Offre n°2025-08795

Ingénieur OCR et Analyse de mise en page

Type de contrat : CDD

Niveau de diplôme exigé : Thèse ou équivalent

Fonction : Ingénieur scientifique contractuel

Niveau d'expérience souhaité : Jeune diplômé

Contexte et atouts du poste

Ce poste d'ingénieur IA est ouvert au sein de l'équipe-projet ALMAnaCH du Centre Inria de Paris. ALMAnaCH est une équipe de recherche d'une cinquantaine de membres, dont 7 membres permanents, spécialisée dans le traitement automatique des langues (TAL) et les humanités numériques (traitement des sources historiques et littéraires par l'informatique). Elle fait partie des 230 équipes-projets d'Inria, l'institut national de recherche en informatique et en automatique, établissement public de recherche regroupant 9 centres de recherche, dont le Centre de Paris auquel appartient ALMAnaCH.

Le poste prend place dans le cadre du projet européen ATRIUM (Advancing fronTier Research In the arts and hUManities), qui vise à relier les principales infrastructures de recherche dans les arts et les humanités (DARIAH), l'archéologie (ARIADNE), les langues (CLARIN) et la communication savante ouverte en sciences humaines et sociales (OPERAS). ATRIUM répond aux défis posés par la diversité des disciplines en sciences humaines et sociales, en proposant des services interopérables adaptés à des communautés aux méthodologies variées.

Dans ce contexte, le travail portera plus particulièrement sur la segmentation des documents complexes (formulaire) et la gestion des caractères rares dans le cadre de la reconnaissance de texte en ensemble semi-ouvert.

L’employé travaillera en collaboration directe avec Thibault Clérice, tout en interagissant avec différents membres de l’équipe impliqués dans la conception d’interfaces utilisateur et d’expériences utilisateur (UI/UX), la définition des lignes directrices pour la segmentation des documents, ainsi que d’autres aspects liés à la reconnaissance optique de caractères (OCR).

Mission confiée

Missions :
Sous la responsabilité de Thibault Clérice, la personne recrutée aura pour mission d'améliorer les capacités de moteurs de segmentation et de reconnaissance automatique de texte (ATR) dans le cadre de leur utilisation dans les plate-formes eScriptorium et dérivées. En particulier, sont concernés:

- la reconnaissance de texte en ensemble semi-ouvert pour les documents archéologiques;
- la reconnaissance en segmentation de formulaire;
- le few-shot learning de segmentation, pour des documents répétitifs.

Collaboration :
La personne recrutée sera en lien avec l'ingénieur de BACK IN TIME pour l'intégration des systèmes dans l'interface eScriptorium, avec une autre ingénieure Inria d'ATRIUM pour l'implémentation des workflows designés dans le cadre du WP4 et de la mise en place des demonstrators du WP5. Plus largement, des collaborations internes avec l'ingénieur OCR du PIQ CLLG et les ingénieurs du projet COLaF sont attendues.

Au national et à l'international, des collaborations avec le groupe responsable d'eScriptorium sont attendues, dont des réunions hebdomadaires, ainsi qu'avec le reste du projet ATRIUM.

Principales activités

Principales activés:

- Développer la reconnaissance de texte en ensemble semi-ouvert pour les documents archéologiques et l'intégrer à des moteurs compatibles avec eScriptorium ou ses dérivés;
- Développer la reconnaissance en segmentation de formulaire et l'intégrer à des moteurs compatibles avec eScriptorium ou ses dérivés;
- Développer e few-shot learning de segmentation, pour des documents répétitifs et l'intégrer à des moteurs compatibles avec eScriptorium ou ses dérivés;
- Participer aux réunions autour des plate-formes eScriptorium et ses dérivés dont Inria est membre ainsi qu'aux réunions du groupe ATRIUM;

- Assurer une documentation des fonctionnalités produites.

Activités complémentaires:

- Aider à la création de recommandation pour la segmentation de documents modernes liés à l'archéologie;
- Entraîner des modèles adaptés aux besoins du projet et les mettre à disposition;
- Aider et maintenir les moteurs concernés dans eScriptorium et ses dérivés.

Missions collectives : participation à la vie de l'équipe et des projets, y compris pour des conférences ou rencontres internationales à l'étranger.

Compétences

Compétences techniques et niveau requis :

- Maîtrise de PyTorch et PyTorch lightning;
- Intérêt pour les sciences humaines et leurs particularités;
- Maîtrise de Git;
- Connaissance des standards de la reconnaissance de texte (ALTO/PageXML).

Langues :

- Anglais B2 minimum
- Français B2 souhaité
- Une autre langue est appréciée, en particulier dans le cadre européen d'ATRIUM.

Compétences relationnelles :

- Bonnes capacités d’organisation.
- Bon relationnel.

Compétences additionnelles appréciées :

Avantages

- Restauration subventionnée
- Transports publics remboursés partiellement
- Congés: 7 semaines de congés annuels + 10 jours de RTT (base temps plein) + possibilité d'autorisations d'absence exceptionnelle (ex : enfants malades, déménagement)
- Possibilité de télétravail et aménagement du temps de travail
- Équipements professionnels à disposition (visioconférence, prêts de matériels informatiques, etc.)

- Prestations sociales, culturelles et sportives (Association de gestion des œuvres sociales d'Inria)
- Accès à la formation professionnelle
- Sécurité sociale

Informations générales

- **Thème/Domaine :** Langue, parole et audio
  Production, traitement et analyse des données (BAP D)
- **Ville :** Paris
- **Centre Inria :** Centre Inria de Paris
- **Date de prise de fonction souhaitée :** 2025-07-01
- **Durée de contrat :** 2 ans
- **Date limite pour postuler :** 2025-05-04

Contacts

- **Équipe Inria :** ALMANACH
- **Recruteur :**
  Clerice Thibault / thibault.clerice@inria.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'e?orce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

L'essentiel pour réussir

- Se sentir à l'aise dans un environnement interdisciplinaire, aimer apprendre et écouter sont des qualités essentielles pour réussir cette mission.
- Interessé par les problématiques de sciences ouvertes et la reproducibilité des sciences.
- Une thèse dans le domaine de l'OCR ou de l'analyse de la mise en page constitue un réel atout.

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.