Numéro SIREA :

Titre : Chercheuse postdoctorale ou chercheur postdoctoral en intelligence artificielle et traitement du langage naturel (programme de recherche SCAI/BnF)

Corps : Contrat postdoctoral (12 mois, renouvelable)

Rattachement : UMR 7222 ISIR

Mot-clés : apprentissage automatique, explicabilité, bases de données, informatique, mathématiques appliquées, statistiques, traitement du langage naturel, recommandation

Les activités qui composent la fiche de poste sont appelées à évoluer en fonction des connaissances du métier et des nécessités de service

Qui sommes-nous ?

Sorbonne Université est une université pluridisciplinaire de recherche créée au 1er janvier 2018 par regroupement des universités Paris-Sorbonne et UPMC. Déployant ses formations auprès de 54 000 étudiants dont 4700 doctorants et 10 200 étudiants étrangers, Elle emploie 6 300 enseignants, enseignants-chercheurs et chercheurs et 4 900 personnels de bibliothèque, administratifs, techniques, sociaux et de santé. Son budget est de 670 M€. Sorbonne Université dispose d'un potentiel de premier plan, principalement situé au cœur de Paris, et étend sa présence dans plus de vingt sites en Île-de-France et en régions. Sorbonne Université présente une organisation originale en trois Facultés de Lettres, de Sciences & Ingénierie et de Médecine qui disposent d'une importante autonomie de mise en œuvre de la stratégie de l'université dans leur périmètre sur la base d'un contrat d'objectifs et de moyens. La gouvernance universitaire se consacre prioritairement à la promotion de la stratégie de l'université, au pilotage, au développement des partenariats et à la diversification des ressources.

Présentation de la structure

Dans un contexte national et international marqué par la compétition autour de l'intelligence artificielle, Sorbonne Université a créé le « Sorbonne Center for Artificial Intelligence » (SCAI), qui réunit dans un lieu unique, situé au cœur du quartier latin, un éventail stratégique des disciplines de l'intelligence artificielle moderne. L'ambition de SCAI est de contribuer significativement à l'excellence de la recherche interdisciplinaire en intelligence artificielle en favorisant les échanges entre enseignants-chercheurs, chercheurs, enseignants, étudiants et industriels.

Le projet de recherche décrit ci-dessous s'inscrit dans le cadre du partenariat stratégique entre Sorbonne Université et la BnF, qui rassemble dans ce cadre précis l'expertise de l'équipe MLIA de l'ISIR à la BnF afin de développer un travail de recherche commun au sujet des systèmes de recommandation.

La Bibliothèque nationale de France (BnF) est l'une des plus grandes bibliothèques patrimoniales du monde, Elle a pour mission de collecter, cataloguer, conserver, enrichir et communiquer le patrimoine documentaire national. Engagée depuis de nombreuses années dans d'ambitieux programmes de numérisation de ses collections, auxquels s'ajoute désormais l'entrée massive de collections nativement numériques, la BnF ne cesse d'enrichir son patrimoine numérique dont la masse, la diversité et le rythme d'accroissement nécessitent de nouveaux outils de traitement et de consultation. Pour permettre au plus grand nombre de découvrir et s'approprier ce patrimoine, la BnF s'implique depuis plusieurs années dans les technologies de l'intelligence artificielle (IA).

Mission et activités principales

Description du projet

Gallica, la bibliothèque numérique de la BnF, rassemble plus de 10 millions de documents numérisés librement accessibles en ligne (18,5 millions de visites par an). Cependant, la plupart des utilisateurs ne savent pas que Gallica contient des documents imprimés, mais aussi des photographies, des enregistrements sonores, des vidéos ou encore des objets 3D. Dans les enquêtes de satisfaction, seule une minorité considère que les réponses du moteur de recherche sont pertinentes et une majorité souhaiterait être mieux guidée dans ses recherches. Dans un contexte où la BnF vient d'adopter un nouveau modèle de données orienté entité-relation pour décrire ses collections (le modèle LRM) et envisage par ailleurs une refonte du moteur de recherche de Gallica, il s'agit d'interroger en quoi le recours à un modèle de langage approprié peut aider les utilisateurs à

