SCALIAN

Vos talents, Nos projets : Innovons ensemble !

OFFRE DE STAGE

Anonymisation de donnée par le Machine Learning et Deep Learning

Toulouse | réf. XXXXX

Contexte Scientifique

Avec l'augmentation du nombre de fuite de données, d'attaques et autres logiciels de rançon de ses dernières années, les préoccupations autour de la sécurité et de la confidentialité des données sont le principal frein au partage des données. Il en résulte un besoin de protéger ces données pour différentes raisons. Si les raisons purement morales ne sont que rarement suffisantes pour justifier des dépenses de sécurisation, les conséquences en termes d'images et le coût associé sont beaucoup plus moteurs. De plus, la législation européenne se durcissant, des conséquences judiciaires ou pénales ne peuvent plus être écartées. Ainsi, le règlement général sur la protection des données (RGPD), règlement européen en vigueur depuis le 24 mai 2016 et applicable à partir du 25 mai 2018, prévoit des sanctions pouvant aller jusqu'à 4 % du chiffre d'affaires annuel mondial et 20 millions d'euros.
Avec l'essor du Big Data et de l'IA, de plus en plus de données sont collectées, de nos téléphones, de nos navigateurs web, de nos rendez-vous médicaux.... Garantir la confidentialité des données sensibles est ainsi un enjeu des plus actuels et des plus complexes.
Le L@B travaille actuellement sur un projet R&D visant à comparer des méthodes d'anonymisation parmi les plus récentes, et les rendre compatibles avec des textes, images et vidéos, en mêlant pour cela Deep Learning et Confidentialité Différentielle. Ces méthodes récentes ont l'avantage de prouver mathématiquement l'anonymisation, et garantissent ainsi l'anonymat de nous, utilisateurs.

Objectifs

L'objectif de ce stage est d'étendre une méthode d'anonymisation (déjà implémentée) issue de la recherche à de nouveaux formats de données. Il s'agit essentiellement de :

- Etudier l'application de des méthodes de l'anonymisation implémentée sur des jeux intégrants des types de données variées (texte, numériques …).
- Implémentation d'un algorithme (sous python) qui traite les données catégorielles, ainsi que son intégration à notre méthode d'anonymisation.
- Utilisation du Deep Learning et Transfert Learning pour étendre la méthode aux jeux de données constitués d'images
- Etudier les performances de ces méthodes sur des datasets contenant du langage familier (Tweets par exemple), comme le langage familier reste un challenge en NLP ;
- Améliorer la méthode pour supporter de plus gros datasets

Profil attendu

- Formation en **statistique/mathématiques appliquées** ou formation en **informatique** avec une expérience en Machine Learning et/ou Deep Learning.
- Niveau Bac+5 (universitaire ou école d'ingénieur)
- Bon niveau de programmation en Python
- Goût pour l'innovation

Encadrement

David Jaidan (L@B) & Maxime Carrere (BU Datascale)

Aspects administratifs

Durée : 6 mois
Lieu : SCALIAN Eurogiciel
Toulouse

Merci de nous adresser vos CV et LM par e-mail à l'adresse suivante : david.jaidan@scalian.com