complex probabilistic models and random sampling to generate results that often differ from one run to the next. This makes systematic evaluation of topic models somewhat problematic and is the subject of ongoing work in Computer Science (Wallach et al., 2009; Newman et al., 2010; Ramirez et al., 2012). For our purposes, LDA is used primarily as a form of EDA, which can hopefully provide insights into the *Encyclopédie*'s discursive makeup over and above its original classification system. In this context, the *Encyclopédie* becomes a crucial tool in deploying and critically assessing a range of new technologies that can in turn uncover hidden patterns and connections in this 250-year-old text. These new techniques thus become themselves heuristic tools – much like d'Alembert's maps and trees – that allow us to explore the *Encyclopédie* in ways previously unimaginable. Experiments described in this article can thus open up new avenues of research on the *Encyclopédie*, attesting to the virtuality and potentiality of the work today, not merely as an inanimate historical artifact, but rather as a living digital edition that harkens back to the editors' original goals.

SUPERVISED MACHINE LEARNING AND THE ENCYCLOPEDIC DISCIPLINES

Collectively, we at ARTFL have been mining the rich vein of the *Encyclopédie* for some time now (Cooney et al., 2008; Roe et al., 2008; Horton et al., 2009; Allen et al., 2010). Our initial attempts at document classification using a Naive Bayesian algorithm to generate a computational ontology of the *Encyclopédie*'s classification scheme, and our critical interpretation of the results, were first presented at the *Digital Humanities 2007* conference in Urbana-Champaign (Horton et al., 2009). We followed these efforts up at *Digital Humanities 2008* in Finland with an expanded toolkit that included *k*-nearest neighbor vector space classifications, a meta-classifying decision tree, and centroid-based clustering (Cooney et al., 2008; Roe et al., 2008). Where we had previously achieved 72% accuracy in categorizing medium length and long articles using Naive Bayesian alone, using these new tools and a combination of multiples classifiers, we were able to get similar rates of accuracy (up to 77%) over the entire encyclopedia, including the very short articles, which are quite difficult to classify due to their dearth of distinctive content. This is, in non-binary classification terms, a very good result, given the large number of categories in play and their significant overlaps ("modern" vs. "ancient geography," for example), as well as the far from coherent nature of the original classification scheme. These experiments were preformed using the PhiloMine machine-learning environment developed at ARTFL in 2007–2008.[2] All 74,000 *Encyclopédie* articles were used for testing and training, though we limited our features to words that occurred in >5% and <95% of the overall corpus.

We initially explored a wide range of classifiers, including Naive Bayesian, Support Vector Machines, and *k*-NN vector space, with a range of parameters for word count normalization and other settings [see Cortes and Vapnik (1995), Domingos and Pazzani (1997), and Hadi et al. (2007)]. After examining hundreds of such combinations, we found two combinations that provided the greatest accuracy in correctly re-classifying articles to their previous classifications: Naive Bayesian, using simple word counts, and *k*-NN, using 50 neighbors and *tf-idf* values for the feature vectors [on *tf-idf*, see Salton and Buckley (1988)]. Each classifier alone was right about 64% of the time (64.3% for Naive Bayesian and 63.8% for *k*-NN) – but together, at least one of them was right 77% of the time. It turned out that *k*-NN was most accurate

[2]https://code.google.com/p/philomine/

on smaller articles and smaller classes, whereas Naive Bayesian worked best on longer articles that belonged to larger classes. Using these observations, we were able to iteratively modify the classification rules based on which classifier should decide class membership when they were at odds with each other. Thus, by feeding the article and class metadata (including word frequency counts) into a simple decision tree classifier, along with the results of each classifier, we were able to teach the system which classifier to prefer for a given decision if there was disagreement on the initial class assignment.

Of course, it is not feasible for classifiers to make the "right" decision every time, but we were able to increase our overall accuracy on previously classified articles to just above 73%, or 9% higher than the average of the individual classifiers alone. The resulting machine-generated ontology was optimized from the original 2,899 classes down to 360 more expansive categories – a loss of specificity that, while necessary for computational efficiency, is nonetheless worrisome from a humanistic perspective. But, our goal in re-classifying the *Encyclopédie* was never to replace the original labels, and indeed they remain the canonical point of reference in our edition. Rather, we wanted to use these supervised machine-learning techniques to generate our own abstraction of the *Encyclopédie*'s classification system based on the lexical similarity of the classes. This is an abstraction, we would argue, that is intellectually in line with d'Alembert's previous graph/tree abstractions.

By making this machine-generated ontology available to *Encyclopédie* users (see http://encyclopedie.uchicago.edu/content/machine-classifications), we hoped, ideally, to provide better access to the entire *Encyclopédie* through the addition of class descriptions (imperfect as they may be) to previously unclassified articles. We also hoped to identify articles that are re-classified differently from their original class, allowing users to find articles through alternative heuristics. And, finally, we aimed to identify interesting patterns in the authors' use of the inherited classification system, again primarily by looking at articles that are re-classified differently. Anecdotally, we know of many instances of this sort of subversive classification practice. One need only think of the Chevalier de Jaucourt's use of the "modern geography" class, wherein biographical entries were often incorporated into a geography article – Newton's biography occurs in the modern geography article "Wolstrope," for example, which is the philosopher's birthplace. Now, however, we can begin to think about mechanisms for identifying more systematically these sorts of authorial practices, due in no small part to our machine-learning experiments.

Although the information gain of these supervised learning approaches has been significant, we remain nonetheless constrained by the editors' original classification scheme, a fact that can lead to neverending, and wholly unproductive, conversations regarding "right" and "wrong" class assignments. These sorts of judgments miss the point, in our opinion, of the experimental nature of this work, designed to push existing boundaries both in terms of the notion of the *Encyclopédie* as a "stable" text and also in the use of algorithmic approaches to bring new knowledge to bear on traditional humanities domains. To this end, we began experimenting with several unsupervised machine-learning algorithms, which are concerned more with the clustering of texts with no *a priori* categorizations. What sort of lexical patterns existed, we wondered, under, between, and through the disciplines that we had previously explored? How could we go about identifying these patterns or lexical clusters, and, once identified, what, if anything, did they mean? To answer these questions, we turned, on the one hand, to the post-structuralist notion of "discourse analysis," and, on the other, to the relatively new unsupervised machine-learning algorithm, latent Dirichlet allocation, more commonly known as "topic modeling."

