Rapport de TP Sprint 2 – Parseur d'articles scientifiques en format texte

---

Membres du groupe :

MEJAI
OUAHSOUNI
CHEMENTEL
JAHID
BOUDOUNT

Durée : 3h

Objectif du Sprint :

Développer un parseur pour analyser les articles scientifiques en format texte et extraire les informations essentielles : le nom du fichier d'origine, le titre du papier, et le résumé.

1. Planification du Sprint

Compte tenu de la contrainte de temps, notre sprint de 3 heures s'est concentré sur l'équivalent de "mêlées quotidiennes" (a intervalle de temps de 30 minutes) en continu pour assurer une communication efficace et des mises à jour en temps réel. Après une discussion initiale, nous avons décidé d'adopter une approche de "mêlée en continu" avec des mises à jour toutes les 30 minutes pour maximiser notre efficacité.

2. Choix du langage

Dès le début, l'équipe a discuté des compétences de chacun, de la familiarité avec les langages de programmation et des connaissances en programmation. Nous avons opté pour Python en raison de sa rapidité d'exécution et de la familiarité de l'ensemble de l'équipe avec ce langage.

3. Attribution des tâches

Nous nous sommes concentrés sur trois tâches principales pour ce sprint :

- Parsage du titre
    - Attribué à : JAHID et BOUDOUNT
    - Tâche : Extraire le titre de l'article à partir du texte converti.
- Parsage du nom du fichier
    - Attribué à : CHEMENTEL
    - Tâche : Identifier et extraire le nom du fichier d'origine.
- Parsage du paragraphe Abstract
    - Attribué à : MEJAI et OUAHSOUNI
    - Tâche : Identifier et extraire le résumé de l'article.

Pour les autres tâches :

- Gestion des exceptions
    - Attribué à : JAHID
    - Tâche : Gérer les scénarios où certaines sections pourraient manquer.

- Création/effacement du sous-dossier
    - Attribué à : BOUDOUNT
    - Tâche : Écrire le code pour créer/effacer le sous-dossier.
- Intégration avec Github et Rédaction du README.md
    - Attribué à : MEJAI, OUAHSOUNI, BOUDOUNT, CHEMENTEL et JAHID
    - Tâche : Gérer les branches, commits et rédiger une documentation pertinente pour le projet.

4. Mêlée en continu

Compte tenu de la durée limitée du sprint, l'équipe s'est mise à jour en continu, équivalent à une mêlée quotidienne, pour s'assurer que tout le monde était aligné sur les objectifs et les progrès.

5. Critères d'acceptation

Le programme doit correctement extraire le titre, le nom du fichier et le résumé.

Il doit gérer les exceptions.

Il doit fonctionner sur GNU/Linux en ligne de commande.

Revue du Sprint

À la fin de notre sprint de 3 heures, l'équipe a réussi à accomplir la majorité des tâches. Les intégrations Github ont été réalisées et le README.md a été rédigé avec succès.

L'extraction des abstracts des fichiers PDF composés de colonnes s'est avérée être un défi majeur. Malgré nos efforts pour ajuster notre méthode de parsage, la structure des colonnes a entravé l'extraction précise du contenu. Nous avons exploré quelques approches pour surmonter ce problème (extraire un paragraphe jusqu'à au mot clé « Introduction » ou encore utiliser l'argument « -layout » lors de la conversion avec pdftotext), mais étant donné la contrainte de temps, une solution complète nécessiterait une exploration plus approfondie dans un sprint ultérieur.

Rétrospective du Sprint

<u>Points forts :</u>

- Une communication ouverte et continue.
- La division des tâches était claire et alignée sur les compétences de chacun.

<u>Axes d'amélioration :</u>

Même si nous avons fait de notre mieux étant donné les contraintes de temps, une planification plus approfondie aurait pu améliorer l'efficacité.

En somme, malgré les défis liés à la contrainte de temps, l'équipe a collaboré efficacement pour atteindre les objectifs du sprint. Une attention particulière a été accordée à la communication et à l'attribution des tâches en fonction des compétences de chacun.