



Sujet :

Analyse automatique de style et contenu sémantique des pages Web

Niveau : Bac + 4, Bac +5 – Master

Durée du stage : 4 à 6 mois

Contexte

Objectifs du stage

Dans le cadre d'une analyse linguistique des millions de pages du Web français, il faut développer un système d'analyse pour reconnaître et filtrer des textes tout-venant (blogs, page d'accueil, presse, Wikipedia) du Web en entrée, selon leurs caractéristiques stylistique. Il faut déterminer quelles caractéristiques sont pertinentes pour le filtrage, et les exploiter automatiquement.

Le stagiaire aura à développer un système robuste qui traitera une grande quantité de texte. Il faudrait que le stagiaire ait une bonne appréciation de différences subtiles de la langue française et puisse mettre en algorithme ces intuitions de façon imaginative.

Domaines de spécialité requis

Informatique

Moyens informatiques mis en œuvre

Langages : Langages de scripts (bash, perl...), html, sql, c++
Logiciels : LIMA, Linux, éditeurs de ressources

Autres moyens mis en œuvre (expériences, méthodes d'analyses, autres...)

Centre : Fontenay-aux-Roses

Pôle ou Direction : DRT - LIST

Dépt/Service/Labo : DTSI/SRCI/LIC2M :

Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue

Web : <http://www-list.cea.fr/>

Encadrement

Nom de l'ingénieur responsable : Kris JACK	Tél. : +33 1 46 54 82 47
E-mail : kris.jack@cea.fr	Fax : +33 1 46 54 75 80
Secrétariat : DTSI/SRCI	Tél. : +33 1 46 54 91 17
Nom du chef de laboratoire : Olivier Mesnard	Tél. : +33 1 46 54 73 38