Encyclopaedic Discourses Revisited: Integrating Topic Models in the 18th Century

Glenn Roe[*] and Clovis Gladstone[+]
[*]Centre for Digital Humanities Research, Australian National University
[+]Textual Optics Lab, University of Chicago

This paper describes attempts at integrating topic models of the 18th-century *Encyclopédie* of Diderot and d'Alembert as heuristic tools for classifying other text collections from the same period. An unsupervised machine learning technique commonly used in the digital humanities,[1] we are using topic modelling here to identify 'discourses' in the *Encyclopédie* that can be traced across various text types. We will thus discuss hyper-parameterization settings and their effects on topic model generation, methods for stopword generation, and finally, the inference of our topic models over Voltaire's universal history, the *Essai sur les mœurs et l'esprit des nations*, as well as the more than 20,000 letters found in Voltaire's correspondence.[2]

Preliminary experiments with large-scale correspondence collections have proven both promising and somewhat frustrating when it comes to topic modelling. Letters, by their very nature, should be almost ideal candidates for topic-modelling classification tasks given the size of the documents and the wide array of topics discussed therein. However, topics are often too general in nature (i.e., concerned with the act of letter-writing itself), or conversely, too limited in terms of overall topic distribution. These limitations have led us to revisit earlier work using machine learning approaches to explore the classification system of the *Encyclopédie*. Specifically, we aim to leverage the ontology of the *Encyclopédie*, with its large range of discourses and disciplines, in order to infer topics and topic distributions first on the *Essai sur les mœurs*, whose contents are fairly well understood, and then subsequently on Voltaire's correspondence, a collection that has never been sufficiently indexed.

In previous work using topic modelling to identify 'discourses' in the *Encyclopédie*,[3] we came to the realization that we had not given enough thought to the necessary pre-processing of our data. For this paper, we wanted to consider more carefully what constitutes a discourse from a semantic and morphological perspective before compiling our stopword list and hyperparameters. Here, as before, we are drawing inspiration from Michel Foucault's work on discourse analysis. Following Foucault, the 18th century was the last century to belong to the 'Classical' epistemic paradigm, which was characterized by the importance it placed on names or nouns: "One might say that it is the Name/Noun that organizes all Classical discourse".[4] Thus, by way of Foucault's insistence on the importance of names/nouns in discourse formation, and given our desire to uncover 'encyclopaedic' discourses in other 18th-century

[1] See, for instance, the special issue of the *Journal of Digital Humanities* (vol. 2, no. 1) edited by Scott Weingart and Elijah Meeks: http://journalofdigitalhumanities.org/2-1/.

[2] We would like to thank the Voltaire Foundation at the University Oxford and the ARTFL Project at the University of Chicago for making these collections available to us.

[3] See Glenn Roe, Clovis Gladstone and Robert Morrissey, "Discourses and Disciplines in the Enlightenment: Topic Modeling the French Encyclopédie", *Frontiers in Digital Humanities* 2.8 (January 2016) [https://www.frontiersin.org/articles/10.3389/fdigh.2015.00008].

[4] Michel Foucault, T*he Order of Things: An Archaeology of the Human Sciences* (New York: Routledge, 2001): 129-30.

collections, we experimented with keeping only the content words, such as nouns and proper nouns,[5] when training our topic model. The vastly reduced vocabulary was then fed into the Scikit-Learn[6] implementation of the Latent Dirichlet Allocation (LDA) topic modelling algorithm, in order to generate general topics akin to Foucauldian discourses.

We also decided to dig deeper into LDA's hyperparameters given their importance in any machine learning approach.[7] There are two important values that can be configured prior to the actual training phase of LDA, commonly known as *alpha*, and *beta*. Both hyperparameters determine the resulting sparsity of the topic and word distributions produced by the algorithm. This was very important in our case, since the *Encyclopédie* discusses a wide range of topics, and hence contains a very diverse vocabulary. By choosing lower *alpha* and *beta* values, we were able to produce highly coherent word clusters, and a very sparse topic distribution for our training corpus.

Settling on the right hyperparameters for our data was only a first step however, as we also wanted to generate topics that could be interpreted as discourses for the analysis of other texts. While in previous work we used topic models with over 250 topics, we opted this time to use a 100-topic model, which yielded more abstract yet very coherent word clusters. This abstraction is very much in line with Foucault's notion of discourses, which describe semantic units far broader than topics or disciplines. After labelling this model, we proceeded to apply it to Voltaire's expansive *Essai sur les mœurs* in a first instance and then, finally, to his letters in order to obtain their individual topic distributions (see Annexe).

Preliminary results demonstrate that topic inference of the *Essai* is far more fruitful than that of the letters, which is perhaps not surprising given that it covers topics covered throughout the *Encyclopédie*: geography, history, religion, etc. The key challenge we face will be translating this method to letter classification, which represents a much sparser text-space. In terms of discourse analysis, we can already see the importance of the discourse of sociability, which is used throughout our text collections, and overwhelmingly in the letters. This is not that surprising as Voltaire is describing human societies in the *Essai*, and the letters are, by definition, social texts. These results raise the question of what topic model integration brings to the table as a distant reading tool. In the case of the *Encyclopédie*, we previously found that it allowed us to uncover subversive discourses in seemingly innocuous articles, and as such, very much functioned as a useful discovery tool. In the case of the letters, and perhaps even the *Essai*, we find that topic modelling does give a good overview of discursive content, but may not provide significant new insights into the collections themselves.

[5] We used the Averaged Perceptron part-of-speech tagger from the Spacy library (https://spacy.io/) to identify and extract all nouns from the *Encyclopédie*.

[6] http://scikit-learn.org/stable/.

[7] The importance of hyperparamater tuning is widely recognized within the machine-learning community, and can have a considerable impact on results. For instance, see Hutter, F., Hoos, H. & Leyton-Brown, K.. (2014). An Efficient Approach for Assessing Hyperparameter Importance. Proceedings of the 31st International Conference on Machine Learning, in PMLR 32(1):754-762.

Annexe: Topic distributions of the top 5 topics across our 3 datasets.

Topic #	Label	Top topic	Top 3 topics
65	Géographie	5653	10664
38	Outils de métier	2225	4215
62	Géographie humaine	2045	4883
42	Botanie	1653	3587
56	Anatomie animale	1624	3045

Table 1. *Encyclopédie* topic distributions (74k articles, 23 million words)

Topic #	Label	Top topic	Top 3 topics
69	Histoire politique	74	162
8	Royauté	31	73
57	Église catholique	24	52
62	Géographie humaine	19	59
94	Sociabilité	15	78

Table 2. *Essai sur les mœurs* topic distributions (197 chapters, 540k words)

Topic #	Label	Top topic	Top 3 topics
94	Sociabilité	11600	14490
0	Lettres	1650	5550
8	Royauté	842	4869
18	Chronologie	333	3119
94	Goût artistique	308	1945

Table 3. Voltaire correspondence topic distributions (21k letters, 11 million words)