Apprentissage de représentations pour la recherche d'information dans les tableaux – H/F

Stage de 6 mois @ EDF R&D Saclay (91120)
Début du stage : mars 2026

Contexte

La R&D d'EDF (2000 chercheurs) a pour missions principales de contribuer à l'amélioration de la performance des unités opérationnelles du groupe EDF ainsi que d'identifier et de préparer les relais de croissance à moyen et long terme. Dans ce cadre, le département Services, Economie, Outils Innovants et IA (SEQUOIA) est un département pluridisciplinaire (sciences de l'ingénieur, sciences humaines et sociales) qui fournit un appui à l'élaboration et au portage des offres, des services et des outils de relation client aux directions opérationnelles du groupe EDF.

Au sein de ce département, ce stage sera rattaché au groupe « Statistiques et Outils d'Aide à la Décision » (SOAD) : cette équipe compte une vingtaine d'ingénieurs chercheurs spécialisés en IA et data science avec des compétences fortes autour du machine learning et du deep learning, du web sémantique, de l'IA symbolique et de l'IA générative (texte, voix, image, multimodalité…), en particulier du NLP (LLM, RAG, data mining…).

Objectifs

La recherche d'information dans les données textuelles a bénéficié des avancées récentes en traitement automatique des langues (NLP). La recherche vectorielle, où les documents sont segmentés en *chunks* et encodés à l'aide d'un modèle de langue, est devenue un standard dans les communautés scientifique et technique (Reimers et Gurevych, 2019).

Cependant, ces approches proposent rarement un traitement spécifique pour la gestion des tableaux. Or, ceux-ci sont omniprésents dans les documents techniques industriels, comme c'est le cas à EDF. L'approche vectorielle classique échoue à représenter correctement les informations qu'ils contiennent pour deux raisons principales :

- **Perte de contexte** : Les valeurs des cellules ne sont compréhensibles qu'en lien avec les en-têtes de colonnes et de lignes. Le chunking standard découpe souvent les tableaux, brisant cette association.
- **Distance physique** : Les en-têtes de tableaux sont parfois très éloignés des valeurs correspondantes, rendant la compréhension de la structure tabulaire difficile pour un modèle de langage (LLM).

Plusieurs travaux de recherche récents s'intéressent à la construction de représentations vectorielles de tableaux en utilisant des méthodes de Deep Learning. Kim et al. (2024), par exemple, présentent un modèle de fondation pour les données tabulaires qui, en combinant représentation orientée graphe et un entraînement auto-supervisé, peut être spécialisé sur des tâches précises (e.g. régression, classification).

L'accès efficace à l'information est un enjeu stratégique pour EDF. Notre système d'informations héberge plusieurs millions de documents techniques contenant de nombreux tableaux. Optimiser la recherche d'information au sein de ce vaste corpus est crucial pour le travail quotidien de nos ingénieurs.

L'objectif de ce stage est d'explorer et d'implémenter des méthodes avancées de création de représentations pour les données tabulaires, spécifiquement adaptées à notre corpus technique dans le domaine du nucléaire.

Encadré(e) par nos ingénieurs-chercheurs, vous aurez pour missions principales de :

- **Réaliser un état de l'art** approfondi sur les méthodes d'apprentissage de représentations pour les données tabulaires.
- **Sélectionner et implémenter** les approches les plus prometteuses pour notre cas d'usage industriel.
- **Développer le pipeline d'entraînement** des modèles sélectionnés.
- **Concevoir et mettre en œuvre** un protocole d'évaluation rigoureux pour mesurer la performance des modèles entraînés sur des tâches de recherche d'information réelles.
- **Contribuer à la valorisation des travaux** (documentation, rapport technique, présentation interne).

Ce stage est une immersion complète en R&D, couvrant l'ensemble du cycle de vie d'un projet de Deep Learning, de la recherche bibliographique à l'évaluation en contexte industriel.

Le/la stagiaire développera une triple compétence pointue :

- **Expertise Technique en NLP (Deep Learning)** : Maîtrise des architectures SOTA (Transformers) et des méthodes d'apprentissage de représentations pour les données tabulaires.
- **Compétence MLOps (Calcul Intensif)** : Capacité à opérer dans un environnement de calcul haute performance (HPC), incluant la gestion de jobs d'entraînement distribué sur plusieurs dizaines de GPU.
- **Compétence Méthodologique** : Capacité à mener un état de l'art sur un sujet de recherche avancé, à proposer et tester des solutions innovantes, et à appliquer une méthodologie scientifique rigoureuse pour répondre à un enjeu industriel critique (la recherche d'information dans les données tabulaires en domaine spécialisé).

Références

Kim, Myung Jun, Léo Grinsztajn, et Gaël Varoquaux. « CARTE: pretraining and transfer for tabular learning ». Proceedings of the 41st International Conference on Machine Learning (Vienna, Austria), ICML'24, vol. 235 (juillet 2024): 23843-66.

Reimers, Nils, et Iryna Gurevych. « Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks ». In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), édité par Kentaro Inui, Jing Jiang, Vincent Ng, et Xiaojun Wan. Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/D19-1410.

Profil recherché

- Étudiant(e) en dernière année d'École d'Ingénieur ou en Master 2 Recherche, avec une spécialisation en Data Science, Machine Learning ou Traitement Automatique du Langage (NLP).
- Solide maîtrise de Python et des librairies de Deep Learning (PyTorch).
- Bonnes connaissances théoriques en Machine Learning et Deep Learning (notamment les Transformers).
- Une première expérience (projets académiques ou stages) en NLP sera appréciée.
- Bon niveau de rédaction en français et en anglais
- Curiosité scientifique, intérêt pour la recherche

Informations pratiques

Début du stage : mars/avril 2026

Durée du stage : 6 mois

Unité d'accueil : Groupe Statistique et Outils d'Aide à la Décision (SOAD), département Services, Economie, Outils Innovants et IA (SEQUOIA) – EDF Lab Paris-Saclay, 7 boulevard Gaspard Monge, 91120 Palaiseau.

Télétravail 2j/semaine

Transmettre par mail un CV et une lettre de motivation à :
leila.hassani@edf.fr, julien.tourille@edf.fr