Mise en place d'un système d'analyse de données textuelles hybride

Stage de fin d'étude (Master2, dernière année école d'Ingénieurs) à pourvoir dès que possible

QUI EST AKTAN ?

Aktan est une société reconnue dans le domaine de l'innovation par les services. Elle travaille notamment avec de grandes entreprises dans des secteurs variés (Safran, RATP, Crédit Agricole, Michelin, Dassault Aviation, …). Dans ce contexte, Aktan a investi dans le domaine du Traitement Automatique du Langage depuis quelques années et oriente sa recherche sur l'utilisation de techniques d'Intelligence Artificielle telles que les Réseaux de Neurones pour mettre en œuvre un nouveau système optimisé de Traitement Automatique du Langage. L'environnement intègre des phases d'analyses grammaticales et d'analyse probabilistes (réseaux de Neurones).

SUJET DU STAGE

Le stage est donc centré sur une participation active à l'étude, au développement et à la mise en œuvre d'un système d'analyses multiples de données textuelles (compte-rendus d'observations, transcriptions d'entretiens téléphoniques, interviews, etc). Il consiste en plusieurs tâches :

- Architecture du système
    - Définition des flux pour appliquer différents types d'analyses (reconnaissance d'entités, relations entre entités, analyse de sentiments, …) à un ensemble de corpus
    - Préparation de « datasets » d'entrainement
        - Pré annotations d'un ensemble de textes
- Optimisation des « plongements de mots » (Word Embeddings)
        - Utilisation des algorithmes Word2Vec et Fasttext
        - Etude des nouveaux algorithmes de « Word Embeddings » contextuels (Elmo, ULMFit, BERT, …)
- Développement de modèles de réseaux de neurones pour implémenter des analyses fines (analyse de sentiments par aspect, …) et labélisation d'entités et/ou de séquences.
- Développements autour de l'outil open source Unitex/Gramlab (notamment agrégation et visualisation de résultats)

La phase de construction des « datasets » d'entrainement s'appuie sur les capacités d'analyses grammaticales du logiciel Open Source Unitex/Gramlab. La partie Apprentissage (profond) est implémentée en Python avec Keras et Tensorflow pour implémentés des réseaux neuronaux récurrents de type LSTM. L'environnement de développement est à ce jour Jupyter mais nous utilisons également Spyder (PyCharm à voir).

PROFIL

Qualités requises :

- Adaptabilité
- Autonomie
- Curiosité
- Force de proposition

- Implication

Compétences requises :
- Python
- Outils de type Jupyter
- Machine learning et Deep learning
- Bon niveau en anglais

MODALITES
Ce stage se déroule principalement à Orléans mais il est possible de travailler à distance car l'équipe est constituée à ce jour d'une linguiste (Orléans) et d'un architecte/développeur/… (Nantes). Il pourra donner lieu à une thèse en convention CIFRE sur le sujet de « l'explicabilité et la justesse des algorithmes de Deep Learning appliqués au Traitement Automatique de Langage ».
Durée du stage : idéalement 6 mois
Avantages :
- Possibilité de tickets restaurants
- Remboursement de 50% des transports en commun pour les trajets domicile / bureau
- Fruits, bonbons, cafés et thés à volonté !

CONTACTS :
Gabrielle Bosshard : Analyste Linguistique, 06.84.80.46.20
Eric Clairambault : Architecte, 06 07 39 14 65