



UNIVERSITÉ D'AVIGNON
ET DES PAYS DE VAUCLUSE

C E N T R E
D'ENSEIGNEMENT
ET DE RECHERCHE
EN INFORMATIQUE



Licence Informatique

ILSEN
UE Génie Logiciel

Projet SCRUM: Séance 1

Groupe 2

Enzo GUENY
Adam SERGHINI
Jarod DURET

CERI - LIA
339 chemin des Meinajariès
BP 1228
84911 AVIGNON Cedex 9
France

Tél. +33 (0)4 90 84 35 00
Fax +33 (0)4 90 84 35 01
<http://ceri.univ-avignon.fr>

Encadrement

Juan Manuel TORRES MORENO

Sommaire

| | |
|----------------------------------|---|
| Titre | 1 |
| Sommaire | 2 |
| 1 Introduction | 3 |
| 2 Choix de l'outil de conversion | 3 |
| 2.1 pdftotext | 3 |
| 2.2 pdf2txt | 3 |
| 2.3 Conclusion | 3 |
| 3 Choix du langage | 3 |

1 Introduction

Ce document est un rapport sur l'avancement du projet et les étapes réalisées lors de la première séance du projet **UAPV TP de Génie logiciel - Scrum** le 28/11/2018.

On commence par lire le sujet et récupérer les ressources disponibles.

On définit le SCRUM Master (DURET Jarod).

Création du dépôt GitHub : -> https://github.com/Team-Rocket-CERI/SCRUM_Project

2 Choix de l'outil de conversion

2.1 pdftotext

Test :

```
pdftotext -enc UTF-8 -layout article.pdf
```

» article.txt

enc détermine l'encodage adapté.

layout maintient la coupe originale du document pdf.

Il s'est avéré que le résultat en sortie, stocké dans le txt, présente peu d'erreurs de conversion. Malgré quelques problèmes mineurs sur les formules mathématiques.

2.2 pdf2txt

Test :

```
pdf2txt -c utf-8 -o output.txt article.pdf
```

output.txt est le fichier texte de destination

article.pdf est le pdf à parser.

Il s'est avéré que le résultat en sortie, stocké dans le txt, présente beaucoup d'erreurs de conversion. Les formules mathématiques sont mal traduites, problèmes d'espacement, mauvais décodage des caractères.

2.3 Conclusion

Nous avons choisis d'utiliser **pdftotext**. Ce dernier garantit une conversion bien plus fidèle et complète que son concurrent.

Cependant quelques inconvénients persistent ; formules mathématiques : perte de signalisation, et certains caractères.

3 Choix du langage

Nous avons choisis d'utiliser **Rust** car cet outil nous permet d'obtenir des performances similaires au C tout en éliminant les erreurs liées à la gestion de la mémoire. De plus, dans le cadre de traitement de gros fichiers, Rust nous permet de facilement faire de la programmation concurrente afin d'obtenir des performances optimales. Pour finir, ce choix nous sort de notre zone de confort en nous confrontant à l'utilisation d'un langage que nous ne connaissons pas.

Le développement se fera intégralement sur Linux pour garantir le bon fonctionnement de l'exécution des scripts sur ce dernier.