se repérer dans la masse des collections et améliorer la visibilité des contenus les plus méconnus. Dans ce projet, la BnF s'engage à adopter une démarche résolument éthique. La potentielle exploitation des logs des utilisateurs doit respecter leur vie privée et garantir à la fois la pertinence et la transparence des algorithmes, en évitant les risques de bulles de filtre. Trois axes de réflexion se dégagent :
1) sur la base des informations disponibles, comprenant à la fois les logs des utilisateurs et les données des documents (plein texte par reconnaissance optique de caractères et métadonnées descriptives), comment évaluer la similarité entre documents ?
2) comment désambiguïser les requêtes des utilisateurs et leur faire des suggestions appropriées grâce à un agent conversationnel utilisant l'intelligence artificielle ?
3) comment instaurer la confiance des utilisateurs en matière de conception et d'audit des algorithmes ?

Missions principales
Ce projet consiste à travailler sur l'accès à l'information dans la bibliothèque Gallica, du point de vue des techniques d'apprentissage machine et profond. Les axes de recherche concernent (1) l'analyse et l'indexation des documents textuels ainsi que (2) l'analyse des traces utilisateur et (3) les systèmes de recommandation. Nous nous intéresserons en particulier aux techniques multimodales qui permettent de contextualiser un document ou une requête à partir des interactions d'utilisateurs.

La candidate ou le candidat retenu.e aura pour mission de :

- Mettre en œuvre des modèles pour apprendre la sémantique des données textuelles dans le but de les vectoriser.
- Développer des algorithmes basés sur des méthodologies d'apprentissage de représentation pour mêler efficacement texte et traces utilisateur.
- Utiliser un grand modèle de langue afin de générer des questions pertinentes pour l'utilisateur.
- Rendre compte et présenter le travail de développement de manière claire et efficace, à la fois pour discuter avec les experts de la BnF et rédiger des publications en machine learning.

La collection des livres imprimés sera prioritairement visée par le programme décrit ci-avant, mais une extension à d'autres collections dotées de descripteurs textuels (en particulier des collections iconographiques) pourra être envisagée.

Formation :
Un diplôme de doctorat en informatique ou équivalent est nécessaire, ainsi qu'un solide dossier scientifique, notamment en NLP et/ou Systèmes de recommandation et/ou Recherche d'information. Une expérience des projets de recherche internationaux et des applications en SHS serait un atout.

Informations générales :
Lieux : Campus Pierre et Marie Curie de Sorbonne Université et Datalab de la BnF
Contrat : à durée déterminée de 12 mois avec possibilité d'un prolongement
Date d'embauche prévue : le plus tôt possible
Quotité de travail : temps complet
Expérience souhaitée : 1 à 3 années
Salaire selon expérience

Principaux interlocuteurs :
Laure Soulier, MCF en informatique à Sorbonne Université, équipe MLIA, ISIR.
Vincent Guigue
Lucie Termignon, cheffe de projet données et Intelligence Artificielle à la BnF.
Jean-Philippe Moreux, Expert scientifique de Gallica à la BnF.

Encadrement : NON
Conduite de projet : OUI

Connaissances et compétences

Une solide formation en traitement du langage naturel ou en analyse de texte est essentielle, et de bonnes compétences en programmation sont requises. Une expérience des systèmes de recommandation est supposée. Une compréhension des enjeux éthiques de tels systèmes est également attendue.
Langue : connaissance du français non obligatoire mais fortement souhaitée

Les candidatures (CV + motivations + références éventuelles) sont à adresser par email à xavier.fresquet@sorbonne-universite.fr avec copie à philippe.chevallier@bnf.fr et laure.soulier@isir.upmc.fr