

Sujet de Master 2 Recherche ou Pro

# Développement de méthodes de fouille de données pour l'analyse de textes

(Février - Juin 2011)

---

*Lieu :* UMR MIA (Mathématiques et Informatique Appliquées) à AgroParisTech  
16, rue Claude Bernard, 75005 Paris

*Direction :* Antoine CORNUÉJOLS (AgroParisTech) ([antoine.cornuejols@agroparistech.fr](mailto:antoine.cornuejols@agroparistech.fr))  
Christine MARTIN (AgroParisTech) ([christine.martin@agroparistech.fr](mailto:christine.martin@agroparistech.fr))

---

## 1. Sujet

On estime que l'univers digital contient actuellement plus que 1000 milliards de Giga-Octets. Sur chaque sujet, il existe ainsi, distribué sur le web, une énorme masse d'information qui pourraient être exploitée. Dès lors, un défi extraordinaire est de savoir découvrir les sources d'information intéressantes et d'en tirer des informations pertinentes.

Le projet dans lequel s'inscrit le sujet de stage proposé s'attaque à ce problème sur une échelle plus réduite. Il s'agit d'extraire automatiquement des informations sur le risque alimentaire associé à certaines substances à partir de textes tirés du web et portant sur ce sujet. Spécifiquement, en partant de textes annotés par des experts, le projet consiste à développer des méthodes d'apprentissage artificiel permettant, d'une part, de détecter les fragments de texte pertinents, et, d'autre part, de les annoter automatiquement. Outre leurs annotations, les experts fournissent des mots clés et des patrons de phrases qu'ils estiment indicatifs. Cependant, il semble aussi qu'ils mettent en œuvre une importante connaissance du domaine et de l'argumentation dans leur travail d'annotation. L'un des enjeux de l'étude proposée est de voir si des techniques relativement simples d'apprentissage peuvent permettre d'atteindre un niveau de performance similaire.

Le travail demandé concernera donc :

- En dehors de la mise au point de petits outils de scripts permettant de traduire des textes d'un format à un autre afin de faciliter les traitements ultérieurs, le stage se concentrera sur le développement et le test de méthodes d'apprentissage sur des corpus annotés afin de déterminer le niveau de performance accessible.
- Le test des mêmes méthodes sur des corpus nouveaux et le contrôle par des experts.

Étant donné l'enjeu de ce projet, le stagiaire bénéficiera d'un encadrement important et très intéressé. L'ampleur du travail et le niveau de sophistication des réalisations dépendront naturellement beaucoup de la motivation du stagiaire. Si les traitements de base à réaliser sont bien identifiés, il est clair que toute idée innovante sera examinée avec beaucoup d'intérêt.

Les développements informatiques seront réalisés si possible en Python.

## 2. Environnement de la thèse

Le stage se déroulera dans le laboratoire d'informatique de l'UMR MIA (Mathématiques et Informatique Appliquées) d'AgroParisTech au 16 rue Claude Bernard, 75005 Paris. Il sera rémunéré sur la base légale (environ 417 euros/mois).