DISCOURSE ANALYSIS

Discourse analysis has over the past half century become a staple method of text analysis in the historical and social sciences, and in particular, for the literary and intellectual history of the French Enlightenment and Revolutionary periods. French discourse analysis (FDA), as distinct from the more general field of discourse analysis in functional linguistics, was nonetheless part and parcel of the "linguistic turn" of the human and social sciences in France during the 1960s. As Glyn Williams tells us: "FDA derives from the philological tradition of textual reflection, but in a more specific sense it derives from structuralism and post-structuralism, which, in turn, drew upon Russian formalism. As such it draws upon Saussurean linguistics, the philosophic work of Foucault, the psychoanalysis of Lacan and Althuserrian Marxism" (Williams, 1999, p. 3). Alongside its decidedly post-structuralist roots, FDA was also one of the first disciplines to embrace computational text processing and analysis. In the mid 1960s, for instance, researchers at the Centre for Political Lexicometry (École Normale Supérieure de Saint-Cloud) began using computers to study political discourse for the first time. A few years later, Pêcheux (1969) – a former student of Louis Althusser – would develop an early computer program called *automatic discourse analysis* (ADA), to identify ideological processes in textual corpora.

Pêcheux and his team sought a computational method for uncovering hidden ideological meanings in text corpora. To do this, he developed a formal, potentially automatic computer program, which he called ADA. The program could, in theory, generate a structuralist description of a discourse by identifying and describing relations of selection and substitution of syntactic elements in a corpus of texts representing that discourse (Helsloot and Hak, 2001). Gradually, Pêcheux would move away from a strictly structuralist approach to discourse by working through criticisms of ADA and attempting to overcome its limitations. In so doing, he developed a more reflective theory of "interdiscourse" in which he tried to account for the discursive dynamics of ideological struggle. According to Pêcheux (1982), the meaning of a discourse "does not exist anywhere except in the metaphorical relationships (realized in substitution effects, paraphrases, synonym formations), which happen to be more or less provisionally located in a given discursive formation: words, expressions, and propositions get their meanings from the discursive formation to which they belong" (p. 188). Here, Pêcheux's notion of "substitution" as it relates to the formation of discourse is quite close to Paul Ricœur's work on the metaphor and the "semantics of

discourse" in *La métaphore vive* (Ricœur, 1975). Unlike Ricœur, however, Pêcheux attempts to recast the Saussurean distinction between *langue* and *parole* as one between "linguistic basis" and "discursive process" wherein certain definable linguistic units (such as the use of relative clauses, for instance) are related to larger discursive and ideological formations and thus to the struggle for political hegemony.

In the same year that Pêcheux published his work on ADA, Foucault (1989), in his ground-breaking work *L'Archéologie du savoir*, likewise attempted to move the concept of "discourse," and of the underlying power politics at play in its formation, away from the purely structural or linguistic acceptations of the day. While Pêcheux's brand of ideological analysis was in many ways too specific in scope and method to be widely taken up by the public, Foucault's broader concept of "archeology" ushered in a radically new, and ultimately widespread, approach to the discursive at an extra-linguistic and even extra-textual level:

> In analyzing discourses themselves, one sees the loosening of the embrace, apparently so tight, of words and things, and the emergence of a group of rules proper to discursive practice […] A task that consists of not – of no longer – treating discourses as groups of signs (signifying elements referring to contents of representations) but as practices that systematically form the objects of which they speak (Foucault, 1989, p. 54).

This expanded notion of discourse, and of the "discursive practices" that writers consciously or unconsciously engage in, would go on to exert a profound influence on French historical studies, and in particular, on the historiography of the French Enlightenment and Revolutionary periods. Beginning in the 1980s, and building upon the work of François Furet, historians such as Lynn Hunt and Keith Baker examined the political discursive practices that shaped not only the formation of the French Revolution for its actors but also our reception and understanding of its cultural meaning over the *longue durée* (Furet, 1978; Hunt, 1984; Baker, 1990). More recently, works by Edelstein (2009, 2010) and de Bolla (2013) have re-introduced the specifically linguistic elements of discourse analysis back into the historian's and literary scholar's toolbox, most notably through the use of newly available historical and natural language databases for textual analysis.

With these previous efforts in mind, and given the rapid growth of available digital text collections, a revisiting of Pêcheux's notion of an "Automatic Discourse Analysis" would again seem warranted. And, although Pêcheux's early attempts at computational discourse analysis were eclipsed by the broader notions of Foucauldian discursive and "enunciative" practices, recent developments in information retrieval, such as latent semantic analysis (LSA) and LDA, or topic modeling, are perhaps finally suited to the task. It is not unreasonable, for instance, to posit that Foucault's concept of archeology, in fact, justifies the "bag of words" analytical model used by topic modeling and other machine-learning algorithms; a model that has often come under scrutiny (for good reason) by humanists. By locating words within a set of discursive practices rather than linguistic rules, Foucault's concept of discourse frees us from exclusive interest in language structure, and what that structure conveys, and orients us more toward the association of the various words, concepts, or "topics" that form a discourse. From this perspective, topic modeling, and the "bag of words" model that underlies it, can be used to identify multiple discourses in text collections based on the probabilistic co-occurrence of words in the same discursive context. Computer scientists call these clusters of co-occurring words "topics," we prefer to think of them as "discourses."

TOPIC MODELING AS A DISCOURSE ANALYSIS TOOL

Topic modeling is an unsupervised machine-learning approach that was originally designed as a way to classify and analyze large amounts of unlabeled data (Blei, 2012). In today's information-rich environment, where written online production and mass digitization efforts are on the rise, this method provides an efficient way to organize and summarize data automatically, with minimal human intervention. Moreover, for historical data, Newman and Block (2006) have demonstrated through their use of probabilistic latent semantic analysis (pLSA, a method similar to topic modeling) on an eighteenth century colonial newspaper, that such unsupervised algorithms can provide a unique overall picture of the contents of a corpus by organizing the data in a manner that avoids "fallible human indexing or their own preconceived identification of topics." In other words, unsupervised learners can provide multiple perspectives onto seemingly intractable data sources – freed, initially at least, from human bias and preconceptions – which can then enable new insights into both unknown resources and already well-studied texts (Noh et al., 2011).

As with any new computational approach, these methods should in no way be accepted without reservation. And, as literary scholars, we should remain resolutely (though hopefully constructively) critical of these techniques, and in particular of the "bag of words" model that underlies them. These reservations notwithstanding, we believe that topic modeling offers real promise for the exploration of discursive practices at work in and between texts over time. In order to compare the analytic possibilities of both supervised and unsupervised learners, then, we chose the LDA topic-modeling algorithm as our point of departure. First described by Blei (2003), LDA is built upon the important premise that documents, however, focused, are never about one single topic, but are instead the result of multiple topics bound together in a single unit of text. Consequently, the documents analyzed by this algorithm will be identified by a unique signature, a distribution of topics that represents the variety of each document's discursive content. In other words, while this algorithm puts forward each document's uniqueness, it also provides the ability to create very different "maps" of texts depending on what topics are being considered, creating what d'Alembert called the "single point of view" from which to grasp the knowledge system from above.

