La minute du PSIR

2 juin 2021 - # 24

OCR (Optical Character Recognition)

NIVEAU DE DIFFICULTÉ

OCR

PERSONNES CONCERNÉES

Cette semaine la Minute du PSIR vous propose une présentation de l'OCR ou ROC pour Reconnaissance Optique de Caractères. Ce procédé consiste à transformer le contenu d'un document numérisé dactylographié en fichier texte pour pouvoir le manipuler, le modifier ou faire des recherches à l'intérieur. Pour les contenus manuscrits, bien plus complexes à transformer, il existe le HTR (Handwritten Text Recognition) souvent réalisé par le biais de réseaux neuronaux.

Nous sommes chaque jour confrontés dans les projets de recherche à la réutilisation de documents anciens, qui ne sont pas tous nativement numériques ou bien dont la version numérique n'est pas disponible. Cependant leur contenu est essentiel au projet, entre donc en jeu l'OCR dont voici le fonctionnement général :

Le fichier d'entrée : il est nécessaire de disposer d'un fichier numérisé de qualité (tif, jpeg, png, pdf etc.). Il doit être lisible, sans trop de traces parasites, si possible sans annotation manuscrite, le texte doit être bien distinct du fond et très contrasté pour obtenir de bons résultats. Le logiciel procèdera tout d'abord à une analyse d'image pour la redresser si besoin.

L'apprentissage : il faut apprendre au logiciel à reconnaitre les différents éléments de la page (n° page, titre, tableau, note de bas de page, illustration…). Puis les caractères en eux-mêmes (sélectionner la ou les langues du document, etc.). Même si tout logiciel ou service dispose de dictionnaires embarqués, il faut vérifier, sur quelques pages, que la reconnaissance est correcte en validant l'interprétation de départ.

La correction : Une fois le traitement terminé, il y a toujours une étape de correction et de vérification pour corriger les erreurs, mauvaises interprétations... On estime qu'un résultat d'OCR est totalement satisfaisant quand son taux de réussite dépasse les 96% sur l'ensemble du document. Il est alors possible d'exporter le résultat sous forme de fichier texte. Abobe Acrobat permet un OCR plus léger, il rajoute une couche d'informations sur le fichier d'entrée et permet uniquement de rechercher au sein d'un pdf, sans export possible du texte.

Initiatives participatives : Il existe également des initiatives de science participative (crowdsourcing) qui mettent à contribution des personnes bénévoles pour effectuer le travail de correction d'OCR rendant ainsi le travail beaucoup plus efficient (Monasterium Collaborative Archive, Bulliod, etc.).

Différents outils et solutions d'OCR et HTR :
ABBYY FineReader (OCR) payant
Kraken (OCR) gratuit
Adobe Acrobat (OCR) payant
eScriptorium (HTR) gratuit

PDF *MINUTE_PSIR_#24*