



OFFRE STAGE EN DATA SCIENCE

Mission : Analyse topologique des données textuelles – H/F

Durée : 6 mois

Lieux : EDF R&D Lab Paris - Saclay (Palaiseau - 91120)

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, d'identifier et de préparer les relais de croissance à moyen et long termes. Dans ce cadre, le département Services, Economie, Questions hUmaines, 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. Ce stage sera rattaché au groupe « Statistiques et Outils d'Aide à la Décision » (SOAD) qui compte une vingtaine d'ingénieurs chercheurs spécialisés en text mining, informatique décisionnelle, data science et data engineering ayant pour missions de construire et mettre en œuvre les méthodes d'analyse, de fouille et d'enrichissement de données volumineuses d'origines multiples, structurées ou complexes.

En effet, le volume des données numériques textuelles, disponibles (enquêtes, centre d'appel etc.), augmente chaque année à un rythme exponentiel et l'analyse de ces informations, structurées ou non, est un impératif stratégique pour une entreprise telle qu'EDF. Dans ce cadre, et dans l'objectif de toujours mieux connaître les besoins de ses clients, l'exploitation de ces documents implique l'utilisation de méthodes et d'outils adaptés. Au cœur de ces problématiques, les outils de text mining sont de plus en plus nombreux et performants, ainsi nous souhaitons étudier les principaux outils évoluant sur le marché.

Objectifs

Depuis 2003, la R&D exploite les données textuelles du groupe EDF grâce à des méthodes d'IA adaptées et performantes. Ce domaine étant en constante évolution, il est nécessaire d'explorer de nouvelles pistes en amont dans l'optique de les adapter à des cas d'applications concrets.

L'analyse topologique de données (TDA) est un domaine d'étude récent qui repose sur des travaux de géométrie computationnelle datant du début des années 2000 (Edelsbrunner, 2002 ; Carlsson, 2009) et repose sur l'idée que la topologie et la géométrie dans des espaces en grande dimension permettent de déduire des informations qualitative et quantitative sur la structure même des représentations. Le TDA a été appliqué sur des jeux de données variés (tabulaire, séries temporelles) mais reste très peu appliqué aux données textuelles et pourrait être utile à des tâches de classifications, de détection de registres, etc.

L'objectif est donc de réaliser un état de l'art des méthodes existantes afin d'identifier le potentiel du TDA appliqué au texte. De plus, il s'agira de mener des tests avec des bibliothèques existantes permettant d'estimer la performance des outils et l'utilité des informations en sortie.

Ce stage se décomposera en 3 parties :

- Etat de l'art des méthodes de TDA
- Evaluation : tests des bibliothèques existantes et développements potentiels de nouvelles briques/modèles spécifiques aux données textuelles.
- Comparaison des résultats avec les outils actuellement en place à EDF.

Profil recherché :

- Etudiant(e) en master 1 ou 2 ou équivalent école d'ingénieur.
- M1 ou M2 en ingénierie informatique/mathématique.
- Connaissances des problématiques et technologies du text mining.
- Connaissances des méthodes de catégorisation et de clustering.
- Connaissances en développement informatique (Python).
- Connaissances des bases de données (MySQL).
- Connaissance des méthodes mathématiques, statistiques, de *machine learning* et *deep learning*.
- Des connaissances des environnements Big Data serait un plus.
- Des connaissances sur le secteur de l'énergie serait un plus.
- Curieux.se, ingénieux.se et motivé.e pour le domaine de la recherche appliquée.
- Bon niveau rédactionnel.

Informations pratiques

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

Le stage sera encadré par un.e ingénieur.e-chercheur.se du département SEQUOIA.

Début du stage souhaité : janvier 2022.

Transmettre par mail CV, diplômes, lettre de motivation et les bulletins de notes à :

clement.christophe@edf.fr (Ingénieur chercheur - Département SEQUOIA).