The results of a topic modeler do not always, however, provide clearly interpretable topics (represented as a list of words ordered by weight), which could lead one to dismiss the entire model based on a perceived incongruity. We would argue, however, that

the usefulness of a topic model does not necessarily rest on its ability to provide meaningful topics (a subjective categorization) for the corpus being analyzed, but rather on the multiplicity of perspectives it can generate and, as a result, on the potential for discovery that some of these topics can offer. As stated above, we are interested in topic modeling primarily as a form of EDA with which to investigate and scrutinize the complex discursive makeup of texts, and the unique distribution of topics that contribute to their semantic content. It is important to note, however, that the words that make up a given topic have no semantic or thematic relationship in and of themselves; they represent the words that most probabilistically co-occur with each other over the entire vocabulary (minus stopwords) of a given corpus of documents. Any labels we assign to these word groupings, or semantic content we posit in them, are thus largely perspectival in nature (in much the same way as d'Alembert's maps mentioned above), and should be treated as such.

Given the above caveats, we are nonetheless confident that topic modeling can be brought to bear fruitfully on the larger field of discourse analysis. In particular, we see many commonalities between LDA and the strain of FDA that runs through the work of Michel Pêcheux and Michel Foucault in the 1960s and 1970s – concerned as they both were with the formation of discourses through the association of certain lexical commonalities, of words and things, deployed in varying discursive contexts. Our contention is that LDA can be used as a discourse analysis tool in which unseen discursive practices can be brought to light through the careful analysis of topic distributions in large heterogeneous document collections.

We certainly do not intend to use every single topic in our topic model (or indeed over the multiple models we generated), as our goal here is more that of a proof-of-concept for the method rather than a systematic application of one particular model, nor do we aim to evaluate fully the performance of LDA as an unsupervised machine-learning approach. Rather, our objective is to uncover trends and patterns of interest within the results provided by the LDA algorithm. As such, the ultimate usefulness of topic modeling as a discourse analysis tool does not rest on a strict evaluation of the performance of the algorithm, which is outside of the scope of our work and expertise, but on its ability to provide valuable insights to researchers working in a variety of interdisciplinary fields. For our purposes, we wanted to think about Pêcheux's notion of an ADA system for use in the broad field of French Enlightenment studies. But, in order to approach this vast subject, we chose to limit ourselves to further experiments with Diderot and d'Alembert's *Encyclopédie*, taken again as an exemplary text both of Enlightenment discourse and of contemporary eighteenth century ideas.

TOPIC MODELING THE *ENCYCLOPÉDIE*

In applying topic-modeling algorithms to the *Encyclopédie*, our aim is to use LDA to go beyond the disciplinary boundaries of the editors' original classification scheme outlined earlier. This will provide us with a more transversal view of this important text and of the discursive makeup of its contents. Whereas Blei (2013) has asked: "What is the likely hidden topical structure that generated my observed documents?," we would add: "what are the non-obvious discourses and discursive practices that span across multiple disciplines in the *Encyclopédie*?"

In order to achieve our goals, we are using the well-known machine-learning toolkit MALLET.[3] As is often necessary in data mining experiments, we created a stopword list in order to filter out function words that tend to occur on a very frequent basis. This has the benefit of significantly reducing the dimensionality of the data model, since far fewer words are being processed. Another necessary step when generating a topic model, one must supply the algorithm with the number of topics that seem to best encompass the entire data set. While the arbitrariness of this choice is often brought up in the literature on LDA (Schmidt, 2012), for our purposes the notion of finding an "optimal" or perfect number of topics is largely irrelevant. In the case of the *Encyclopédie*, building multiple models based on a different number of topics simply results in so many new maps to be explored. As the *philosophers* themselves were mapmakers, we contend that topic models constitute another way of charting the *Encyclopédie*. With this in mind, we decided to generate several of these maps, with 280, 300, 330, and 360 topics. Admittedly, these numbers are not completely random, and were consistent with our previous machine classification experiments, whose generated ontology included 360 classes.

Once our topic models were generated, we stored the results for each article in a SQLite table, along with all of their corresponding metadata. We wrote a web interface in Python to query this database and run searches against the original metadata, such as finding the most important topic for any given author. Using this interface, we examined our four separate topic models, based on 280, 300, 330, and 360 topics, respectively. At this preliminary stage, we found that the topics in the 280-topic model were sufficiently coherent for labeling and searching. In other words, the new map of 280 topics we had generated provided, from our perspective, an appropriate overview of the discursive makeup of the text. We were thus able to identify a great many of the disciplinary discourses of the *Encyclopédie* in our topic lists of keywords – i.e., the 20 most significant words for each generated topic, a common method for scrutinizing the topic model. We should note, however, that only showing the first 20 or so words of a topic, while common practice, is not without its detractors [see Schmidt (2013)] nor is the act of labeling topics in any way straightforward or free from interpretive bias. But, as we are concerned primarily with the notion of topics as "discursive formations" (to borrow again a term from Foucault) and their deployment, rather than any definitive interpretation of their content, we find both the word lists and labels useful shorthand for distinguishing topics and the discursive practices they contain.

If we examine how of our topic model processes the short article "Fish" (*Poisson*), for example, which belongs to the discipline of "Sacred criticism" in the *Encyclopédie*, we notice that there are two main and two secondary topics that contribute to its discursive makeup (**Table 1**). A close reading reveals that this article is mostly (and perhaps predictably) about the anatomy of fish, along with related practices of the Ancient Jews. The topic distribution thus reflects a more complex discursive signature than indicated

[3] http://mallet.cs.umass.edu/

TABLE 1 | Article "fish" and its distribution of topics.

Poisson (*Critiq. sacrée*.) Moise met les poissons au nombre des reptiles; l'Histoire naturelle n'étoit pas encore cultivée chez les Juifs dans le tems du regne de ce législateur. Comme il y a des poissons qui ont des écailles sans nageoires, & d'autres qui n'ont ni nageoires ni écailles, Moise fonda sur cette différence sa distinction des poissons purs & immondes. Il mit ceux qui n'ont ni nageoires ni écailles au rang des poissons impurs, & defendit d'en manger, ne permettant l'usage que des poissons qui ont des nageoires & des écailles L'Ecriture désigne quelquefois figurément les hommes sous le nom de poissons; les poissons de vos rivieres tiendront à vos écailles, dit Ezéchiel xxix. 4. c'est – a – dire la perte de vos sujets sera inséparable de la vôtre La porte des poissons, Sophon. j. 2. étoit une porte de Jérusalem, ainsi nommée parce que c'étoit par – là qu'on apportoit le poisson dans la ville	Topic #233 (Fish anatomy): 0.329 Topic #194 (Jewish practices): 0.229 Topic #65 (Ancients' practices): 0.0719 Topic #134 (Animal anatomy): 0.072

TABLE 2 | Article "firmness" and its distribution of topics.

