**Offre de stage M2 / PFE**

# Classification automatique des domaines de connaissance d'entrées lexicographiques

---

## Encadrants


- Ludovic Moncla, LIRIS - INSA Lyon (https://ludovicmoncla.github.io)
- Julien Velcin, ERIC - Université Lumière Lyon 2 (https://eric.univ-lyon2.fr/jvelcin/)


## Contexte et objectifs

Ce stage s'inscrit dans un projet interdisciplinaire dont l'objectif consiste à conduire des études exploratoires en traitement automatique de données lexicographiques extraites du Dictionnaire Universel François-Latin de Trévoux (DUFLT). Dans ce contexte, le travail de stage s'intéressera à l'expérimentation de méthodes d'apprentissage automatique pour l'entraînement de modèles de classification afin d'identifier automatiquement les domaines de connaissance dans les articles du DUFLT. De manière générale, nous souhaitons dresser une liste des domaines et sous-domaines de connaissances mentionnés dans chacune des éditions du corpus, afin de quantifier et de comparer la place qu'ils occupent. Cela permettra de mettre en évidence l'évolution qualitative et quantitative de ces domaines dans la série DUFLT entre 1704 et 1771. Dans le cadre du stage, l'expérimentation portera sur les éditions de 1743 et 1771 que nous avons au format numérique.

Le ou la stagiaire devra s'appuyer sur les récentes avancées en intelligence artificielle et en TAL pour proposer des solutions pour la classification des textes. Nous nous intéresserons en particulier aux approches neuronales pour la modélisation thématique et aux plongements de mots (ainsi que d'unités plus grandes : phrases, alinéas, articles) pour la modélisation et la spécialisation de modèles de langues. Le volume limité et la segmentation temporelle d'un corpus historique en ancien français rendra difficile l'utilisation pure et simple des modèles pré-entraînés sur des données modernes comme CamemBERT, FlauBERT, BARthez. Un premier objectif sera alors d'évaluer les performances de ces modèles de langues pour la tâche de classification supervisée et de comparer les résultats entre les deux éditions du corpus. Pour cette tâche, le ou la stagiaire pourra s'appuyer sur nos premiers résultats obtenus dans le cadre du projet GEODE [1] sur l'Encyclopédie de Diderot et d'Alembert [Brenon et al., 2022]. Pour réaliser cette tâche, il sera au préalable nécessaire de construire le jeu de données labellisé à partir des marqueurs de domaines identifiés au sein des documents. La présence des marqueurs (labels) n'est pas systématique dans la version structurée de nos données et peut être implicite dans le contenu des articles. Il sera ainsi nécessaire de développer une méthode pour enrichir la classification du jeu de données en s'appuyant sur des ressources telle qu'une liste de marqueurs de domaines constituée par des experts de ce dictionnaire. Cette étape pourra faire appel aux modèles de langue et à des calculs de similarité sémantique pour annoter les entrées non classées du dictionnaire et constituer les jeux d'entraînement et d'évaluation. En complément, nous nous intéresserons à la comparaison des modèles entraînés indépendamment sur les deux éditions du corpus et sur l'analyse de l'évolution des domaines d'une édition à l'autre. Ces analyses pourront amener au développement de solutions adaptées pour intégrer la dimension diachronique du corpus et permettre l'étude de l'évolution de la sémantique et stylistique en fonction de leurs contextes [Christophe et al., 2021, Terreau et al., 2021]. L'enrichissement des données devrait nous aider à construire des espaces de représentation du corpus, possiblement à plusieurs niveaux de granularité (phrase, alinéa, article), qui pourront nous aider à mieux identifier les domaines. Pour

1. https://geode-project.github.io

cela, étant donné les volumes dont nous disposons et du côté fluctuant des catégories, des approches non ou peu supervisées, comme le clustering par contrainte pourront être explorer.

Ce stage sera réalisé dans l'équipe DM2L (Data Mining et Machine Learning) du laboratoire LIRIS [2] et l'équipe DMD (Data Mining & Decision) du laboratoire ERIC [3] et en collaboration avec le laboratoire ICAR [4].

## Candidatures

Des compétences sont attendues en programmation et en science des données (Machine Learning et Deep Learning). Des connaissances en traitement automatique de la langue (TAL) seront appréciées.

**Profil recherché** Master 2 Informatique

**Lieu du stage** Laboratoire ERIC, Université Lyon 2, Bron (principalement), avec des visites au laboratoire LIRIS, INSA Lyon, Campus La Doua, Villeurbanne.

**Période de stage** 5 à 6 mois entre février et juillet 2024

**Candidature** Envoyer un mail présentant votre parcours et vos motivations, votre CV et vos derniers relevés de notes à : ludovic.moncla@insa-lyon.fr et julien.velcin@univ-lyon2.fr Date limite de candidature : **20 décembre 2023**. Les candidatures seront examinées au fil de l'eau.

# Références


[Brenon et al., 2022] Brenon, A., Moncla, L., and McDonough, K. (2022). Classifying encyclopedia articles : Comparing machine and deep learning methods and exploring their predictions. *Data & Knowledge Engineering*, 142 :102098.

