SCRUM

1. Pas de style (stylisation)

- **PDF** : Texte avec mise en forme (gras, italique, tailles, alignements...).
- **TXT** : Le texte est brut, **aucune mise en forme** n'est conservée.

Exemple : les titres de sections (comme "Abstract" ou "1 Introduction") apparaissent sans distinction (pas de gras ou indentation).

---

2. Alignement toujours à gauche

- Le texte dans le `.txt` est **entièrement aligné à gauche**.
- Dans le PDF, certains éléments sont centrés (titres, auteurs) ou justifiés.

---

3. Les parties restent collées (pas de séparation visuelle nette)

- Il n'y a **aucune séparation** claire entre les sections.
- Exemple : l'**auteur**, le **titre** et l'**abstract** sont souvent collés sans saut de ligne net.

---

4. Disparition ou mauvaise interprétation de symboles

- Des symboles comme `{}` **, les guillemets " ", ou les flèches →** peuvent :
  - Disparaître
  - Être remplacés par des caractères bizarres (`(cid:0)` ou `(cid:1)` par exemple)

---

5. Pas de soulignement, pas de colonnes, pas de tableau

- Le texte converti **perd complètement les lignes, les colonnes**, et **les tableaux**.
- Les éléments sont **plats et désordonnés** dans le `.txt`.

Exemple : certaines colonnes sont lues complètement avant de passer à la suivante, ou une ligne de chaque colonne est lue successivement.

---

6. Les numérotations sautent parfois des lignes

- Exemple :
  - Dans le PDF : les listes sont bien structurées (`1.`, `2.`...).
  - Dans le `.txt` : il y a parfois des sauts de ligne ou des numéros mal positionnés.

---

7. Les tableaux sont très mal rendus

- Le convertisseur lit parfois :
  - **Toutes les lignes d'une colonne**, puis les colonnes suivantes.
  - Ou **les en-têtes d'abord**, puis un texte désordonné.

---

8. Perte de la structure de paragraphes

- Dans le PDF, les paragraphes sont **clairement séparés par des espaces** ou indentation.
- Dans le `.txt` :
  - **Tout est "plat"**, on a du mal à voir les débuts et fins de paragraphes.
  - Parfois, des mots sont collés entre deux lignes.

---

9. Formules ou caractères spéciaux mal convertis

- Parfois, des **formules mathématiques ou références** sont **représentées par des codes** comme :

```
scss
CopierModifier
(cid:3)(cid:5)(cid:4)(cid:7)(cid:6)
```

le signe de somme n'exist pas

10. certain fichier sont incompréhensible

Example : le fichier Boudin-Torres-2006

> A Scalable MMR Approach to Sentence Scoring
> for Multi-Document Update Summarization

> Florian Boudin \ and Marc El-Bèze \\
> Laboratoire Informatique d’Avignon
> ...

est transferé à

> AScalableMMRApproachtoSentenceScoringforMulti-DocumentUpdateSummarizationFlorianBoudin\andMarcEl-B`eze\\LaboratoireInformatiqued’Avignon...

11. Quelques différence dans l’ecriture

- Un simple mot qui est “Artificial “ esr transformé comme si dessus

Artiï¬cial

•les petites 1 , 2 en haut ça transforme en 1 et 2 normales
la lettre F ça transforme en ça

ï¬

•quand il yas les doubles quottes ça transforme en ça

â€œparse-and-trimâ€

•le å se transforme en ça dans le mot Mårdh

MËšardh

•le - est transformé en ça â€”

•le , est transformé en ça â€™

• le • est transformé en ça â€¢

12.le code Pour Pdf2text utilisée :

```
import os
# Dossiers d'entrée et de sortie

input_folder = "pdf"

output_folder = "output_pdf2text"

# Créer le dossier de sortie s’il n’existe pas

os.makedirs(output_folder, exist_ok=True)

def convertir_pdf_en_txt(pdf_path, txt_path):

commande = f"pdf2txt.py \"{pdf_path}\" > \"{txt_path}\""

print(f"Conversion de {pdf_path} en {txt_path}...")

os.system(commande)

print("Conversion terminée.")

# Parcourir tous les fichiers dans le dossier "pdf"

for filename in os.listdir(input_folder):

if filename.endswith(".pdf"):

pdf_path = os.path.join(input_folder, filename)

txt_filename = filename.replace(".pdf", ".txt")

txt_path = os.path.join(output_folder, txt_filename)
```

```
convertir_pdf_en_txt(pdf_path, txt_path)
```

II. Avec Pdftotext :

1. Code utilisée

```
import subprocess
import os
# Dossiers (relatifs au dossier courant)

input_dir = "CORPUS_TRAIN"

output_dir = "output"

# Créer le dossier de sortie s’il n’existe pas

os.makedirs(output_dir, exist_ok=True)

# Liste des fichiers PDF dans CORPUS_TRAIN

for filename in os.listdir(input_dir):

if filename.endswith(".pdf"):

input_path = os.path.join(input_dir, filename)

output_path = os.path.join(output_dir, filename.replace(".pdf", ".txt"))

# Commande pdftotext

command = [

"pdftotext",

"-layout",
```

```
"-enc", "UTF-8",

"-nopgbrk",

input_path,

output_path

]

# Exécution

print(f"Conversion de {filename}...")

result = subprocess.run(command, capture_output=True, text=True)

if result.returncode != 0:

print(f"❌ Erreur avec {filename} : {result.stderr}")

else:

print(f"✔ {filename} converti avec succès.")
```

2. différences trouvée dans la conversion :
   perte de mise en forme (comme le style de texte et taille)
   perte des éléments graphiques
   les tableaux devient en txt avec les TAB
   caractere speciaux(”,;-_à) ils sont supprime ou transforme en caractere special
   parfois mal place les sauts de ligne
   perte les formule mathematique comme valeur absolue ,division
   perte les grand parenthese

mal ecrit les termes mathematique comme terme C0 devient c0