FERMETE, s. f. (*Gramm. & Littér.*) vient de *ferme*, & signifie autre chose que *solidité* & *dureté*. Une toile serrée, un sable battu, ont de la *fermeté* sans être durs ni solides. Il faut toûjours se souvenir que les modifications de l'ame ne peuvent s'exprimer que par des images physiques: on dit *la fermeté de l'ame, de l'esprit;* ce qui ne signifie pas plus *solidité* ou *dureté* qu'au propre. La *fermeté* est l'exercice du courage de l'esprit; elle suppose une résolution éclairée: l'opiniâtreté au contraire suppose de l'aveuglement. Ceux qui ont loüé la *fermeté* du style de Tacite, n'ont pas tant de tort que le prétend le P. Bouhours: c'est un terme hasardé, mais placé, qui exprime l'énergie & la force des pensées & du style. On peut dire que la Bruyere a un *style ferme*, & que d'autres écrivains n'ont qu'un style dur. *Article de M. de Voltaire*	Topic #227 (Morals): 0.224 Topic #118 (Eloquence): 0.209 Topic #265 (Properties of matter), 0.179 Topic #225 (Sensualism): 0.135

by its assigned "sacred criticism" label, as the fish topic belongs to the broader discourse on animal anatomy, and the Jewish topic to the discussion of the Ancients.

As a second example – using the article "Firmness" (*fermeté*) that belongs to the Grammar and Literature disciplines – the topic distribution once again captures the variety of what is actually discussed in the article by Voltaire (**Table 2**).

These two examples, among the many others we found, thus confirmed our initial intuition that topic modeling can be used as an effective tool for identifying the various discourses at play within individual articles. Next, we were eager to apply this approach on a larger scale: could we identify, for instance, the discursive makeup of whole classes of knowledge, or even of the entire *Encyclopédie*? Our first step in this direction was to examine the most highly prevalent topics across the whole text, those that occur in the greatest number of articles. Unsurprisingly, these are fairly unremarkable in terms of discursive content. For the most part, these large-scale topics all attest to the function of the *Encyclopédie* as a reference work, concerned, in the most general sense, with making comparisons and distinctions (topic #272, found in 11,106 articles), providing the meaning of words (topic #260, found in 10,120 articles), and appealing to the authority of the ancients (topic #65, found in 6,512 articles):

> Topic #272: différentes appelle usage lieu nombre différens sortes peuvent rapport seulement selon général non-premiere maniere espece particulier égard savoir doivent …
>
> Topic #260: terme signifie quelquefois sens chose appelle adj sert usage pris exprimer mots latin signifier entend désigner prend usité employé termes …
>
> Topic #65: chez romains usage anciens appelloit eux là donnoit quelquefois uns falloit gens seulement pouvoient grecs sortes selon parmi premiers lieu …

Once we move beyond these meta-discourses and the most prevalent topics, we can begin to identify specific topics that correspond more or less to the various disciplines treated in the *Encyclopédie*. Not surprisingly, the "chemistry" topic is found most in chemistry articles, the "botany" topic in botanical articles, mathematics in mathematics, etc. What interests us, however, are topics that are both distinct in nature – i.e., identifiable with a particular "discourse" or set of discursive practices – and that span multiple disciplinary boundaries. Mapping these discourses through the various classes and articles in which they are prevalent can thus lead to a greater understanding of the dialogical and discursive elements at play beneath the surface of the encyclopedic classification system.

Consider the class "Grammaire," for instance, which was known to be a sort of clearing house for Diderot in which to fit controversial material (Leca-Tsiomis, 1999). The topic we have identified with the discourse on natural rights ("droit naturel") is present in more than 60 grammar articles, almost double that of its own class of knowledge (see **Table 3**). Interestingly, what this topic distribution also demonstrates is how the discussion of natural rights in the *Encyclopédie* is primarily found in unclassified articles, a phenomenon that might be explained by the dangers of making the discourse around natural rights a prominent subject in the context of the Old Regime's strict censorship laws and absolutist politics.

Following again in d'Alembert's footsteps, we can think geographically about the distribution of the natural rights discourse throughout the *Encyclopédie* – a distribution that is quite independent of the existing discipline of "droit naturel" – and that underscores its truly inter- or trans-disciplinary breadth (see **Figure 2**).

If we use the above map – generated using the popular D3.js library – to move from the global to the local, from the general distribution of the disciplinary islands to the 61 articles that make up the particular island of Grammar, we can draw attention to articles that contain the natural rights discourse but that would have otherwise gone unnoticed. For example, our list of the most highly weighted articles for topic #56 includes the small, unsigned article "Inviolable" that belongs to the Grammar class and that has since been attributed to Diderot (Schwab et al., 1984). In it, along with the grammatical definition of the term, we find a usage example that reads: "La liberté de conscience est un privilege inviolable" (vol. 8, p. 864), which subtly places the freedom of thought

on the same plane as other "natural" and unalienable rights. We find a similar treatment of natural rights in the Grammar article "Supplanter," also unsigned in the original and later attributed to Diderot, which moves quickly from a properly grammatical consideration into a condemnation of tyranny as an unnatural state of governance (vol. 15, p. 671).

Other classes function in much the same way as Grammar, allowing the *philosophers* to smuggle controversial or even heretical opinions into articles with a seemingly non-ideological scope. By tracing the presence of these subversive discourses in an interdisciplinary manner, we can finally begin to uncover the various discursive and ideological practices in play over the entirety of the *Encyclopédie*. The discourse around morality, for instance, is found in no 240 articles from the geography class, both ancient and modern (see **Table 4**).

TABLE 3 | Most frequent classes in which topic #56 is found.

Topic	Top classes, number of articles
Topic #56 "droit naturel": *droit lois nature société loi hommes raison choses état homme justice naturel naturelle juste vie gens devoirs morale vertu souverain …*	Unclassified, 191 Grammaire, 61 Jurisprudence, 56 Morale, 55 Droit naturel, 30 Géographie moderne, 28 Théologie, 27 Géographie, 25 Droit politique, 23 Histoire moderne, 22

Here again (**Figure 3**), if we use our "maps" as a way of navigating through the different disciplinary islands toward the Ancient Geography articles, Diderot once more proves to be exemplary in his discursive acrobatics. While describing a tribe of ancient Thracians in the article "Dranses," for example, he quickly turns the discussion toward moral relativism – in a move the prefigures his later work, *Le Supplément au voyage de Bougainville* – with the assertion that: "It is not nature, it is tyranny that places on the heads of men a weight that causes them to moan and hate their condition" (vol. 5, p. 106).

