Poste de Post-Doctorat

Classification automatique de documents d'entreprise

Le LabCom IDEAS lance un appel à candidatures pour un poste de post-doctorant en informatique dans le domaine de l'apprentissage automatique pour la classification de documents d'entreprise.

Durée : 12 mois (avec des possibilités de renouvellement)
Date d'embauche souhaitée : 1er Septembre 2020 (pouvant être modifiée en fonction de la situation sanitaire)
Salaire : 2100 € net / mois
Lieu de travail : LabCom IDEAS, dans les locaux du laboratoire L3i à La Rochelle, France
Spécialités : Informatique / Apprentissage automatique / Classification / Analyse d'images / Traitement Automatique de la Langue

Description du LabCom :

Les travaux menés par le candidat s'inscriront dans le cadre du LabCom IDEAS, co-financé par l'Agence Nationale de la Recherche (ANR) et la région Nouvelle-Aquitaine, et regroupant l'entreprise Yooz et le laboratoire L3i. Ce LabCom a pour objectif d'imaginer, inventer, concevoir, développer, optimiser et entrainer les meilleurs algorithmes de traitement automatiques des documents d'entreprise pour offrir un service d'intelligence artificielle capable de comprendre un maximum de document d'entreprise.

Le post-doctorant sera basé au sein du LabCom, localisé dans les locaux du Laboratoire Informatique, Image et Interaction (L3i), à La Rochelle.

Le laboratoire L3i, EA 2118 créé en 1993, représente la seule composante de recherche du domaine STIC à l'Université de la Rochelle associant les chercheurs de l'IUT de la Rochelle, et du Pôle Sciences en informatique. Dans le cadre de la politique quadriennale (désormais quinquennale) de l'université de la Rochelle, le L3i a été évalué A par l'AERES.
Le large déploiement des technologies numériques et la multiplicité des processus d'acquisition et de diffusion de l'information engendrent un développement rapide et diversifié des modes de production et de consommation de contenus numériques, ainsi qu'une croissance exponentielle de la volumétrie des données. Par ailleurs, l'avènement des dispositifs nomades interactifs augmente encore plus les problématiques de positionnement de l'utilisateur dans la gestion et la navigation au sein de contenus numériques.
Il s'agit, pour le L3i, de mettre en synergie les compétences établies dans le laboratoire afin d'aborder la problématique de la valorisation des contenus numériques sous un angle systémique. Cela revient, en particulier, à une exploitation croisée des compétences en matière d'applications interactives, d'indexation par le contenu, et de représentation de connaissances. Le laboratoire se structure autour de trois thématiques scientifiques (Ingénierie des connaissances, Analyse et gestion de contenus, Interactivité et dynamique

des systèmes), toutes centrées sur la problématique de la gestion interactive et intelligente des contenus numériques.

Yooz, partenaire industriel du Labcom, est fournisseur d'un service Cloud d'automatisation des processus d'achat et de paiement. Yooz intègre des technologies d'Intelligence Artificielle pour automatiser les processus et le traitement des documents impliqués dans ces processus. Le service yooz est utilisé quotidiennement par près de 3000 utilisateurs.

Le travail de recherche et développement mené au sein du LabCom s'articule autour de 3 grands axes :
- Classification de documents
- Fouille de documents
- Détection de fraude documentaire

Description du poste :

Le travail du post-doc recruté s'inscrira dans le cadre de l'axe "Classification de documents". Il s'agit de concevoir des approches innovantes pour la classification de documents (selon leur nature : facture, devis, RIB, …) dans des flux documentaires multicanaux non-structurés et de créer, à partir de ces approches, un prototype de laboratoire.

Les verrous scientifiques découlant de ce contexte applicatif, et relevant essentiellement du domaine de l'apprentissage automatique, sont nombreux :

- les classes de documents sont généralement très déséquilibrées dans les corpus existants. En effet, certaines classes sont très bien représentées dans la base d'apprentissage, tandis que d'autres le sont beaucoup moins (voire pas du tout). En conséquence, les approches développées jusqu'à présent offrent des taux de précision très inégaux entre classes.

- la variabilité intra-classes est très grande (parfois même supérieure à la variabilité entre différentes classes). On peut citer à titre d'exemple le fait que deux documents de classes différentes (par exemple une facture et un devis) provenant de la même entreprise peuvent être, visuellement et en termes de contenu textuel, plus proches que deux factures d'entreprises différentes.

Ce travail de post-doctorat s'appuiera sur un état de l'art détaillé des approches existantes, pour en identifier les limites, et proposer des approches innovantes qui permettent de contourner les difficultés mentionnées ci-dessus. Pour résoudre ces problèmes, nous envisageons d'utiliser des techniques d'apprentissage automatique qui, basées sur des techniques existantes de classification d'images et/ou de contenu textuel, permettent de :

- prendre conjointement en compte ces deux modalités (image et texte) pour la classification de documents (multi-modalité), afin d'améliorer la précision pour la plupart des classes

- apprendre une classe à partir de très peu d'exemples, voire d'aucun exemple (*zero-shot learning*), le cas échéant à la volée, dans le flux de documents

- mettre en œuvre efficacement une stratégie de rejet, dès lors que le document à classer est trop éloigné des classes existantes, ou bien lorsque l'ambiguïté entre classes est trop importante (et ce, avec des seuils fixés de manière automatique ou semi-automatique, en fonction du corpus).

Si le chercheur souhaite acquérir/renforcer une expérience en milieu industriel, il serait possible d'organiser de courts séjours de travail collaboratif au sein de l'entreprise Yooz, sur le site d'Aimargues (côte méditerranéenne).

Profil recherché :

Le candidat, titulaire d'un doctorat dans les domaines de l'informatique, du génie informatique et traitement du signal, ou des mathématiques appliquées, devra justifier d'une expérience de recherche dans au moins deux des domaines suivants :

- Apprentissage automatique / classification
- Analyse d'images
- Reconnaissance de formes

Des connaissances en Traitement Automatique de la Langue seraient appréciées.

Les compétences du candidat incluront :

- Maitrise nécessaire d'un ou plusieurs langages de programmation (Java, Python, C/C++...)
- Très bonnes aptitudes au travail en équipe, une connaissance des méthodes Agile serait un plus (le travail sera mené à la fois en lien avec les chercheurs du laboratoire L3i et le service R&D de l'entreprise Yooz)
- Bonne aptitude à la rédaction d'articles scientifiques et maîtrise de l'anglais écrit et parlé

Pour postuler :

Les candidats à ce poste devront envoyer un CV et une lettre de motivation (les noms et coordonnées de références seraient un plus) à :

- muriel.visani [chez] univ-lr.fr
- nicolas.sidere [chez] univ-lr.fr
- Vincent.PoulaindAndecy [chez] getyooz.com
- Aurelie.Joseph [chez] getyooz.com

Les candidatures seront étudiées au fil de l'eau et donc il n'y a pas de date limite de candidature. Néanmoins, nous attirons votre attention sur le fait que notre objectif est, idéalement, d'avoir sélectionné le meilleur candidat pour la mi-juillet.