Appel à candidatures - Mission d'études et de recherches

Titre : Développement d'un outil de transcription vocale pour la documentation des données du patrimoine

Type de poste : Ingénieur·e de Recherche en NLP

Cadre : L'employeur est la Fondation des Sciences du Patrimoine

Contexte :

Créée en 2013 pour assurer la gouvernance du laboratoire d'excellence Patrima et de l'équipement d'excellence Patrimex, la Fondation des sciences du patrimoine (FSP) regroupe des établissements d'enseignement supérieur, des institutions patrimoniales ainsi que des laboratoires et des unités de service qui dépendent, entre autres, du ministère de la Culture et du Centre national de la recherche scientifique.

Retenu pour la programmation scientifique 2023 de la FSP, le projet **METAREVE**, fruit de la collaboration entre 3 laboratoires de recherche membres de la fondation, vise à répondre à des enjeux de traçabilité de données numériques dans le domaine des sciences du patrimoine. Il porte sur le développement d'une application dédiée à l'extraction automatique de métadonnées et de paradonnées, afin de simplifier le travail de documentation des données analytiques. L'approche proposée repose sur l'intelligence artificielle et plus particulièrement le NLP, à travers deux modules : l'un étant dédié à la **reconnaissance vocale**, et l'autre à **l'extraction d'entités nommées** à partir de **données non structurées**.

La complexité du projet réside dans le fait que les acteurs, les institutions, les qualificatifs, les techniques, ou les protocoles susceptibles d'être mentionnés sont très spécifiques à des domaines d'expertise et à des contextes d'analyse. Ils font souvent référence à des éléments de vocabulaire n'étant pas répertoriés dans les corpus d'apprentissage usuels, et évoluent trop rapidement pour que l'entrainement d'un modèle unique et stable soit une option viable. Ainsi, le défi principal concernera la définition de stratégies pour **l'adaptation thématique du modèle de langue** ainsi que le recours à des méthodes de **désambiguïsation** (WSD), qui pourront s'appuyer sur des thésaurus et des ontologies existants.

Activités :

Intégré à une équipe pluridisciplinaire, l'ingénieur.e aura pour mission de développer des algorithmes de transcription vocale et d'extraction d'entités nommées à partir de données non structurées (textes, rapports d'expertise). Leur implémentation sous forme de prototype sera menée en collaboration avec un ingénieur en informatique ainsi qu'un stagiaire en sciences des données intégrés à l'équipe.

Compétences :

- De formation supérieure BAC+5 ou doctorat avec un profil scientifique.
- Expertise en IA et Machine Learning, avec au moins une expérience réussie en traitement automatique du langage naturel
- Intérêt pour la recherche interdisciplinaire, capacité à interagir avec d'autres disciplines
- Français courant
- Bon niveau d'anglais

Caractéristiques du poste :

Catégorie : Ingénieur·e de recherche
Type : Contrat à Durée Déterminée d'un an (renouvelable)
Quotité : 100%
Affectation : Cergy Paris Université (site de Neuville). Possibilité de télétravail hors nécessité de service et déplacements fréquents à prévoir en Île-de-France
Rémunération : 40k€ brut annuel (en fonction de l'expertise du candidat)
Équipe de supervision scientifique et interactions : Violette Abergel (MAP, CNRS), Besma Zeddini, (SATIE, CY Cergy Paris Université), Vincent Detalle, (SATIE, CY Cergy Paris Université), Olivier Malavergne (LRMH, CRC) et Ruven Pillay (C2RMF).

Pour candidater :

Pour candidater, envoyer un document PDF (un seul fichier) incluant :

+ Curriculum Vitae détaillé
+ lettre de motivation
+ Diplômes

à l'équipe l'adresse suivante :
violette.abergel@map.cnrs.fr

Date de début du contrat : novembre-décembre 2023 (selon disponibilité du candidat)