

## **OUTIL DE SCRAPING DE SITES INTERNET**

### **ORIENTATION : LOGICIELS ET SYSTÈMES COMPLEXES**

#### **Descriptif :**

A l'ère d'Internet, la quantité d'information disponible en ligne a fortement augmenté. La récupération de ces données sur le web s'avère souvent longue et fastidieuse. Ce travail propose d'implémenter un outil de scraping automatisé permettant à l'utilisateur de décrire de façon naturelle les documents qu'il souhaite récupérer sur un site.

Ce dernier se veut le plus simple possible, de façon à rendre l'utilisation d'un tel système par n'importe quel utilisateur imaginable. Aucun prérequis en programmation ou en informatique générale n'est donc attendu de sa part. La plateforme doit être aussi flexible que possible, de façon à pouvoir fonctionner sur la très grande majorité des sites web.

Le travail présentera les prérequis à la réalisation d'une telle plateforme, l'architecture et principalement son implémentation.

#### **Travail demandé :**

Dans les limites du temps imparti, les tâches suivantes seront réalisées :

- Implémentation d'une interface utilisateur se basant sur le travail de semestre Inari;
- Création d'une extension navigateur dans le but de faciliter l'utilisation de la plateforme;
- Transformation automatique de la description des documents à récupérer fournie par l'utilisateur en code;
- Distribution de la récupération des documents à récupérer;
- Application de la plateforme sur des sites web dissemblables proposant des documents à télécharger;
- Mesures des performances de la plateforme.

Candidat :

**PIRKL THÉO**

Filière d'études : ITI

Professeur responsable :

**ORESTIS MALASPINAS**

**En collaboration avec :**

Travail de bachelor soumis à une convention de stage en entreprise : non

Travail de bachelor soumis à un contrat de confidentialité : non