

Le Mans School of AI

PDF to Datas : Extraction et qualification de la donnée dans un pdf scanné ou numérique

Sommaire

Pourquoi extraire des données ?

Étape 1 : OCR

Étape 2 : Table extraction

Étape 3 : Géographie

Étape 4 : NLP & Classification

Aller plus loin & Sources

Pourquoi extraire des données ?

Début de chaîne informatique et logistique

Extraire et qualifier la donnée

Automatiser le processus de saisie manuel

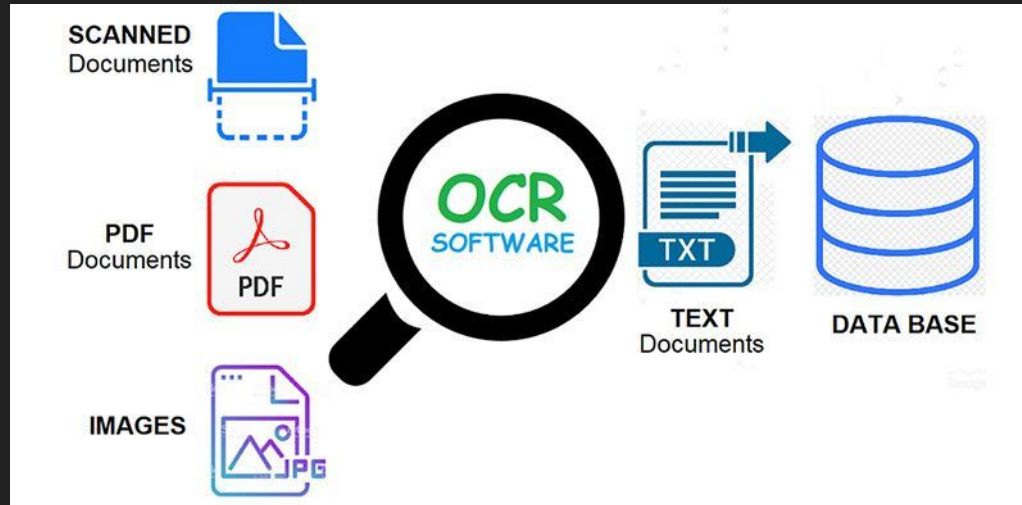


Pourquoi extraire des données ?



