Projet de Développement

INF1603 réalisé

13 mars 2021

Table des matières

Sprint 3 - Rapport

13 mars 2021

1 Explications sur l'analyse des articles

1.1 Les références

Le format des références des articles de recherche n'est pas toujours normalisé.

Nous avons pu récupérer l'intégralité des références, mais la mise en forme sous une ligne unique par référence ne fonctionne pas toujours, car il n'existe pas de règle universelle permettant d'identifier les références une à une.

Par ailleurs, il existe certaine liste de références avec des codages caractères différents (MacOS, UTF-8, Windows, ...) et c'est pourquoi nous nous retrouvons de temps en temps avec plus espaces avant le début de chaque références.

1.2 Les adresses email

Nous avons choisi de filtrer directement les adresses emails à partir de la 1ère page du fichier pdf, transformé en texte.

Ce filtrage nous permet de récupérer une liste d'adresses emails, transformée en chaine de caractères, ajoutée à la fin de notre chaine "auteur".

Comme pour les références, il existe trop de formats différents d'adresses emails pour trouver une règle unique permettant d'identifier 100% des adresses emails.

Notre règle est efficace à 98%.

2 Les formats de sortie

À travers les options `-t` ou `-x`, nous pouvons générer des sorties TXT ou XML comme demandé.

L'option `-a` permet d'avoir les deux sorties simultanément.