A similar deployment of the discourse around "le culte religieux" – a subject on which the *encyclopédistes* were forced to tread lightly due to the strict censorship rules of the period – can be found in more than 100 articles drawn from Modern History (see **Table 5**).

Oriented toward the otherwise unremarkable Modern History article "Schooubiak" by the topic model map (**Figure 4**), we find another unsigned (but later attributed) article by Diderot, in which he describes an Islamic sect that practices a very unusual form of religious tolerance. This seeming incongruence with the accepted religious stereotypes of the time allows Diderot to raise the issue of religious intolerance by using the sect as a proxy for the *philosophers* themselves:

> Thus we see that if madness is in every land, reason also is in every land. Here we find men as much, or more obstinate of their religion as any other people on earth,

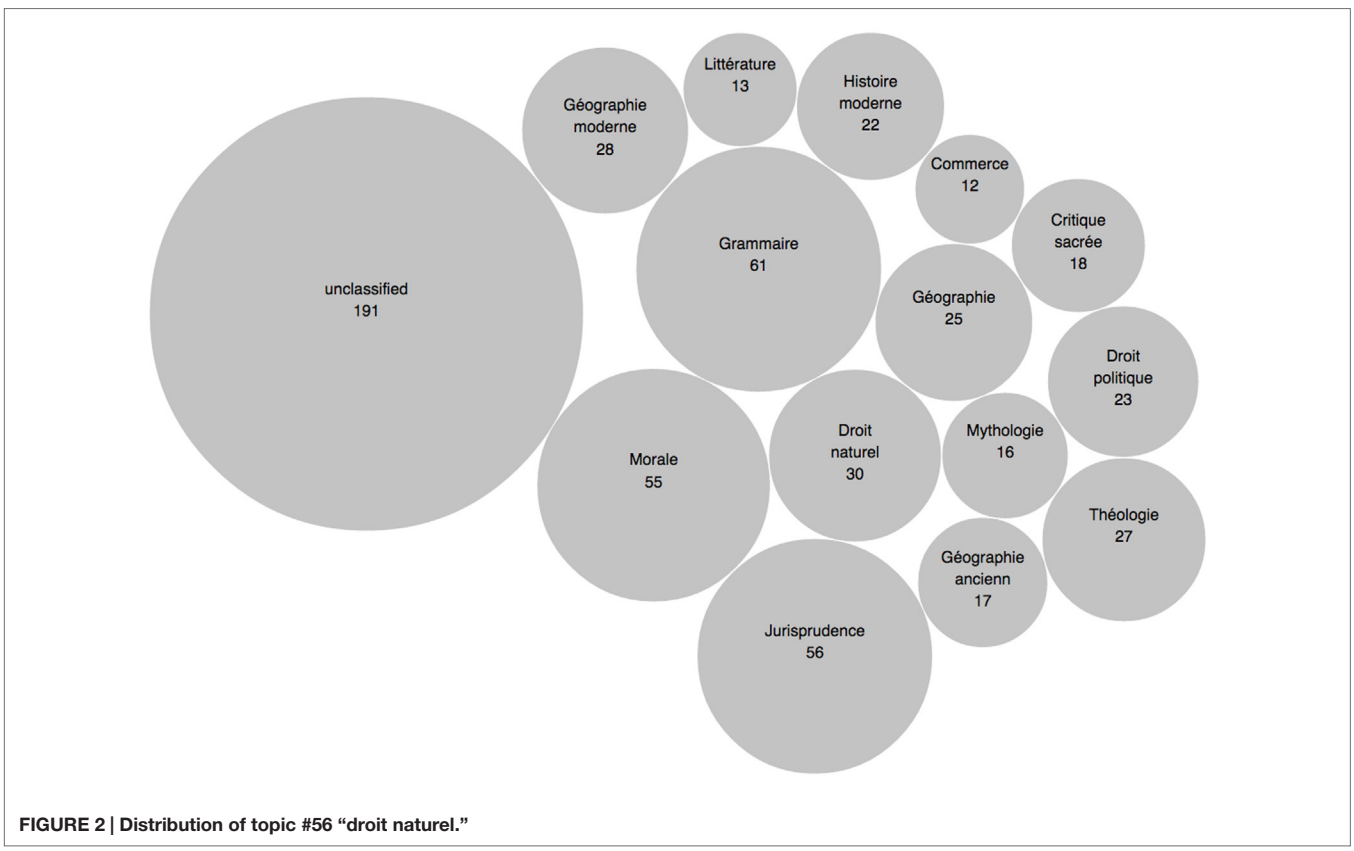

FIGURE 2 | Distribution of topic #56 "droit naturel."

> preaching tolerance to their peers; they are accused, of course, of disbelief, of indifference and atheism; they are forced to hide their doctrine; they are persecuted; and all because as priests are the same everywhere, it follows that tolerance should be everywhere hated (vol. 14, p. 778).

While these few examples represent well-known instances of subversive tactics on the part of Diderot and the other *encyclopédistes* [see Proust (1995), Leca-Tsiomis (1999), and Edelstein et al. (2013)], the novel manner in which we uncovered them is what we would like to stress here. By identifying discourses that occur in many disciplinary contexts – the discussion of natural rights in Jurisprudence and Grammar, for instance – we can finally move beyond the editors' original classification scheme and begin to draw out the multi-layered discursive practices that contribute to the rich dialogical texture of the *Encyclopédie*. The complexity of these practices, whether we find them using traditional methods of close reading or through the algorithmic processes described above, only serve to underscore the truly revolutionary nature of Diderot and d'Alembert's enterprise. Thanks to our vantage point atop the shoulders of these giants, to paraphrase Newton, we can now use digital methods to draw again on the rich multiplicity of points of view that d'Alembert situates in his various abstract systems of knowledge. To this end, machine-learning algorithms such as LDA can help us construct a host of new graphs, maps, and trees better to understand the *Encyclopédie*'s complex epistemology. These techniques may seem far afield from the original concerns of the eighteenth century *encyclopédistes*. But, as we have endeavored to demonstrate above, the mapping of knowledge and how this knowledge is communicated – either via traditional or computational methods – is a gesture wholly in keeping with "systematic spirit" of the Enlightenment.

TABLE 4 | Most frequent classes in which topic #227 is found.

