Internship position at CNAM Paris on Machine Learning for complex documents analysis

Host ISID team at Centre d'études et de recherche en informatique et communications (CEDRIC), CNAM Paris

Starting date From February 2021

Duration 6 months

Gratification around 600 € / month

* **Candidate profile***
As a minimum requirement, the successful candidate should have:

- A master degree in one or more of the following areas: machine learning, natural language processing, symbolic AI, semantic web.
- Excellent programming skills (Java or Python)
- Excellent command of English
- Experience with machine learning and natural language processing

How to apply
The application should be formatted as **a single pdf file** and should include:

- A complete and detailed curriculum vitae
- A cover letter
- a report you wrote (ideally an internship or project report)
- The content of M1 and M2 courses and the corresponding grades
- The contact of two referees and a recommandation letter if possible

The pdf file should be sent to nada.mimouni@cnam.fr

Keywords machine learning, natural language processing, semantic web

Supervision Nada Mimouni (Cnam), Fayçal Hamdi (Cnam), Thomas Francart (Sparna)

Context
A collection of documents in specific domains is characterized by the abundance and variety of links between the documents where a document cannot be interpreted without its context i.e. the ones to which it is connected.
Besides the rich semantic information contained in each document, analyzing a collection of interlinked documents requires taking the network dimension into consideration in the analysis process in order to have the full picture.
In the last years, with the expansion of machine learning and deep learning models, a lot of algorithms have been used with such kind of data in order handle the complexity related to its double dimension: semantic and interlinks. These algorithms were applied into a variety of application such as information retrieval, navigation and exploration, automatic summary, topics extraction, recommendation, etc. Depending on the type of the document, the targeted application, one algorithm is used.

Objective

The objective of this internship is to identify, through a broad study of the state of the art, which algorithm is used for which application and on which type of data (language, structure, etc.). A comparison will be performed between the identified algorithms on the basis of their performances regarding two characteristics of the used data (mainly the language and the structure).

The aim is to be able to select, for a given application, the most appropriate algorithm or model according to the studied data.

References

Mimouni, N., Nazarenko, A. and Salotti, S. Answering Complex Queries on Legal Networks: A Direct and a Structured IR Approaches. In AI Approaches to the Complexity of Legal Systems. AICOL 2015, 2016, 2017, Lecture Notes in Computer Science, vol 10791. Springer., pages 451-464, 2018.

Mimouni, N., Nazarenko, A., Paul, E. and Salotti, S. Towards Graph-based and Semantic Search In Legal Information Access Systems. In Twenty-Seventh Annual Conference on Legal Knowledge and Information Systems (JURIX 2014), pages 163-168, Krakow, Poland, 2014.

Alschner, Wolfgang, AI and Legal Analytics (November 2, 2020). in Florian Martin-Bariteau & Teresa Scassa, eds., Artificial Intelligence and the Law in Canada (Toronto: LexisNexis Canada, 2021), Available at SSRN: https://ssrn.com/abstract=3733957.

Tarasconi, F., Botros, M., Caserio, M., Sportelli, G., Giacalone, G., Uttini, C., Vignati, L., and Zanetta, F. (2020). Natural Language Processing Applications in Case-Law Text Publishing. JURIX.

Chalkidis, I., Kampas, D. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artificial Intelligence and Law 27, 171–198 (2019). https://doi.org/10.1007/s10506-018-9238-9.

Francesconi, Enrico and Küster, Marc and Gratz, Patrick and Thelen, Sebastian. (2015). The Ontology-Based Approach of the Publications Office of the EU for Document Accessibility and Open Data Services. 9265. 29-39. 10.1007/978-3-319-22389-6_3.

Leone V, Di Caro L, Villata S. Taking stock of legal ontologies: a feature-based comparative analysis. Artificial Intelligence and Law. 2019 Jun 13:1-29.

Francart T, Dann J, Pappalardo R, Malagon C, Pellegrino M. The European Legislation Identifier. Knowledge of the Law in the Big Data Age. 2019 Jul 23;317:137-48.

Bibal, A., Lognoul, M., de Streel, A. et al. Legal requirements on explainability in machine learning. Artificial Intelligence and Law (2020). https://doi.org/10.1007/s10506-020-09270-4.

B. Waltl, G. Bonczek, E. Scepankova and F. Matthes, Semantic types of legal norms in German laws: classification and analysis using local linear explanations. Artificial Intelligence and Law. 2019 27 (1), 43-71.

Agnoloni, Tommaso, Lorenzo Bacci, Ginevra Peruginelli, Marc van Opijnen, Jos van den Oever, Monica Palmirani, Luca Cervone et al. "Linking European Case Law: BO-ECLI Parser, an Open Framework for the Automatic Extraction of Legal Links." In Jurix, pp. 113-118. 2017.

Medvedeva, M., Vols, M. and Wieling, M. Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law 28, 237–266 (2020). https://doi.org/10.1007/s10506-019-09255-y.