

Post-correction d'OCR pour les documents administratifs

Mots clés : OCR, post-OCR, deep learning

La société EasyChain est une jeune société d'édition d'outils utilisant l'Intelligence Artificielle basée à Niort. Si ses premiers clients sont les courtiers en crédit immobilier, la société adresse toute la sphère immobilière qui va de l'acheteur, au vendeur, en passant notamment par les notaires, agents immobiliers, banques et huissiers. Avec notamment l'arrivée des milleniums sur le marché du travail, les métiers liés à l'immobilier vont connaître d'importantes mutations au cours des prochaines années : la digitalisation et l'intelligence artificielle en sont les principaux enjeux. Afin de répondre à ces enjeux futurs, des premiers outils ont été développés par la société :

- un outil de reconnaissance et de classification de documents standard et non standard, avec extraction d'informations importantes, alimentant entre autres, un CRM métier ;
- une application mobile permettant un transfert d'information orale et écrite, chiffré de bout en bout. C'est un « WhatsApp » très augmenté, dédié à la sphère du crédit immobilier.

L'outil de reconnaissance et de classification de documents a besoin d'effectuer d'une OCR (Optical Character Recognition) pour extraire le contenu textuel dans tous les pages. En se basant sur ces contenus, un modèle d'apprentissage a été construit pour pouvoir classifier les pages en différents types. Cependant, lors de la phase d'OCR, il s'agit des erreurs à cause de la qualité de la page, ou erreur du moteur d'OCR. Ses erreurs ont beaucoup impacté sur le résultat du modèle de classification et il faut les corriger.

L'enjeu du projet est de comprendre comment des connaissances linguistiques peuvent aider à optimiser automatiquement ou quasi-automatiquement la qualité orthographique des textes, avec des objectifs proches du **zéro-faute** en sortie, tout en maintenant un temps de traitement très bas compatible avec les très gros volumes de données à traiter.

Plus précisément, l'objectif est d'exploiter des connaissances, outils et ressources dédiées à l'analyse linguistique de surface (étiquetage morphosyntaxique, analyse morphologique, connaissances lexicales, et autres) pour identifier automatiquement les (rares) erreurs orthographiques issues de l'OCR (ou d'autres méthodes de capture textuelle) et pour les corriger automatiquement (ou semi-automatiquement), avec pour objectif une correction aussi parfaite que possible.

Les techniques utilisées seront : DeepLearning (Recurrent Neurone Network, LSTM, etc.), des bibliothèques telles que Spacy, NLTK ou d'autres bibliothèques nécessaires.

La langue étudiée sera prioritairement le français.