Stage fin d'études pré-embauche : Ingénieur Développement Algorithme de clustering

La société :

Basée à Paris, GEOLSemantics est une jeune entreprise innovante qui développe et commerciale des logiciels de traitement automatique du langage naturel pour satisfaire des besoins de veille stratégique. Les applications sont multiples et principalement centrées sur la détection automatique de menace, de rumeur ou d'opinion à partir de l'analyse systématique des documents multi-lingues publiés sur le WEB, dans les sites, les blogs, les forums ou les réseaux sociaux.

Sujet du stage

GEOLSemantics recherche un stagiaire bac +5 en informatique pour la réalisation d'un algorithme de clustering. Intégré(e) au sein d'une équipe à taille humaine, vous aurez la possibilité de vous épanouir dans le monde du Web 3.0.

Contenu du stage : pré-embauche

De façon à évaluer la qualité des outils de regroupement d'information concernant des personnes (comme 123people), la communauté scientifique a organisé plusieurs compétitions internationales (campagnes WEBS). A partir d'un ensemble de textes qui concernent des personnes avec des homonymes (personnes différentes de même nom), une première phase consiste à regrouper les documents parlant de la même personne par des technologies de clustering et dans un deuxième temps pour chaque cluster qui concerne une seule personne (les homonymes étant sensé avoir été distingués par le clustering) on extrait un certain nombre d'information concernant la personne (date et lieu de naissance, adresse, téléphone, diplômes, …).

La société GEOLSemantics a développé des traitements linguistiques automatiques qui permettent d'identifier dans les textes les éléments significatifs et d'extraire des informations concernant les personnes.

En s'appuyant sur les outils existant dans la société, il est demandé au stagiaire de développer un algorithme de clustering qui s'appuiera sur les résultats de l'analyse linguistique des textes décrivant les personnes. Ensuite de compléter les règles d'extraction des connaissances pour extraire de chaque cluster les informations demandées par la campagne d'évaluation WEBS. Les résultats pourront être comparés aux résultats obtenus pas les participants aux campagnes précédentes.

La langue de la campagne étant l'anglais, il est demandé au stagiaire d'avoir une maîtrise de cette langue.

Vos capacités d'autonomie, vos aptitudes à vous exprimer et à rédiger vous permettront rapidement d'être responsable de la réalisation et du suivi complet de vos travaux sous la conduite d'un ingénieur expérimenté.

Le stage est placé sous la responsabilité du **Directeur Scientifique**

Profil du stagiaire :

Compétences techniques requises :

Profil demandé :

- Master 2 ou ingénieur en informatique ayant une connaissance en traitement automatique des langues.
- Goût de la compétition (il faut sortir de meilleurs résultats que les autres). Les résultats pourront faire l'objet d'une publication scientifique.
- Capacité à appréhender un domaine nouveau (le Web sémantique)
- Bonnes compétences en Java
- Maîtrise du XML et de RDF

Qualités requises : Bon rédactionnel. Forte curiosité fonctionnelle

Langue : Anglais exigé

- Durée : 6 mois minimum
- Début du stage : dès que possible
- Lieu : Paris 15ème
- Indemnités : à débattre

Candidature à adresser (CV, lettre de motivation, photo) à :
Emmanuel Dupont, Directeur des Opérations
Société Cadège/GEOLSemantics
32, rue Brancion
75015 Paris
emmanuel.dupont@geolsemantics.com

Site WEB : en cours de construction

Décembre 2011