scrum master : Pascal Flores
product owner : Sébastien Ré
équipe de développement : les deux

Sprint Backlog :

- comparer les options disponibles de chaque outil
- arriver à lancer chaque outil
- convertir les textes du corpus en texte avec chaque outil avec les bonnes options
- mettre les fichiers texte sur github, chacun sa branche
- merge les branches
- comparer les résultats des outils (pdftotext vs pdf2txt)

notre priorité est de satisfaire le client donc nous allons privilégié pour avoir un logiciel opationnel dès que possible que l'on perfectionnera

Lequel de 2 logiciels est le plus adapté ? Quels options ?

critères d'évaluation de l'outil :

- doit être capable de convertir le texte en .txt
- gestion du texte en double colonne
- doit être capable de lire les formules
- capable d'identifier le titre, les auteurs, l'introduction, le développement (les différentes parties en somme)

Description de l'outil pdftotext :

- est très rapide ~100ms en moyenne
- option -layout qui permet de rendre un doc semblable à l'original
- pas encore trouvé comment gérer les formules mathématiques

Description de l'outil pdf2txt :

- Dispose de beaucoup d'options différentes, notamment une nommée all-texts, qui est censée récupérer même le texte contenu dans les formules mathématiques etc
- option nécessaire pour écrire dans un fichier, sinon écrit sur la console directement
- est écrit en python, sera probablement assez lent
- lors d'une mauvaise utilisation de l'outil, affiche une aide sommaire et incomplète, il faut utiliser l'option --help pour une liste complète des commandes

Comparatif des performances:

- pdf2txt : ~1.5s pour le fichier ACL2004-HEADLINE
- pdftotext : ~100ms pour le même fichier

Comparatif des résultats :

- pdf2txt : résultats sur une seule colonne, formules non présentes, séparateurs de lignes coupant les mots, lignes de texte correspondant sans doute à de la mise en forme du format pdf, etc…

- pdftotext : fichier texte ressemblant au pdf original, pas de mots coupés ou manquants, formules imparfaites mais lisibles

conclusion : En plus d'être plus rapide, l'outil pdftotext est bien meilleur en termes de résultats que l'outil pdf2txt, il est celui à sélectionner pour la suite.

Lien du dépôt git : https://github.com/Pascal-Flores/PDF-To-Text-CLI