[Christophe et al., 2021] Christophe, C., Velcin, J., Cugliari, J., Boumghar, M., and Suignard, P. (2021). Monitoring geometrical properties of word embeddings for detecting the emergence of new topics. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 994–1003, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Terreau et al., 2021] Terreau, E., Gourru, A., and Velcin, J. (2021). Writing style author embedding evaluation. In Gao, Y., Eger, S., Zhao, W., Lertvittayakumjorn, P., and Fomicheva, M., editors, *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 84–93, Punta Cana, Dominican Republic. Association for Computational Linguistics.


2. https://liris.cnrs.fr
3. https://eric.msh-lse.fr/
4. http://icar.cnrs.fr

**Master's Internship Offer**

# Automatic Classification of Knowledge Domains for Lexicographic Entries

## Supervisors


- Ludovic Moncla, LIRIS - INSA Lyon (https://ludovicmoncla.github.io)
- Julien Velcin, ERIC - Université Lumière Lyon 2 (https://eric.univ-lyon2.fr/jvelcin/)


## Context and objectives

This internship is part of an interdisciplinary project with the aim of conducting exploratory studies in the automatic processing of lexicographic data extracted from the Dictionnaire Universel François-Latin de Trévoux (DUFLT). In this context, the internship work will focus on experimenting with machine learning methods for training classification models to automatically identify knowledge domains in articles from DUFLT. In general, we aim to compile a list of knowledge domains and sub-domains mentioned in each edition of the corpus to quantify and compare their prominence. This will highlight the qualitative and quantitative evolution of these domains in the DUFLT series between 1704 and 1771. The internship will specifically focus on the 1743 and 1771 editions available in digital format.

The intern will need to leverage recent advances in artificial intelligence and natural language processing to propose solutions for text classification. We will focus particularly on neural approaches for thematic modeling and word embeddings (as well as larger units : sentences, paragraphs, articles) for modeling and specializing language models. The limited volume and temporal segmentation of a historical corpus in Old French will make the straightforward use of pre-trained models on modern data, such as CamemBERT, FlauBERT, BARthez, difficult. A primary objective will be to evaluate the performance of these language models for the task of supervised classification and compare the results between the two editions of the corpus. For this task, the intern can rely on our initial results obtained in the context of the GEODE project [5] on Diderot and d'Alembert's Encyclopedia [Brenon et al., 2022]. To accomplish this task, it will be necessary to construct the labeled dataset from domain markers identified within the documents. The presence of markers (labels) is not systematic in the structured version of our data and may be implicit in the content of the articles. Thus, it will be necessary to develop a method to enrich the classification of the dataset by relying on resources such as a list of domain markers compiled by experts in this dictionary. This step may involve language models and semantic similarity calculations to annotate uncategorized entries in the dictionary and create training and evaluation sets. Additionally, we will explore the comparison of models trained independently on the two editions of the corpus and analyze the evolution of domains from one edition to another. These analyses may lead to the development of adapted solutions to incorporate the diachronic dimension of the corpus and enable the study of semantic and stylistic evolution based on their contexts [Christophe et al., 2021, Terreau et al., 2021]. Data enrichment should help us build representation spaces for the corpus, possibly at several levels of granularity (sentence, paragraph, article), which can aid in better identifying domains. Given the volumes at our disposal and the fluctuating nature of categories, non- or minimally supervised approaches, such as constraint-based clustering, may be explored.

This internshipwill be carried out in the DM2L (Data Mining and Machine Learning) team of the LIRIS laboratory and the DMD (Data Mining & Decision) team of the ERIC laboratory, in collaboration with the ICAR laboratory.

5. https://geode-project.github.io

## Applications

Skills in programming and data science (Machine Learning and Deep Learning) are expected. Knowledge in Natural Language Processing (NLP) would be appreciated.

**Profile Sought** Master 2 Computer Science

**Internship Location** ERIC laboratory, Université Lyon 2, Bron (mainly), with visits at the LIRIS Laboratory, INSA Lyon, Campus La Doua, Villeurbanne.

**Internship Duration** 5 to 6 months between February and July 2024.

**Application** Send an email presenting your background and motivation, your CV, and your most recent transcripts to : ludovic.moncla@insa-lyon.fr and julien.velcin@univ-lyon2.fr Application deadline : **December 20, 2023**. Applications will be reviewed on a rolling basis.

# Références


[Brenon et al., 2022] Brenon, A., Moncla, L., and McDonough, K. (2022). Classifying encyclopedia articles : Comparing machine and deep learning methods and exploring their predictions. *Data & Knowledge Engineering*, 142 :102098.

[Christophe et al., 2021] Christophe, C., Velcin, J., Cugliari, J., Boumghar, M., and Suignard, P. (2021). Monitoring geometrical properties of word embeddings for detecting the emergence of new topics. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 994–1003, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Terreau et al., 2021] Terreau, E., Gourru, A., and Velcin, J. (2021). Writing style author embedding evaluation. In Gao, Y., Eger, S., Zhao, W., Lertvittayakumjorn, P., and Fomicheva, M., editors, *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 84–93, Punta Cana, Dominican Republic. Association for Computational Linguistics.