**Projet Python**

Lebreton Flavien
Gerault Thomas

Le programme est plus optimal si les fichiers PDF sont renseignés en premier et ensuite les HTML. Et il ne traite pas le cas où le fichier PDF n'a pas d'image quand on lui demande d'extraire des images.
À chaque ouverture, il demande à l'utilisateur de rentrer le chemin du document texte et du fichier qu'il aura créé préalablement. Il ne pourra pas traiter en cas de faute de l'utilisateur sur les chemins.
Le programme propose 2 choix : « Nuage » qui est le fait de faire un (ou des) nuage(s) de mots avec les documents et « Image » qui traite les images des documents afin de les exporter sur un fichier HTML ouvert sur une page web.
Si l'utilisateur veut garder la page HTML des résultats en faisant d'autres programmes, il doit renseigner un nouveau répertoire.

Classe Ressource :
Cette classe prend un lien pour un HTML ou un PDF. Premièrement, elle retourne si le document est un PDF ou un HTML avec la fonction type et ensuite elle retourne l'adresse sous forme d'un texte avec la fonction text.
Pour prendre le lien pour un document PDF, nous avons utilisé le programme d'Ilyas Tasli et Théo Legruel.
Pour que la classe ressource marche parfaitement, il faut que la fonction type soit faite pour que la fonction text fonctionne.
Il est possible que l'utilisateur est besoin d'installer les bibliothèques beautifulsoup4 et pdfminer.

Classe Collecte :
Cette classe prend plusieurs liens HTML ou PDF. Elle récupère les ressources en leur extrayant leur texte et elle renvoie une liste de contenus sous forme d'un tableau.

Classe Traitement :
Cette classe nous sert à traiter les documents HTML et PDF et à leur extraire les images sur un document HTML.
La fonction load_html permet de créer une page HTML dans un fichier demandé par l'utilisateur. Ensuite elle rajoute les images, dont les liens et chemins sont enregistrés dans une liste, sur cette page.
La fonction enregistrement_image_html permet d'extraire les images d'un document HTML sous forme d'une liste de lien. Chaque lien représente une image.
La fonction enregistrement_image_pdf permet d'extraire les images d'un document PDF sous forme d'une liste de chemin. Dans cette fonction, les images sont enregistrées sur le fichier prédéfini par l'utilisateur afin de pouvoir enregistrer leurs chemins.
Il est possible que l'utilisateur est besoin d'installer les bibliothèques os, pthlib,urllib et fitz.

Classe Prisme :
Cette classe est la principale. Elle prend un fichier txt contenant les liens des PDF et HTML, extrait les liens et traite les documents. Elle traite soit les images en les extrayant des documents ou soit les mots en faisant un nuage de mots. Pour finir elle exporte les résultats sur une page web.

La fonction run permet d'extraire les liens présents dans le fichier txt renseigné par l'utilisateur et enregistre les liens dans une liste (utilisée pour « Image »). Ensuite il utilise la classe Collecte pour extraire le texte des documents en le sauvegardant sur une nouvelle liste (utilisée pour « Nuage »).
La fonction show permet d'extraire les mots sous un(ou des) nuage(s) de mots ou les images dans une page HTML ouverte sur une page web. Premièrement, on supprime tout le contenu présent dans le fichier renseigné par l'utilisateur pour éviter tout problème avec d'anciens documents et pour éviter de prendre trop de place sur l'ordinateur.
Si l'utilisateur choisit « Nuage », il doit renseigner le maximum de mot présent dans le nuage de mot et s'il veut un nuage de mots global ou un pour chaque document. Lorsque qu'il y en un seul de nuage de mots, on ouvre directement l'image sur une page web. Mais lorsque qu'il y en a plusieurs, on enregistre les images sur une page HTML afin d'exporter le résultat sur une page web.
Si l'utilisateur choisit « Image », alors le programme va extraire les images sous forme d'une liste de liens d'image et les enregistre sur une page HTML afin d'exporter le résultat sur une page web.
Nous avons décidé de séparer le cas où le document est un HTML ou un PDF et pour traiter les documents on utilise la classe Traitement.