# Rapport de fin de Sprint

## Sprint numéro : 2

Comme convenu avec le client, le deuxième Sprint consistait à extraire quatre informations d'un article scientifique. Notre équipe a concrétisé les attentes du client en réussissant à extraire les quatre informations suivantes :

- Le nom du fichier d'origine (dans une ligne),
- Le titre du papier (dans une ligne),
- Le ou les auteurs (dans une ligne),
- Le résumé ou « abstract » (dans une ligne).

Une part importante du travail a été l'achèvement de tests.

En effet, notre équipe a réalisé des tests unitaires afin de vérifier si notre programme qui analyse l'article scientifique capture correctement les informations correspondantes à ce que nous voulions. Par exemple, lorsque nous voulions prendre uniquement le titre, il devait y avoir une correspondance parfaite avec ce que nous attendions en résultat.

En conclusion des tests unitaires menés tout au long du Sprint 2, nous pouvons considérer que le programme est fonctionnel pour analyser, traiter et rendre le résultat attendu pour les quatre informations qui nous été demandé de traiter.

Nous avons aussi réalisé des tests de performance et de charge, notamment en essayant d'analyser les fichiers du corpus un par un ou en passant par des « threads » qui vont chacun analyser un fichier et donc diviser le nombre de fichiers à analyser par le nombre de « threads ». Concernant les tests de charge pour voir comment le programme réagissait et gérait une charge importante de fichier à examiner, un corpus de plus de 80 fichiers « .pdf » a été analysé plusieurs fois.

En conclusion des tests de performance et de charge, analyser le corpus en prenant les fichiers un par un prend autant de temps que passer par « threads », soit un peu moins d'une seconde. De plus, il n'y a pas de différence significative lorsque nous analysons un grand nombre de fichiers « .pdf ». L'analyse du grand corpus de plus de 80 fichiers prenaient entre 7 et 8 secondes pour les deux approches différentes.

Afin de diversifier nos analyses, nous avons jugé nécessaire de faire un deuxième corpus pour voir si nous pouvions rencontrer de nouvelles difficultés auxquels nous n'aurions pas penser. Créer un deuxième corpus a donc permis de mettre en évidence de nouvelles contraintes à traiter afin d'avoir un bon résultat en sortie.

En conclusion, le programme est robuste et efficace dans le traitement des 4 informations suivantes : le nom du fichier d'origine, le titre du papier, le ou les auteurs et le résumé ou « abstract ».