OFFRE DE STAGE EN NLP, IA, Data Science

Mission : Expliquer des « Embeddings » de texte à l'aide d'un Sparse Auto Encoder (SAE)

Durée : 6 mois - Début du stage souhaité : mars 2026 - Lieu : EDF R&D Lab Saclay (91120)

Contexte et objectifs

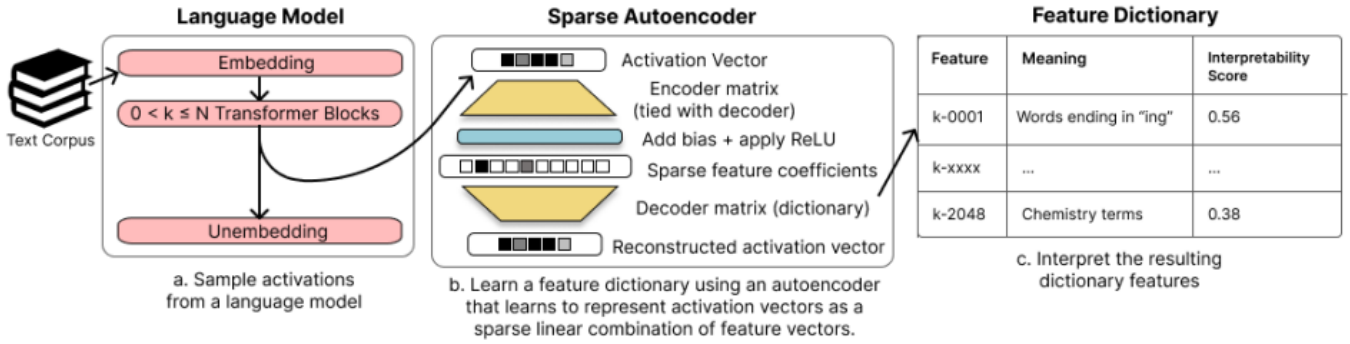
La R&D d'EDF (2000 chercheurs) a pour missions principales de contribuer à l'amélioration de la performance des unités opérationnelles du groupe EDF, d'identifier et de préparer les relais de croissance à moyen et long terme. Dans ce cadre, le département Services, Economie, Outils Innovants et IA (SEQUOIA) 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).

Ce stage concerne le domaine du TAL (Traitement Automatique des Langues) et plus particulièrement :

- les « **embeddings** » de texte obtenus à partir de la transformation d'un document textuel sous la forme d'un vecteur ;
- l'explication de ces « embeddings », à savoir essayer de mieux comprendre, expliquer ou interpréter les informations réellement encodées dans ces vecteurs.

Ces « embeddinsg » de texte sont maintenant très utilisés dans différentes tâches comme la classification (supervisée ou non), la détection de documents similaires, etc. Ils peuvent être génériques (gratuits ou payants) et éventuellement fine-tunés sur des données métiers. Mais comment savoir si un modèle est meilleur qu'un autre ? Comment choisir le meilleur modèle adapté à ses données ? Il est difficile de répondre à ces questions car la production d'embeddings s'apparente à un mécanisme de type « boîte noire ». Les SAE pour « Sparse Auto Encoder », apparues il y a quelques années pourraient être une solution pour mieux comprendre ces embeddings. C'est ce que propose de réaliser ce stage. Ici un schéma global décrivant un SAE :



Objectifs

Après une partie bibliographique sur le sujet des SAE, il s'agira d'entraîner un SAE[1] sur des données EDF non confidentielles (données synthétiques). Ces données seront converties sous la forme d'embeddings (par exemple avec Camembert). Une fois le SAE entrainé, il s'agira d'expliquer les nouvelles features ou vecteur à partir de techniques de visualisation par exemple.

Etapes du stage :

- Partie bibliographique sur les SAE ;
- Choix d'une implémentation existante sur GitHub ;
- Conversion des données textuelles sous la forme d'embeddings ;
- Entrainement d'un SAE ;
- Analyses, interprétation des nouveaux vecteurs (notamment à partir de visualisation) ;
- Rédaction d'un rapport.

Profil recherché :

- Formation : master machine-learning, IA data avec TAL + projet TAL.
- Niveau de langue : anglais (niveau B2 minimum), français oral et écrit.
- Compétences solides en programmation, en particulier en Python.
- Connaissance des techniques de traitement du langage naturel (NLP) et des modèles d'IA générative.
- Un goût pour la recherche est indispensable (autonomie, curiosité scientifique).

Références

[1] Sparse Autoencoders Find Highly Interpretable Features in Language Models, https://arxiv.org/abs/2309.08600

[2] A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models, https://arxiv.org/pdf/2503.05613v1

[3] Towards Interpretable Scientific Foundation Models: Sparse Autoencoders for Disentangling Dense Embeddings of Scientific Concepts, https://openreview.net/pdf?id=mPq3R6jdtD

Informations pratiques

Unité d'accueil : Groupe SOAD (Statistique et Outils d'Aide à la Décision), département SEQUOIA d'EDF Lab Paris-Saclay, 7 boulevard Gaspard Monge, 91120 Palaiseau. Le stage sera encadré par des ingénieurs-chercheurs Data Scientist du département SEQUOIA.

Transmettre par mail un CV et une lettre de motivation à:

philippe.suignard@edf.fr (Département SEQUOIA).

[1] - Des implémentations de SAE en python son disponible sur GitHub