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.

Post-Doctoral job offer

Automatic classification of business documents

The LabCom IDEAS calls for applications for a post-doctoral position in computer science, in the field of machine learning, for the classification of business documents.

***Duration**: 12 months (with possibilities of renewal)*
***Desired hiring date**: September 1st, 2020 (may be modified according to the sanitary situation)*
***Take-home salary**: 2100 € / month (French public healthcare coverage included)*
***Workplace**: LabCom IDEAS, on the premises of the L3i laboratory in La Rochelle, France*
***Specialities**: Computer Science / Machine Learning / Classification / Image Analysis / Automatic Language Processing*

Description of the LabCom:

The work carried out by the candidate will be part of the LabCom IDEAS. IDEAS is co-funded by the French National Research Agency (ANR) and the Région Nouvelle-Aquitaine, and brings together the Yooz company and the L3i laboratory. The IDEAS objective is to imagine, invent, design, develop, optimise and train the best algorithms for processing automatically business documents. The goal is to offer services, based on artificial intelligence, that are capable of automatically analysing and understanding various types of business documents.

The post-doc fellow will be based in the LabCom, located in the premises of the Laboratory L3i, in La Rochelle, France.

The L3i laboratory, created in 1993 at La Rochelle University brings together researchers in Computer Science and Signal Processing from different faculties. The L3i brings together the skills of its researchers in order to address the issues of digital content enhancement from a systemic perspective. This relies, in particular, on a cross exploitation of interactive applications, content indexing and knowledge representation. The laboratory is structured around three scientific themes (Knowledge Engineering, Content Analysis and Management, Interactivity and Dynamic Systems), centred on the common goal of interactive and intelligent management of digital content.

Yooz, Labcom's industrial partner, is a provider of a Cloud service for automating purchasing and payment processes. Yooz integrates Artificial Intelligence technologies to automate the processes and document processing involved in these processes. The yooz service is used daily by nearly 3000 users.

The research and development carried out within the LabCom focuses on 3 main areas:
- Document classification
- Document search
- Document fraud detection

Job description:

The work of the post-doc fellow will fall within the framework of the area "Document Classification". The aim is to design innovative approaches for classifying documents (according to their nature: invoice, quotation, bank account details, ...) in multi-channel incoming document flows and to create, from these approaches, a prototype.

There are many scientific bottlenecks arising from this applicative context, mainly in the field of machine learning:

- document classes are generally very unbalanced in existing corpuses. Indeed, some classes are very well represented in the learning base (with many training documents), while others are much less represented (if present at all). As a result, the approaches developed so far offer very uneven rates of accuracy between classes.

- the intra-class variability is very large (sometimes even greater than the inter-class variability). An example of this is the fact that two documents from the same company but from different classes (*e.g.* an invoice and a quotation) may be closer than two invoices from different companies in the representation space, both visually and in terms of textual content.

This post-doctoral work will be based on a detailed state of the art of existing approaches, to identify their limits and propose innovative approaches that will help to overcome the bottlenecks mentioned above. To solve these problems, we plan to use machine learning techniques that, based on existing image and/or text content classification techniques, allow to:

- take both modalities (image and text) into account jointly for document classification (multi-modality), in order to improve accuracy for most classes

- learn a class from very few (or even 0) examples (zero-shot learning), possibly on the fly, in the document flow

- effectively implement a rejection strategy when the document to be classified is too “far” from existing classes, or when the ambiguity between classes is too great (with thresholds that will be et automatically or semi-automatically, depending on the corpus).

If the post-doc fellow would like to acquire/reinforce his/her experience of working in a private company, we could arrange short collaborative working stays in the premises of Yooz company, at Aimargues (Mediterranean coast).

Candidate Profile:

The candidate, who holds a Ph.D. in the fields of computer science, computer engineering, signal processing, or applied mathematics, must have a significant research experience in at least two of the following areas:

- Machine learning / classification
- Image analysis
- Pattern recognition

Moreover, knowledge or experience of Automatic Language Processing would be appreciated.

The candidate's skills will include:
- Mastering one or more programming languages (Java, Python, C/C++...)
- Very good teamwork skills, having knowledge or experience of Agile methods would be a plus (a the work will be carried out both in conjunction with researchers from the L3i laboratory and the R&D department of the Yooz company).
- Good scientific writing skills, and fluency in writing and speaking English.

To apply:

Candidates for this position should send a CV and a cover letter (names and reference details would be appreciated) to:

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

Applications will be considered as they arise, so there is no strict deadline for applying, even though we would prefer to have selected the best applicant by mid-July.