Projet de développement

Sprint 1

Equipe : Romain Le Talour, Corentin Leconte, Thibaut Quilleré.

Pour ce projet de développement, la méthodologie de travail est basée sur l'agile avec une approche SCRUM.

Maitre SCRUM : Romain Le Talour

Ce premier sprint est basé sur la mise en place de l'équipe, du test des parser pdftotext et pdf2txt ainsi que le résumé de leurs différentes options. (Commande *man* utilisée ici).

Voici un rapport des différentes options proposées par ces deux outils :

1) PDFTOTEXT

Pdftotext est un convertisseur de fichier PDF en fichier texte via le terminal par ligne de commande.

Les options sont :

-f *nombre* Spécifie la première page à convertir.

-l *nombre* Spécifie la dernière page à convertir.

-r *nombre* Spécifie la résolution en DPI. Par défaut est à 72.

-x *nombre* Spécifie la coordonnée x du coin supérieur gauche de la zone de recadrage.

-y *nombre* Spécifie la coordonnée y du coin supérieur gauche de la zone de recadrage.

-W *nombre* Spécifie la largeur de la zone de recadrage en pixels. Valeur par défaut à 0.

-H *nombre* Spécifie la hauteur de la zone de recadrage en pixels. Valeur par défaut à 0.

-layout Conserve le mieux possible la mise en page originale du texte. La valeur par défaut consiste à annuler la disposition des colonnes, coupures de mots etc., et à afficher le texte dans l'ordre de lecture.

-raw Garde le texte dans l'ordre du flux de contenu. Il s'agit d'un hack qui ne défait le formatage des colonnes etc. Utilisation non recommandée.

-htmlmeta Génère un simple document HTML, y compris les méta-informations (encapsule les <pre> et </pre> dans le texte.

-enc Definit l'encodage à utiliser pour la sortie texte. *UTF-8* est utilisé par défaut.

-listenc Fait la liste des encodages disponibles.

-eol Sert à définir la convention de fin de ligne à utiliser pour la sortie texte.

-nopgbrk N'insère pas les sauts de page ni même les caractères de saut de page.

-opw *mot de passe* Spécifier un mot de passe du propriétaire. Détourne toutes restrictions de sécurité.

-upw *mot de passe* Spécifier un mot passe utilisateur.

-q N'affiche rien, ni même les erreurs.

-v Affiche le numéro de version et le copyright.

-h équivalent à -help ou --help.

2) PDF2TXT.PY

Tout comme PDFTOTEXT, il permet de parser un document PDF en document texte.

-P *password* : Mot de passe du PDF.

-o *output* : Pour définir le nom du fichier en sortie.

-t *text*|*html*|*xml*|*tag* : Type de sortie. (Par défaut : automatiquement déduit du nom du fichier de sortie.)

-O *output_dir* : Répertoire de sortie pour les images extraites.

-c *encoding* : Encodage de sortie. (Par défaut : UTF-8)

-s *scale* : Echelle de sortie.

-R *rotation* : Effectue une rotation de la page en degrés.

-Y *normal*|*loose*|*exact* : Spécifie le mode de mise en page. (Uniquement pour la sortie HTML.)

-p *pagenos* : Ne traite qu'un nombre de pages défini.

-m *maxpages* : Limite un nombre maximum de pages à traiter.

-S : type de contrôle de bandes.

-C : Mise en cache des ressources désactivée.

-n : Désactive l'analyse de la disposition.

-A : Applique une analyse de mise en page à tous les textes, y compris les figures.

-V : Détecte automatiquement l'écriture verticale.

-M *char_margin* : Spécifie la marge du caractère.

-W *word_margin* : Spécifie la marge du mot.

-L *line_margin* : Spécifie la marge de la ligne.

-F *boxes_flow* : Spécifie le rapport de débit de la boîte.

-d : Turns on Debug output.

Ces deux outils ont chacun leurs avantages et inconvénients.

Avantages et inconvénients de PDFTOTEXT :

- Est flexible d'utilisation, beaucoup d'options.
- Facile d'exécution, pas besoin d'installation particulière.
- Beaucoup de sorties supportées ainsi que d'encodages.
- Peut parfois être imprécis et ne pas supporter les caractères spéciaux

Avantages et inconvénients de PDF2TXT.PY :

- Très efficace dans la détection de la localisation du texte des informations de mise en page.
- Automatique dans l'analyse de la disposition
- Beaucoup de sorties supportées ainsi que d'encodages.
- A des options de sécurité cryptographique.
- Nécessite d'installer python sur sa machine pour faire fonctionner le programme, plus fastidieux à prendre en main.
- Assez peu documenté.