



FICHE DE PROPOSITION DE STAGE

Intitulé du stage : Détection de concepts et classification de documents appliquées à la détection de fraude

Domaine : Data Science

Contexte et déroulement du stage :

Stage à temps plein de 35h/semaine d'une durée de 6 mois.

Tracfin est un Service de renseignement rattaché au Ministère des Finances et des Comptes publics. Il concourt au développement d'une économie saine en luttant contre les circuits financiers clandestins, le blanchiment d'argent et le financement du terrorisme. Le Service est chargé de recueillir, analyser et enrichir les déclarations de soupçons que les professionnels assujettis sont tenus, par la loi, de lui déclarer.

Rattaché au chef de la Mission des Systèmes d'Information du Service, le stagiaire aura pour mission principale d'analyser un corpus de textes constitué des déclarations de soupçons, dans l'optique de faciliter leur orientation et d'aider au processus d'enquête.

Le stagiaire devra ainsi :

- Mettre en œuvre des méthodes non-supervisées afin de dresser une première cartographie du corpus, notamment des algorithmes de détection de concepts (topic models) permettant de repérer des classes thématiquement homogènes.
- Comparer la représentation des documents en « sacs de mots » avec les nouvelles approches de type neuronal (*word2vec*) pour l'annotation sémantique ou la catégorisation de documents.
- Constituer une base d'apprentissage pour la classification supervisée de futures déclarations et tester des méthodes de classification automatique (arbres de décision, réseaux de neurones, etc.)
- Réaliser un rapport final présentant ses résultats.

Compétences indispensables :

- Statistiques (ACP, analyse factorielle)
- Machine learning (clustering, classification, réseaux de neurones)
- Programmation en Python (pandas, scikit-learn, ...)
- Esprit d'initiative, autonomie
- Discrétion (le candidat retenu fera l'objet d'une habilitation « Confidentiel Défense »)

Compétences souhaitables :

- Intérêt pour le traitement automatique du langage et la fouille de textes
- Big Data : Spark, écosystème Hadoop

Contact : M. Rémi DE LA VIEUVILLE, Chef de la Mission Système d'Information,
remi.de-la-vieuville@finances.gouv.fr