📕

Rapport 1 - Analyse des outils

Informations générales :

Équipe : Walid Medouaz / Abdelghani Rechidi / Hicham Guessab

Maître scrum : Hicham Guessab

Lien Gitlab : https://gitlab.com/HichamGsb/pdftotxt

Les outils :

📌 pdftotext

> *Pour chaque fichier pdf, on le converti en txt en gardant sa mise en page (layout).*

```
# Convertir un seul fichier pdf en txt
pdftotext -layout fichier.pdf

# Convertir tous les fichiers pdf d'un répertoire en txt
for file in *.pdf; do pdftotext -layout "$file"; done
```

Options :

-layout garder la même mise en page.

📌 pdf2txt

> *Pour chaque fichier pdf, on le converti en précisant le nom (-o) et le type de fichier souhaité (-t).*

```
# Convertir un seul fichier pdf en txt
pdf2txt -o fichier.txt -t text fichier.pdf

# Convertir tous les fichiers pdf d'un répertoire en txt
for file in *.pdf; do pdf2txt -o "$file".txt -t text "$file"; done
```

Options :

-o renommer le nom des fichiers en sortie

-t type de fichier

Tout d’abord, certaines imperfection sont communes aux résultats des deux outils.

Il s’agit tout d’abord de certains signes mathématiques qui ne sont pas du tout écrits comme la somme, la division et les indices. De plus, les graphiques et les images ne sont pas du tout traités. On pourrait les afficher grâce à des points.

D’une part, l’outil **pdf2txt** n’est pas très approprié à l’utilisation de LIA. En effet, la mise en page n’est pas conservé, les données sont écrites à la ligne avec souvent un saut à la ligne très exagéré, qui ne sert à rien. De plus, certains signes mathématiques ne sont pas affichés et sont remplacés par des “cid:nbr”.

Les tableaux ne sont pas du tout mis en forme, ils sont quasiment illisibles, une chose très inappropriée pour un laboratoire scientifique.

Parfois, la conversion des fichiers est mal réalisée, ainsi toutes les lignes sont collées et illisibles.

D’autre part, l’outil **pdftotext** permet une présentation plus propre pour la plupart des documents. En effet, la présentation des documents (layout) est quasiment conservée, une aération des paragraphe est également notable ainsi qu’un saut de page marqué par le nom des pages.

Nous pouvons aussi remarquer un certain effort non négligeable par rapport à l’outil pdf2txt où par exemple, même si c’est assez mal réalisé, nous conservons une double page pour le fichier “*stolcke_1996_Automatic_linguistic.txt”.*

La mise en forme des tableaux est conservée même s’il reste des imperfections et qu’il n’y a pas de bordures, le tableau reste bel et bien compréhensible.

Les chiffres des grandes parties ne sont pas affichés (affichage de l’encodage au lieu du chiffre ex : FF au lieu de 2 pour le fichier “*probabilistic_sentence_reduction.txt”*.

Nous pouvons conclure que l’outil **pdftotext** est beaucoup plus approprié à l’utilisation du LIA étant donné le besoin scientifique et d’aération et malgré les quelques défauts qu’il peut proposer.