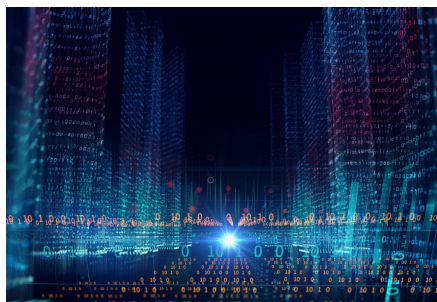


Résumé

Rester informé sur Internet devient de plus en plus difficile. Ce réseau continue à s'étendre et à proposer une quantité d'informations qui est impossible à assimiler par un être humain. Une solution est le traitement et la récupération en masse de documents, qu'on nomme scraping. Le but de ce projet est de proposer un logiciel de scraping, modulaire nommé Inari. Ce dernier doit être capable de se connecter à tout site, capable de résister aux erreurs et être utilisable par tous et toutes. Ce logiciel se basera sur l'infrastructure du même nom, créée lors du travail de semestre de septembre du même auteur. L'infrastructure existant déjà grâce au travail de semestre, ce travail se concentre principalement sur le rapprochement entre l'utilisateur et l'architecture en implémentant une interface graphique et des outils de connexion à l'utilisateur. L'infrastructure sera améliorée de façon à intégrer l'interface graphique dans les composants Inari. Pour vérifier si Inari répond à ces critères, ce travail étudie les performances de récupération de documents sur différents sites. La capacité d'utilisation par un utilisateur final est aussi étudiée. Les résultats semblent indiquer qu'un scraper peut accélérer dans des conditions moyennes de deux fois la récupération de documents par rapport à un humain. Au maximum, un gain de 1614% de vitesse a été observé. La vitesse d'Inari dépasse donc celle d'un humain même dans le pire des cas. La vitesse dépend non seulement du site mais aussi des paramètres appliqués à ce dernier : il a été observé que le site ralentit quand trop de documents sont téléchargés. Au vu de sa vitesse et de sa capacité à être utilisé par un utilisateur quelconque, il est théoriquement possible à Inari d'être utilisé en entreprise.



Candidat :

Théo PIRKL

Filière d'études : ITI

Professeur-e(s) responsable(s) :

Dr 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