

Sujet de Master 2 (Recherche)

# Catégorisation à partir de phrases

(Mars - Août 2010)

---

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

*Direction :* Antoine CORNUÉJOLS (AgroParisTech) ([antoine@lri.fr](mailto:antoine@lri.fr))

---

## 1. Sujet

Le sujet de stage proposé s'inscrit dans le contexte d'un projet d'intégration de données pour l'estimation de risque alimentaire et de décision (ANR « Holyrisk »). L'une des tâches de ce projet concerne la catégorisation de « bouts » de textes dans un petit nombre de classes (environ une vingtaine).

L'une des approches étudiées actuellement est d'avoir recours à des techniques d'apprentissage artificiel pour induire un classifieur à partir d'un échantillon d'apprentissage, c'est-à-dire de textes pour lesquels les catégories sont connues. On peut chercher directement une fonction de classification, ou bien une fonction calculant un rang pour les différentes catégories possibles. L'une des difficultés de ce problème d'apprentissage vient du fait que la plupart des techniques d'apprentissage ont été développées pour des données sous forme vectorielle, donc décrits à l'aide d'un nombre fixé de descripteurs. Or les textes sont des objets de taille variable.

Dans ce stage, on s'orientera *a priori* sur une approche par méthodes à noyaux. On étudiera des fonctions noyau adaptées pour la mesure de similarité dans le domaine concerné (e.g. « sacs de mots », n-grams, arbres, ...). Il faudra également résoudre le problème de la classification multi-classes.

Dans un deuxième temps, éventuellement, on étudiera des méthodes d'apprentissage incrémental en supposant que l'échantillon d'apprentissage s'agrandit ou évolue avec le temps.

Ce stage s'inscrit dans un axe prioritaire des équipes d'accueil et bénéficiera d'un plein soutien. Il est rémunéré.

## 2. Environnement de la thèse

La thèse se déroulera dans le laboratoire d'informatique de l'UMR MIA (Mathématiques et Informatique Appliquées) d'AgroParisTech.