Représentation et Traitement de Documents Longs sous Forme de Graphes pour des Applications IA de Type RAG – H/F

Stage de 6 mois @ EDF R&D Saclay (91120)
Début du stage : mars 2025

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…).

Le stage s'inscrit dans un projet de recherche mené par les équipes R&D d'EDF en collaboration avec le Laboratoire Interdisciplinaire des Sciences du Numérique (LISN). Le projet vise à développer de nouvelles méthodes de représentations des documents longs.

Objectifs

Les modèles d'Intelligence Artificielle pour le traitement des données textuelles et orales ont connu des avancées techniques importantes depuis la création de l'architecture Transformer (Vaswani *et al.*, 2017). Cependant, le traitement des textes longs reste une tâche difficile tant pour le résumé automatique que pour les applications de question/réponse (Liu *et al.*, 2024).

Dans ce stage nous proposons d'étudier la création de représentations de documents longs sous forme de graphes qui pourront être utilisées par la suite dans des applications de type RAG (Edge *et al.*, 2024 ; Plenz *et al.*, 2024).

Le/la candidat.e sélectionné.e aura pour missions de :

- Dresser un état de l'art des méthodes de construction et de complétion automatique de graphes à partir de textes
- Implémenter et tester une ou plusieurs méthodes sur un corpus de documents
- Sélectionner des méthodes d'évaluation existantes et/ou proposer une nouvelle méthode

Edge, Darren, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, et Jonathan Larson (2024) « From Local to Global: A Graph RAG Approach to Query-Focused Summarization ». arXiv.2404.16130.

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, et Percy Liang (2024) « Lost in the Middle: How Language Models Use Long Contexts ». Transactions of the Association for Computational Linguistics.

Plenz, Moritz, et Anette Frank (2024) « Graph Language Models ». In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, et Illia Polosukhin (2017) « Attention is All you Need ». In Advances in Neural Information Processing Systems. Curran Associates, Inc.

Profil recherché

- Etudiant(e) de Master 2 en informatique/TAL ou dernière année d'école d'ingénieur
- Compétences en Traitement Automatique des Langues
- Compétences en Apprentissage Automatique (Deep Learning)
- Bon niveau en Python
- Bon niveau de rédaction en français et en anglais
- Bonne connaissance de pytorch
- Curiosité scientifique, intérêt pour la recherche

Informations pratiques

Début du stage : mars/avril 2025

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 :
eve.sauvage@edf.fr, sabrina.campano@edf.fr, julien.tourille@edf.fr