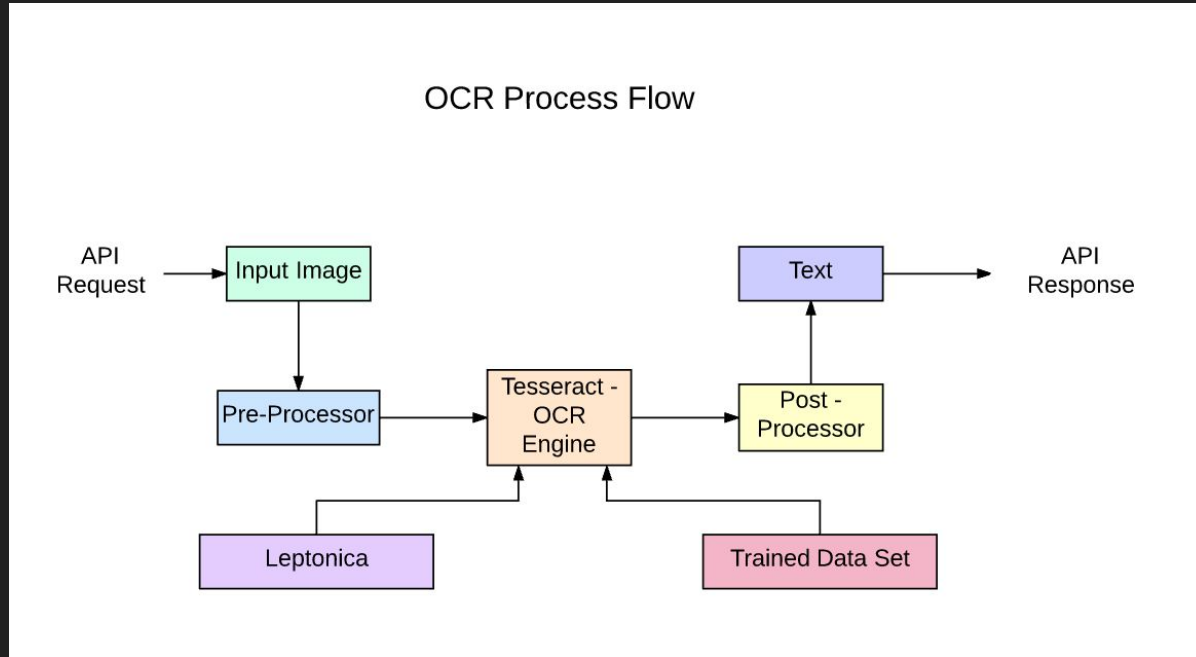
Étape 1 : OCR

L'OCR (Optical Character Recognition) est un processus de détection de lettres et de mots dans une image.



Étape 1 : OCR

Tesseract-OCR est le logiciel opensource le plus populaire.

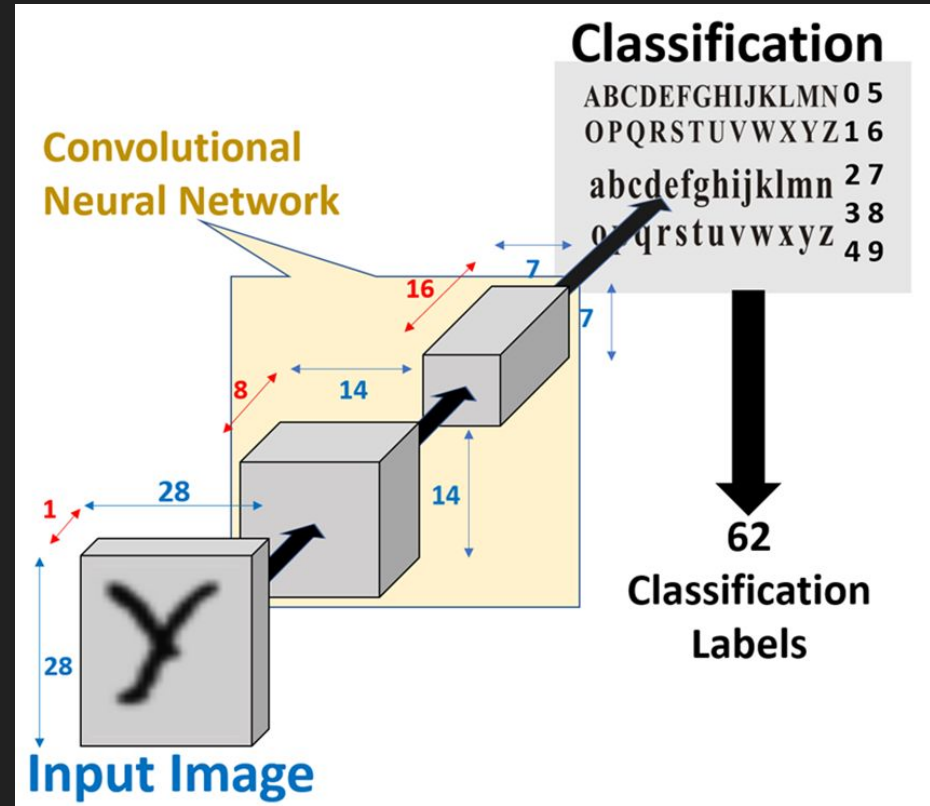


Étape 1 : OCR

Classification sur les caractères latins, mais aujourd'hui élargi à l'ensemble des caractères (même en Klingon ou en langue elfique de Tolkien), basée sur un CNN.

Base pratique en python :


<https://machinelearningmastery.com/handwritten-digit-recognition-using-convolutional-neural-networks-python-keras/>



Étape 1 : OCR

Chaque lettre est ensuite regroupée avec ses plus proches voisins pour former des mots.


On peut alors extraire les word bounded boxes. Selon le paramétrage et la qualité de l'image scannée (fonctionne très bien sur du caractère imprimé, mais beaucoup moins sur du caractère écrit à la main), on peut avoir des incohérences, des mots superposés, etc.



