OCR
Optical Character Regnition

Introduction

De nos jours, de nombreux documents papier sont transformés en format électronique, ce qui facilite le traitement de l'information, comme la recherche, l'analyse et la conversion.

De nombreuses entreprises et autres institutions décident de numériser leurs documents. travailler avec des fichiers est moins cher que de traiter des documents traditionnels, car il n'y a pas d'espace requis pour le stockage de documents.

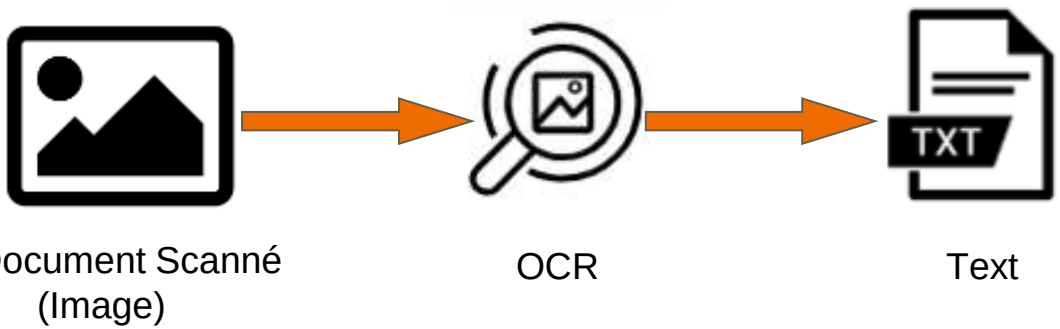
- Cependant le problème qui se pose est comment faire des recherches sur ces documents?
- Quels sont les outils utilisés pour y parvenir?

Le plan

- C’est quoi l’OCR?
- Les systèmes de l’OCR.
- Tesseract (le moteur de l’OCR)
- Comment améliorer la précision de l’OCR?
- Définition de l’analyse sémantique.
- Conclusion.

L’OCR qu'est-ce que c’est?

La reconnaissance de caractères est utilisée pour décrire des algorithmes et techniques (à la fois électroniques et mécaniques) pour convertir des images de text en text.



Comment fonctionne l'ocr

Les systèmes de l'OCR

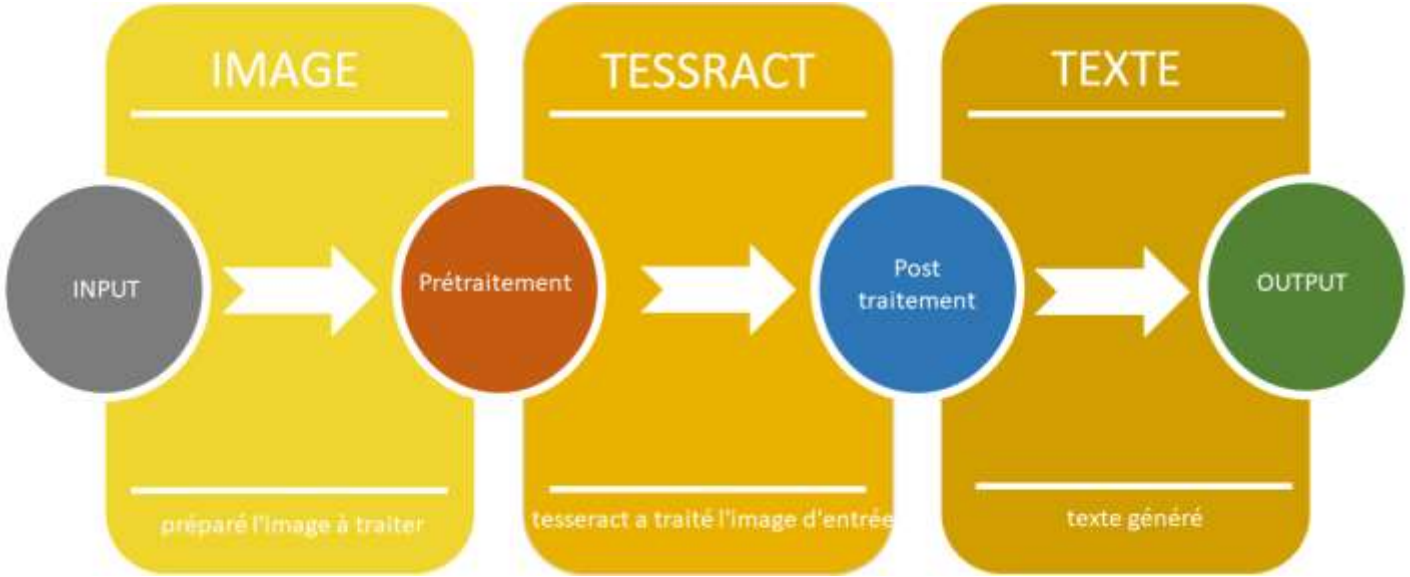
Logiciel
Comme TESSERACT

Dispositif d'OCR physique
(smartpen)

Tesseract (le moteur de l'OCR)

Tesseract est un moteur de reconnaissance optique de caractères pour divers systèmes d'exploitation. C'est un logiciel libre.

OCR se décompose généralement en plusieurs sous-processus exécutés de manière séquentielle:

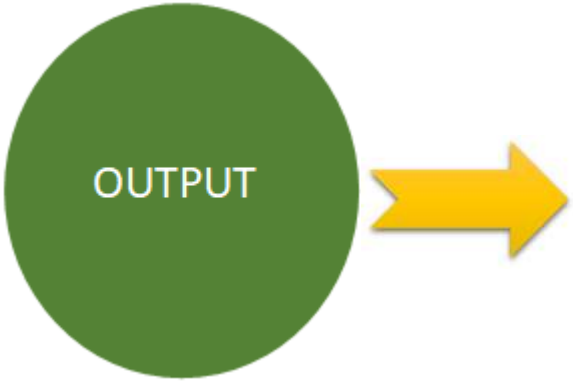


Prétraitement de l’image, comment augmenter la précision?

Il exist plusieurs étapes pour amélioré la précision de l’OCR:

- Redresser l’image.
- Supprimer la bordure autour de l’image produite par le redressement de l’image.
- Rendre l’image transparente: supprimer l’arrière-plan de l’image.

Le produit finale: créer un fichier PDF



La dernière étape est de créer un fichier PDF consultable avec calque de text invisible placé en haut de l’image.

Natural Language Processing Semantic Analysis

L’analyse sémantique consiste à **déterminer le sens d’une expression ou d’un texte en analysant finement les combinaisons de mots et le contexte**.

Conclusion

Ce travail nous a permis de:

- découvrire c’est quoi l’OCR.
- Le traitement de l’image nous a permis d’avoir plus de précision et par conséquent avoir des résultats plus adéquats.
- créer un fichier pdf consultable.