Topic	Top classes, number of articles
Topic #227 "morale": *homme esprit hommes amour vertu morale notre caractere coeur ame mal moeurs raison passions bonheur société vice choses plaisir nos …*	Grammaire, 532 Unclassified, 414 Morale, 220 Géographie moderne, 146 Géographie ancienne, 94 Histoire moderne, 91 Mythologie, 90 Géographie, 83 Synonymes, 77 Jurisprudence, 67

FUTURE WORK

While exploring the paths along which our topic models led us, we came to the realization that we had not, perhaps, given enough

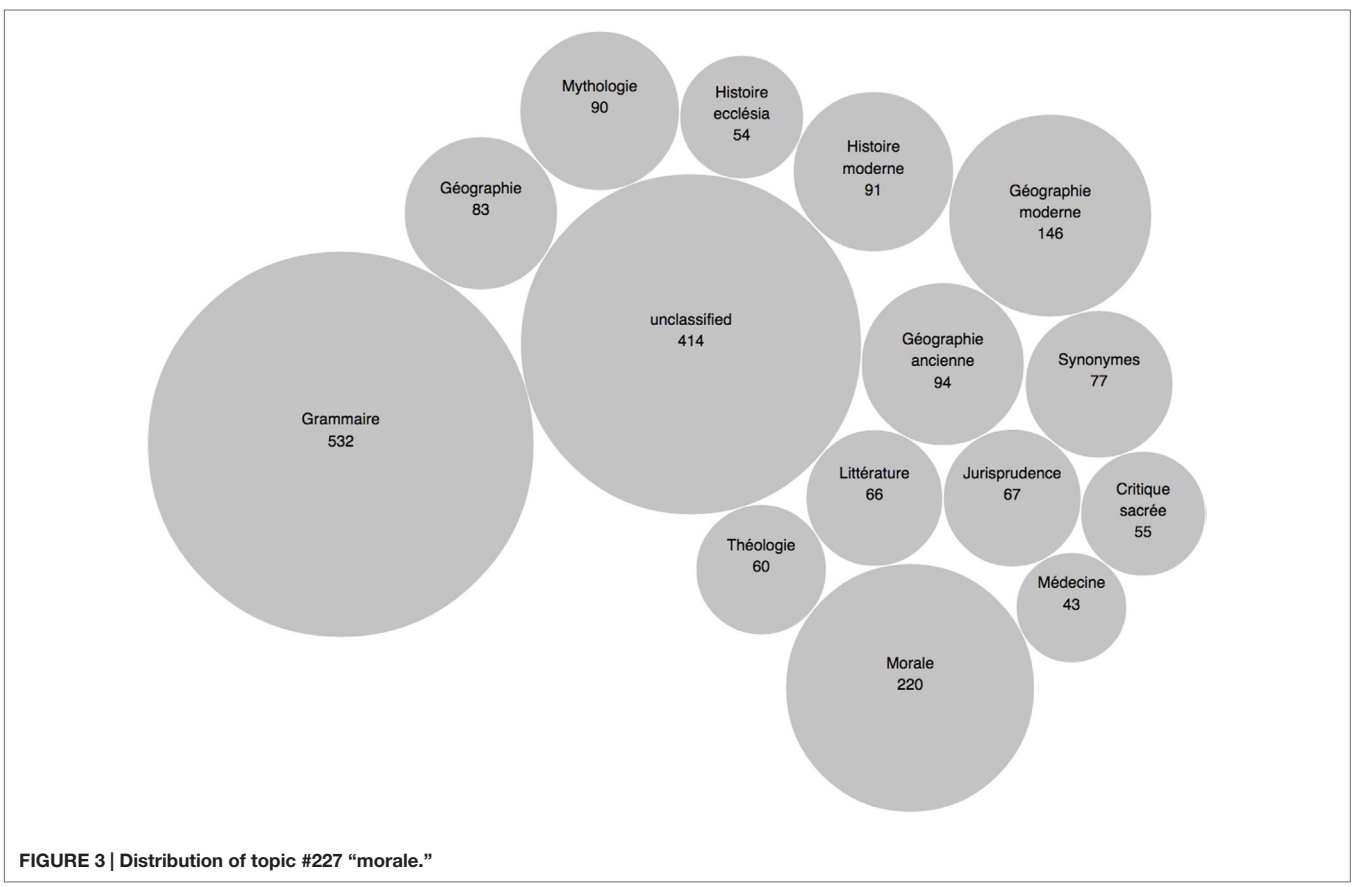

FIGURE 3 | Distribution of topic #227 "morale."

thought to the necessary pre-processing of our data. By this, we mean that in the future we should consider more carefully what constitutes a discourse from a semantic and morphological perspective before compiling our stopword lists, i.e., those words that should be eliminated before generating topic models. Here again, we can draw inspiration from Foucault, who, in *The Order of Things*, defined the *epistémè* as a framework of thought that both defines and determines the type of ideas that can be expressed at any given moment in time. The eighteenth century, which he identifies as part of the Classical age, was characterized by the importance it placed on names/nouns (unfortunately, the playful ambiguity between the homonyms *nom*/*nom* in French is lost in translation): "One might say that it is the Name that organizes all Classical discourse"; "the name is the end of discourse" [*le nom c'est le terme du discours*] (Foucault, 2001, pp. 129–30).

Indeed, as we have seen above in the ontology of the *Encyclopédie*, the construction and organization of knowledge begins with the naming (or labeling) of things, thus identifying and placing them alongside other objects (nouns/names) in a specific order. Classification, and to a larger degree speaking and thinking, are therefore discursive processes of nomination, in which the act of naming makes things be:

> The word designates, that is, in its very nature it is a noun or name. A proper noun, since it is directed always towards a particular representation, and towards no other. So, in contrast to the uniformity of the verb, which is never more than the universal expression of attribution, nouns proliferate in endless differentiation […] The generality of the noun is as necessary to the parts of discourse as is the designation of being to the form of the proposition (Foucault, 2001, p. 107).

Thus, by way of Foucault's insistence on the importance of names/nouns in discourse formation, and given our desire to uncover discourses in eighteenth century thought, we will experiment with keeping only the content words, such as nouns

TABLE 5 | Most frequent classes in which topic #242 is found.

Topic	Top classes, number of articles
Topic #242 "culte religieux": *religion dieu hommes culte dieux chrétiens ciel eux divinité christianisme terre monde esprit superstition vie homme payens doctrine mysteres opinions …*	Unclassified, 177 Théologie, 140 Histoire moderne, 110 Histoire ecclésiastique, 104 Grammaire, 89 Géographie moderne, 76 Mythologie, 75 Critique sacrée, 50 Géographie, 49 Géographie ancienne, 49

