Offre n°2024-07298

Doctorant F/H Analyse de la géométrie des plongements de mots et applications

Type de contrat : Fixed-term contract

Niveau de diplôme exigé : Graduate degree or equivalent

Fonction : PhD Position

Contexte et atouts du poste

Dans le cadre d'un projet de recherche concernant les humanités numériques (projet Apollon)

L' objectif est de procéder à une analyse fine des techniques de plongements de mots (algorithmes de type word2vec, Bert, ...) en s'attachant à la géométrie des plongements, afin de déterminer les meilleures métriques pour procéder à du clustering.

Des applications à l'étude de textes antiques seront développées dans le cadre du projet Apollon et cette thèse comporte un volet *interdisciplinaire*.

Des déplacements réguliers sont prévus pour ce poste ?

Déplacements réguliers entre Nancy et Strasbourg.

Mission confiée

Missions :
La personne recrutée sera amenée à analyser des algorithmes de plongements de mots et proposer des techniques de clustering qui se fondent sur la géométrie de ces plongements.

Grâce aux méthodes de plongement (Word2Vec, Context2Vec, Google BERT, etc.) qui permettent d'identifier les mots d'un texte à une série de vecteurs, le traitement automatique des langues a énormément progressé ses dernières années. On peut même dire que ces méthodes ont introduit une véritable révolution puisqu'elles ont levé les limites qu'imposaient les modèles de langues avec N-grams. Les résultats obtenus sont néanmoins à relativiser très fortement. En effet, le fonctionnement de ces méthodes n'a jusqu'alors été établi que de manière empirique sur des jeux de données précis et volumineux, et principalement en langue anglaise, langue sans signes diacritiques. Le but de cette thèse est d'étudier numériquement et théoriquement l'évolution des géométries formées par les plongements tout au long de l'entraînement d'une méthode de plongement. Les méthodes pourront être retravaillées afin d'augmenter la convergence, mais surtout la stabilité, au niveau empirique. Ensuite, le but est d'établir des résultats de convergence sur les plongements afin de démontrer la robustesse et la stabilité de ces méthodes. Enfin, le doctorant pourra également chercher à établir des ponts entre les méthodes de plongement et les méthodes de réduction de dimension classiques (Analyse en composantes principales, analyse des correspondances, méthodes spectrales, etc.), et les appliquer à des methodes d'agrégation de données (clustering).

Pour une meilleure connaissance du sujet de recherche proposé :
Un état de l'art, une bibliographie, des références scientifiques sont disponibles à l'URL suivante, n'hésitez à pas à vous y connecter : https://project.inria.fr/aexapollon/

Collaboration :
La personne recrutée sera co-encadrée par Lionel Lenotre (Université de Haute-Alsace) et prendra part au projet d'humanités numériques Apollon sur l'étude des textes antiques. Par ailleurs, la personne recrutée prendra part à l'activité scientifique du projet PASTA, ainsi qu'à celle de l'Institut Elie Cartan de Lorraine.

Responsabilités :
La personne recrutée a la charge de veiller au bon développement méthodologique et technologiques et prendra des initiatives pour mettre en oeuvre les algorithmes proposés en les appliquant à des données réelles.

Principales activités

Principales activés :

- analyser les algorithmes de plongements de mots, notamment en ce qui concerne les aspects géométriques
- proposer de nouvelles méthodes de clustering
- communiquer les résultats sous forme d'article et de conférences, vers différentes communautés disciplinaires.
- mettre en oeuvre de nouvelles techniques, les programmer et les tester sur des données réelles, notamment sur des textes issus de l'histoire antique.

Compétences

Compétences techniques et niveau requis : Master 2 en mathématiques appliquées avec vers la statistique et la science des données, en particulier, des connaissances sur les méthodes de régressions, les réseaux de neurones et les méthodes de réduction de dimension. Des connaissances en aggrégation de données (clustering), en géométrie différentielle (espaces géodésiques, espaces de longueur) seront un plus.

Langues : français / anglais

Compétences relationnelles : capacité à communiquer à des publics différents : mathématiques, spécialistes en SHS, informatique, ...

Compétences additionnelles appréciées : programmation, notamment en python et utilisation de tensorflow, scikit-learn ou équivalent..

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 (après 6 mois d'ancienneté) 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

Rémunération

2100 € brut/mois la 1ère année

Informations générales

- **Thème/Domaine :** Stochastic approaches
  Statistics (Big data) (BAP E)
- **Ville :** Villers lès Nancy
- **Centre Inria :** Centre Inria de l'Université de Lorraine
- **Date de prise de fonction souhaitée :** 2024-10-01
- **Durée de contrat :** 3 years
- **Date limite pour postuler :** 2024-05-12

Contacts

- **Équipe Inria :** PASTA
- **Directeur de thèse :**
  Lejay Antoine / Antoine.Lejay@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'efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

L'essentiel pour réussir

La personne recrutée devra avoir un goût pour la science des données, la théorie mathématique, aussi bien du côté méthodologique que du côté pratique, et être capable de programmer et mettre en pratique les avancées proposées sur des données réelles.

Elle devra aussi être capable de communiquer à différents publics, aussi bien mathématiques que des spécialistes des sciences humaines et sociales, notamment des spécialistes de l'histoire antique.

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.