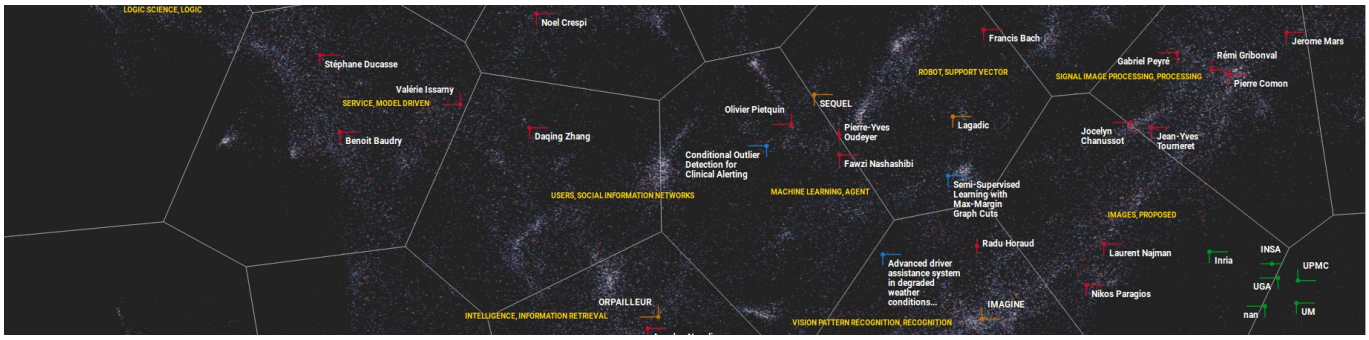
Offre de Stage Master ou ingénieur

Evaluation et apprentissage d'hyperparamètres pour la visualisation de grandes masses de données



Missions :

Les missions du stagiaire consistent à mettre en place un protocole de test pour évaluer les résultats de Cartographie scientifique obtenus par Cartolabe (cartolabe.fr). Dans un premier temps, des indicateurs de qualité et un protocole de validation seront mis en place. L'utilisation d'autres jeux de données tel que wikipedia et la comparaison à d'autres moteurs de recherche scientifique permettra de valider les indicateurs et protocoles de test. Dans un second temps, l'optimisation des hyperparamètres de la cartographie permettra d'améliorer la qualité obtenue.

Activités :

Cartolabe est un projet développé en commun par le LRI, le CNRS et l'INRIA afin de visualiser un grand nombre de publications, d'auteurs, laboratoires et équipes sur une même carte.
L'application Cartolabe calcule une distance entre ces entités liées à des publications à partir du texte des articles. Un pipe-line de traitement de données extrait les données depuis HAL (https://hal.archives-ouvertes.fr/: aujourd'hui 750 000 articles et auteurs) puis les traite en utilisant des techniques de machine learning. Un unique fichier json est produit en sortie du pipe-line. Ensuite, une deuxième partie du logiciel (application web) se charge de visualiser cet ensemble de points en une carte de chaleur annotée et zoomable. Il est possible à partir du client web de naviguer et d'explorer la carte.

Un exemple d'indicateur de qualité intrinsèque assez naturel est de compter parmi les articles voisins d'un auteur, le pourcentage de ceux dont il est lui-même auteur.
Des indicateurs de qualité extrinsèques peuvent être établis en soumettant des requêtes identiques à des applications indépendantes de Cartolabe comme google scholar ou LookInLabs (https://lookinlabs4halinria.cominlabs.u-bretagneloire.fr/).

1 Version du 23 janvier 2019

Des indicateurs de qualité manuels sont également envisageables en interrogeant des scientifiques au cours de sessions enregistrées et en confrontant leur appréciation personnelle des distances entre entités avec les résultats proposés par Cartolabe.
Certains indicateurs de qualités peuvent par ailleurs être validés sur des jeux de données disposant d'autres indicateurs de proximité, tels que les liens entre articles de wikipedia ou les citations croisées entre articles scientifique.

Une fois les indicateurs de qualité mis en place sur la base Cartolabe, une deuxième partie du stage consistera à réaliser une optimisation des hyperparamètres de Cartolabe afin d'étudier leur impact sur les différents indicateurs et d'améliorer le résultat obtenu. Les hyperparamètres à considérer peuvent aussi bien être des choix d'algorithmes (LDA/LSA, choix du type de voisinage, projection UMAP ou TSNE) que leurs paramètres (en particulier le nombre de dimensions latentes utilisées pour calculer la similarité sous jacente).

Domaines de connaissances souhaitées:

- Langage et des outils de programmation Python (Anaconda, scikit-learn, pandas);
- Pratique des environnements de développement logiciels (forges, git) ;
- Notions appréciées dans l'un des domaines suivants : visualisation de grandes masses de données, machine learning, traitement automatique des langues ; information retrieval : recall versus precision.
- Maîtrise de l'anglais scientifique ;
- Méthode, curiosité et aptitude au travail en équipe sont requis pour ce stage.

Contexte:

TAU (TAckling the Underspecified) est une équipe projet Inria commune avec le LRI. Le Laboratoire de Recherche en Informatique est une unité mixte de recherche rattachée à l'INS2I du CNRS et au département STIC de l'Université Paris-Saclay ayant des liens de partenariat avec Inria et CentraleSupelec. Le laboratoire accueille près de 300 personnes dont 133 permanents et 90 doctorants.
Le projet Cartolabe est un projet à fort potentiel porté par une équipe d'ingénieurs et de scientifiques du LRI et d'INRIA. Le moteur de l'application et le module de visualisation sont conçus de façon très ouverte pour être applicable à de nombreux domaines. La mise au point des hyperparamètres qui fait l'objet de la deuxième partie du stage est essentielle car c'est elle qui permet la validation d'une instance de Cartolabe dans une domaine donné, comme les publications scientifiques par exemple.

Niveau : Stage de master M1 ou M2 ou ingénieur 3ème ou 4ème année.

Convention de stage : Université Paris Saclay ou INRIA. 550€ environ de gratification mensuelle, selon le nombre de jours dans le mois.

Lieu : Equipe-projet INRIA TAU. LRI, Laboratoire de Recherche en Informatique – Université Paris Sud – Bâtiment 660 - Shannon.

Durée (entre mars et Août 2019): 4 à 6 mois.

Envoyez CV et lettre de motivation à

Philippe Caillou : caillou@lri.fr et Anne-Catherine Letournel : acl@lri.fr