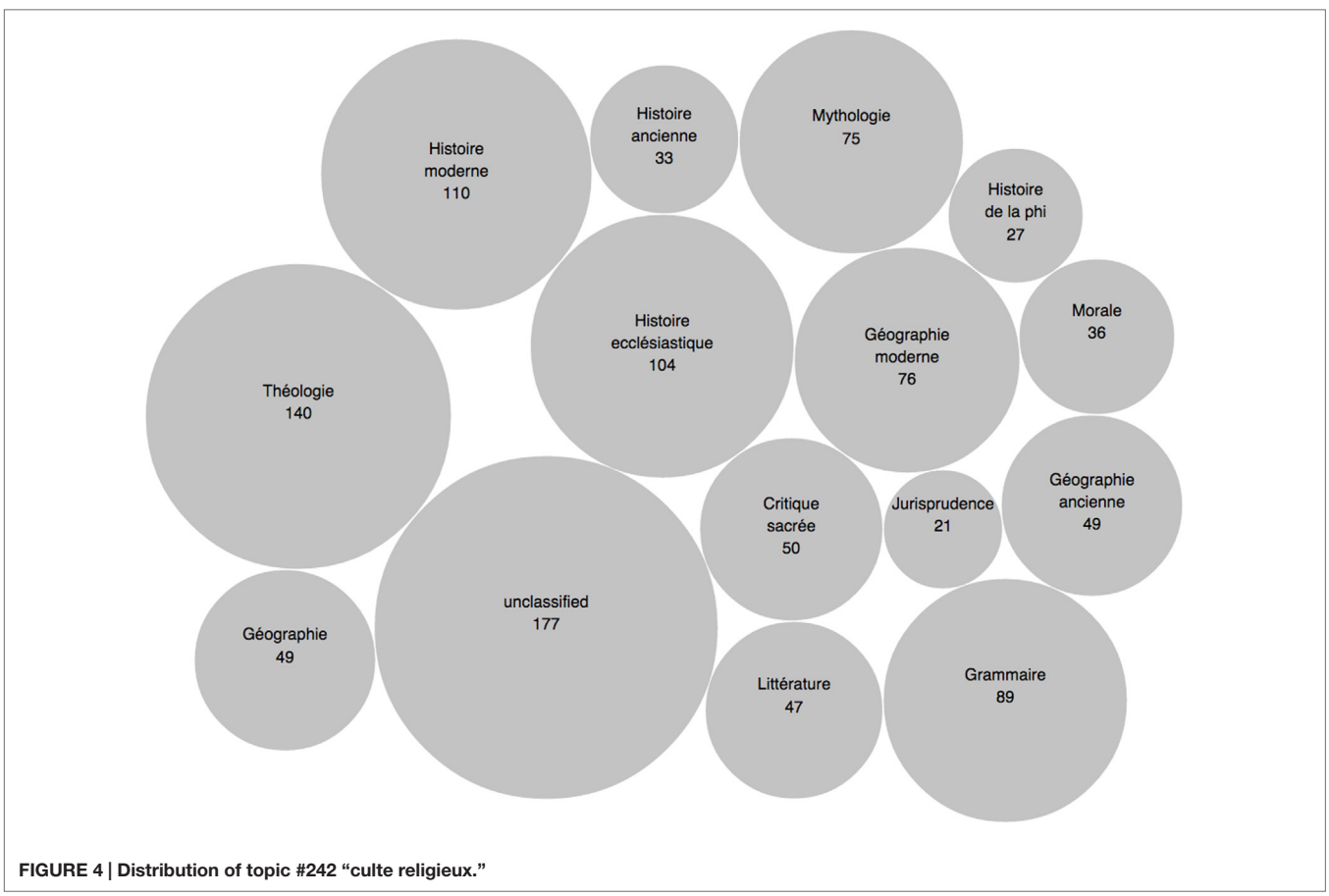

FIGURE 4 | Distribution of topic #242 "culte religieux."

and proper nouns, in our topic model. Jockers (2013), for one, has explored the notion of keeping only content words (even excluding proper nouns) for topic modeling, leading presumably to topics that are easier to parse and less "functional" form a grammatical perspective. This is not to say that we fully subscribe to Foucault's interpretation of the eighteenth century's "Classical" mode of discourse, but it would, if anything, allow us to evaluate his perspective and to generate more "readable" topics. It would also, we think, eliminate many of the more "functional" topics that we mention above – those that are, in fact, most prevalent in our current models – leading to topics that are more thematically consistent.

Furthermore, though we have focused, for practical reasons, exclusively on the *Encyclopédie* in this article, our broader interests lie in eighteenth century French literary culture as a whole. As such, we hope to use the *Encyclopédie* as a useful springboard from which to trace its various discourses in other contemporary texts and contexts. This move, in fact, builds upon our previous work applying the ontologies built from the *Encyclopédie*'s classification system to the *Journal de Trévoux*, an influential academic journal that appeared monthly in France between 1701 and 1782 (Horton et al., 2009). We would like to extend this work by identifying the presence of encyclopedic discourses in texts that both precede and follow the publication of the *Encyclopédie*. Here, our hope is that David Blei's concept of "Dynamic Topic Models" [see Blei and Lafferty (2006)], a derivative of LDA that tracks changes within topics across time, will allow us to look at the diachronic evolution of discourses identified in the *Encyclopédie*, and gain a better appreciation of their deployment over the long eighteenth century.

Finally, as we have always done in the past, we intend to make the results of our present and future experiments available online. As these are meant primarily to serve as research tools for the larger scholarly community, we will explore various interface paradigms in order to allow for an easy and intuitive exploration of our topic models and the discursive networks they underpin. These new "maps" of knowledge will not replace traditional methods of navigation in the *Encyclopédie* and other Enlightenment texts but will rather supplement them by providing multiple points of entry and a new transversal perspective on these already well-known datasets.

ACKNOWLEDGMENTS

The authors would like to acknowledge the generous financial and intellectual support provided by the University of Chicago's ARTFL Project over the duration of this research project.

REFERENCES

Allen, T., Douard, S., Cooney, C., Horton, R., Morrissey, R., Olsen, M., et al. (2010). Plundering philosophers: identifying sources of the Encyclopédie. *J. Assoc. Hist. Comput.* 13. Available at: http://quod.lib.umich.edu/j/jahc/3310410.0013.107/–plundering-philosophers-identifying-sources?rgn=main;view=fulltext

Andreev, L., Iverson, J., and Olsen, M. (1999). Re-engineering a war machine: ARTFL's Encyclopédie. *Lit. Ling. Comput.* 14: 11–28. doi:10.1093/llc/14.1.11

Baker, K. (1990). *Inventing the French Revolution: Essays on French Political Culture in the Eighteenth Century*. Cambridge: Cambridge University Press.

Bianco, J.-F. (2002). Diderot a-t-il inventé le web? *Recherches sur Diderot et sur l'Encyclopédie* 31-32: 15–25.

Blanchard, G., and Olsen, M. (2002). 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: 45–70. doi:10.4000/rde.122

Blei, D. (2012). Probabilistic topic models. *Commun. ACM* 55: 77–84. doi:10.1145/2133806.2133826

Blei, D. (2013). Topic modeling and digital humanities. *J. Digit. Humanit.* 2. Available at: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/

Blei, D., and Lafferty, J. (2006). Dynamic topic models. *Proc. Int. Conf. Mach. Learn.* 6: 113–20. doi:10.1145/1143844.1143859

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3: 993–1022. doi:10.1162/jmlr.2003.3.4-5.993

Brian, E. (1998). L'ancêtre de l'hypertexte. *Cahiers de Science & Vie* 47: 28–38.

