Rapport de Sprint 1 – Parseur d'articles scientifiques en texte brut

Lucas, Gautier, Gael, Ilias, Quentin

Mars 2025

Objectif du projet

Le projet consiste à comparer plusieurs outils de conversion open source, à sélectionner la meilleure solution, puis à créer un script capable d'extraire le texte et les métadonnées pertinentes.

Le tout est développé dans un environnement GNU/Linux, en ligne de commande, dans une approche agile selon la méthodologie SCRUM.

Méthodologie SCRUM et répartition

Notre équipe est composée de 5 membres :
— **Lucas et Gautier** ont travaillé sur la solution basée sur `pdf2txt.py` (librairie pdfminer.six en Python).
— **Gael, Ilias et Quentin** ont exploré `pdftotext` (outil en ligne de commande via Poppler).
Chaque membre a proposé un script, tous les essais ont été regroupés dans le dossier `tests_technologies/`.
La version finale choisie est placée dans le dossier `Final_version/`.

Tests des technologies

Nous avons comparé deux outils :

1. pdf2txt.py

Commande utilisée : `pdf2txt.py fichier.pdf > fichier.txt`
Avantages :
— Outil Python pur (intégrable dans un pipeline)
— Possibilité de sortie XML ou HTML
Inconvénients :
— Très mauvaise gestion des articles en colonnes
— Espaces manquants, mots collés, désordre du texte
— Peu lisible sans post-traitement important

2. pdftotext (Poppler)

Commande retenue :

```
pdftotext -layout -enc UTF-8 -nopgbrk -eol unix fichier.pdf fichier.txt
```

Avantages :

— Excellente gestion de la mise en page multi-colonne
— Texte bien découpé, phrases lisibles
— Encodage UTF-8 et retour à la ligne propres

Inconvénients :

— Nécessite une installation système (Poppler)
— Pas de support XML/JSON ou structuration intégrée

Choix final

Au vu des résultats, nous avons retenu **pdftotext** avec les options suivantes :

```
-layout -enc UTF-8 -nopgbrk -eol unix
```

Elles garantissent une conversion fidèle à la mise en page d'origine, indispensable pour traiter des articles scientifiques à colonnes.

Fonctionnalités de la version finale

— Conversion de tous les fichiers PDF d'un dossier donné
— Génération d'un fichier texte par article

Conclusion

Notre système est maintenant capable de générer automatiquement des fichiers texte structurés à partir de PDF scientifiques. La mise en page est conservée et les sections sont clairement séparées. Le parseur est prêt à être utilisé dans un pipeline de TAL ou d'analyse scientifique automatisée.

Dépôt GitHub du projet : https://github.com/Conception-Logicielle/Parseur-articles-scientifiques-en-format-texte