

## Projet multi-disciplinaire M2 Mass Filtrage d'emails par techniques d'apprentissage automatique

### Description du sujet

Ce projet est à réaliser par un groupe de 2 étudiants.

### Objectif

Un des gros problèmes des messageries électroniques est la quantité de mails non désirés (spam) que chacun reçoit chaque jour. Il est donc intéressant de disposer d'outils informatiques capables d'identifier automatiquement la nature spam/non spams d'emails.

L'objectif de ce projet est de programmer un algorithme d'apprentissage automatique pour le filtrage d'emails. L'algorithme d'apprentissage qui sera implémenté est un algorithme d'apprentissage en ligne, c'est-à-dire qui s'adapte avec le temps aux informations qui lui sont fournies, proposé dans [1] (<http://citeseer.ist.psu.edu/629949.html>). Après avoir implémenté cet algorithme et mesuré son efficacité, des extensions en seront éventuellement proposées et programmées.

**Mots-clés** : filtrage d'emails, programmation, apprentissage en ligne.

### Phases du projet

Le projet se découpera selon les phases suivantes :

1. étude approfondie de la référence bibliographique et programmation de l'algorithme (langage au choix) ;
2. création d'une base d'emails pour l'apprentissage et codage de ces emails sous la forme de vecteurs « sac de mots » ;
3. évaluation de la qualité de l'algorithme ;
4. amélioration de l'algorithme.

### Contact

**encadrant** : Liva Ralaivola

**adresse** : LIF, CMI, 39, rue Joliot-Curie, 13013 Marseille

**email** : [liva.ralaivola@lif.univ-mrs.fr](mailto:liva.ralaivola@lif.univ-mrs.fr)

**tél** : 04 91 11 35 77

**url** : <http://www.lif.univ-mrs.fr/~liva>

### Références

- [1] K. Crammer, O. Dekel, S. Shalev-Shwartz, and Yoram Singer. Online Passive-Aggressive Algorithms. In *Adv. in Neural Information Processing Systems*, volume 16, 2004.