Cassirer, E. (2009). *The Philosophy of the Enlightenment*. Princeton, NJ: Princeton University Press.

Cooney, C., Horton, R., Olsen, M., Roe, G., and Voyer, R. (2008). Hidden roads and twisted paths: intertextual discovery using clusters, classifications, and similarities. In *Digital Humanities 2008*, Edited by L. Opas-Hänninen, M. Jokelainen, I. Juuso, and T. Seppänen, 93–94. Oulu: University of Oulu Press.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20: 273–97. doi:10.1007/BF00994018

D'Alembert, J. (1751). Preliminary discourse. In *Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers, etc.*, Edited by D. Diderot and J. D'Alembert. University of Chicago; *ARTFL Encyclopédie Project (Spring 2013 Edition)*, Edited by R. Morrissey and G. Roe. Available at: http://encyclopedie.uchicago.edu/; Translations provided by The Encyclopedia of Diderot & d'Alembert Collaborative Translation Project, Translated by R. Schwab and W. Rex. Ann Arbor, MI: University of Michigan Library. Available at: http://hdl.handle.net/2027/spo.did2222.0001.083

de Bolla, P. (2013). *The Architecture of Concepts: The Historical Formation of Human Rights*. New York, NY: Fordham University Press.

Domingos, P., and Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.* 29: 103–37. doi:10.1023/A:1007413511361

Edelstein, D. (2009). *The Terror of Natural Right: Republicanism, the Cult of Nature, and the French Revolution*. Chicago, IL: University of Chicago Press.

Edelstein, D. (2010). *The Enlightenment: A Genealogy*. Chicago, IL: University of Chicago Press.

Edelstein, D., Morrissey, R., and Roe, G. (2013). To quote or not to quote: citation strategies in the Encyclopédie. *J. Hist. Ideas* 74: 213–36. doi:10.1353/jhi.2013.0012

Foucault, M. (1989). *The Archaeology of Knowledge*, Translated by Sheridan Smith, A. M. New York, NY: Routledge.

Foucault, M. (2001). *The Order of Things: An Archaeology of the Human Sciences*. New York, NY: Routledge.

Furet, F. (1978). *Penser la Révolution française*. Paris: Gallimard.

Guénard, F., Markovits, F., and Spallanzani, M. eds. (2006). *L'ordre des renvois dans l'Encyclopédie*. Paris: Corpus, revue de philosophie. 51.

Hadi, W., Thabtah, F., and Abdel-Jaber, H. (2007). A comparative study using vector space model with k-nearest neighbor on text categorization data. *World Congress Eng.* 2007: 296–301.

Helsloot, N., and Hak, T. (2001). La contribution de Michel Pêcheux à l'analyse de discours. *Langage et Société* 91: 5–33. doi:10.3917/ls.091.0005

Horton, R., Morrissey, R., Olsen, M., Roe, G., and Voyer, R. (2009). Mining eighteenth century ontologies: machine learning and knowledge classification in the Encyclopédie. *Digit. Humanit. Q.* 3. Available at: http://digitalhumanities.org:8080/dhq/vol/3/2/000044/000044.html

Hunt, L. (1984). *Politics, Culture and Class in the French Revolution*. Berkeley, CA: University of California Press.

Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana-Champaign, IL: University of Illinois Press.

Leca-Tsiomis, M. (1999). *Écrire l'Encyclopédie: Diderot: de l'usage des dictionnaires à la grammaire philosophique*. Oxford: Voltaire Foundation.

Lima, M. (2014). *The Book of Trees: Visualizing Branches of Knowledge*. Princeton, NJ: Princeton Architectural Press.

Melançon, B. (2004). Sommes-nous les premiers lecteurs de l'encyclopédie? In *Les défis de la publication sur le Web: hyperlectures, cybertextes et méta-éditions*, Edited by J.-M. Jean-Michel Salaün and C. Vandendorpe, 145–165. Lyon: Presses de l'ENSSIB.

Morrissey, R., Iverson, J., and Olsen, M. (2001). Présentation: L'Encyclopédie electronique. In *L'Encyclopédie du réseau au livre et du livre au réseau*, Edited by R. Morrissey and P. Roger, 17–27. Paris: Champion.

Newman, D., and Block, S. (2006). Probabilistic topic decomposition of an eighteenth-century American newspaper. *J. Am. Soc. Inform. Sci. Technol.* 57: 753–67. doi:10.1002/asi.20342

Newman, D., Lau, J.H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*. Los Angeles, California, 100–108.

Noh, Y., Hagedorn, K., and Newman, D. (2011). Are learned topics more useful than subject headings? In *JCDL '11 Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (New York, NY: ACM), 411–412.

Pêcheux, M. (1969). *Analyse automatique du discours*. Paris: Dunod.

Pêcheux, M. (1982). *Language, Semantics and Ideology*. New York, NY: St. Martin's Press.

Proust, J. (1995). *Diderot et l'Encyclopédie*. Paris: Albin Michel.

Ramirez, E.H., Brena, R., Magatti, D., and Stella, F. (2012). Topic model validation. *Neurocomputing* 76: 125–33. doi:10.1016/j.neucom.2011.04.032

Ricœur, P. (1975). *La métaphore vive*. Paris: Seuil.

Roe, G., Cooney, C., Horton, R., Olsen, M., and Voyer, R. (2008). Re-engineering the tree of knowledge: vector space analysis and centroid-based clustering in the Encyclopédie. In *Digital Humanities 2008*, Edited by L. Opas-Hänninen, M. Jokelainen, I. Juuso, and T. Seppänen, 179–180. Oulu: University of Oulu Press.

Salton, G., and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inform. Process. Manag.* 24: 513–23. doi:10.1016/0306-4573(88)90021-0

Schmidt, B. (2012). When you have a MALLET, everything looks like a nail. *Blog Post*, November 2. Available at: http://sappingattention.blogspot.com.au/2012/11/when-you-have-mallet-everything-looks.html

Schmidt, B. (2013). Keeping the words in topic models. *Blog Post*, January 9. Available at: http://sappingattention.blogspot.com.au/2013/01/keeping-words-in-topic-models.html

Schwab, R., Rex, W., and Lough, J. (1984). *Inventory of Diderot's Encyclopédie*. Oxford: Studies on Voltaire and the Eighteenth Century.

Wallach, H., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th International Conference on Machine Learning (ICML)* (New York, NY: ACM).

Williams, G. (1999). *French Discourse Analysis: The Method of Post-Structuralism*. New York, NY: Routledge.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.