Invoice

your_company (11) Address (25) State (3) City (10) Country (3) 1112223333 1112223334

http://mrs.invoice.com



BILL TO:
1000 1000
Alpha Bravo Road 100
1112223333 1112223334
client@example.net

SHIPPING TO:
1000 1000 Office
Office Road 100
1112223333 1112223334
Office@example.net

| | |
|---------------|------------------|
| Invoice # | 00000 |
| Invoice Date | 12/12/2001 |
| Name of Rep | 1000 |
| Contact Phone | 101-102-103 |
| Payment Terms | Cash on Delivery |

Amount Due \$4,170

| NO | PRODUCT / SERVICE | QUANTITY | RATE / UNIT | AMOUNT |
|-------------|-------------------|----------|-------------|---------|
| 1 | Tyre | 0 | \$20 | \$40 |
| 2 | steering wheel | 0 | \$10 | \$50 |
| 3 | engine oil | 10 | \$15 | \$150 |
| 4 | brake pad | 20 | \$100 | \$2,400 |
| Subtotal | | | | \$275 |
| TAX 10% | | | | \$27.5 |
| Grand Total | | | | \$302.5 |

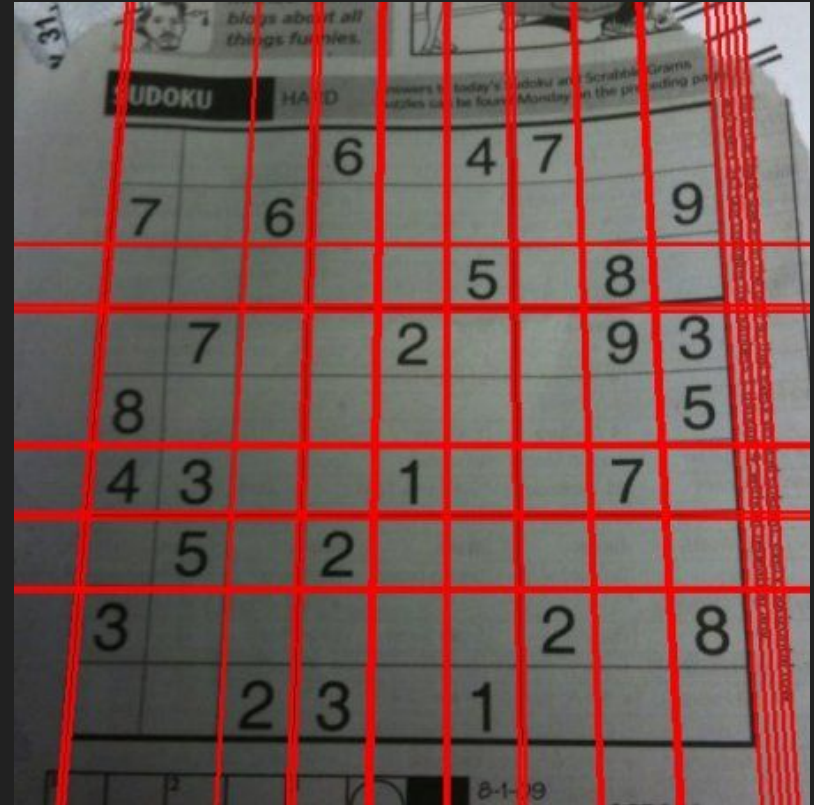
THANK YOU FOR YOUR BUSINESS

Étape 1 : OCR

On peut également procéder à un repérage de ligne car les tableaux ont une part importante dans la structure des informations.

Voici un petit tuto avec openCV :

https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_houghlines/py_houghlines.html



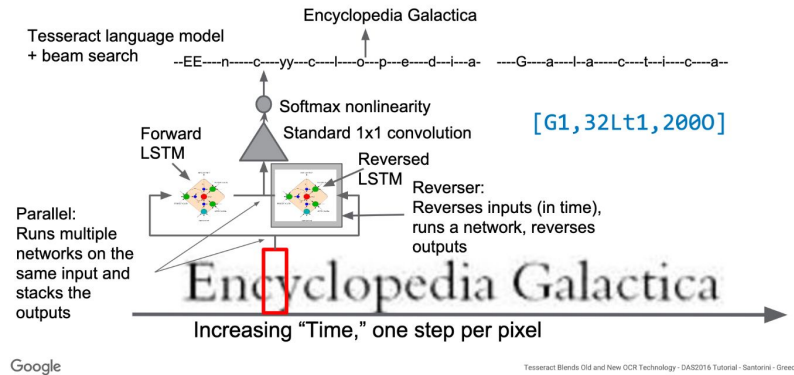
Étape 1 : OCR

Tesseract 4 embarque un LSTM en fin de processus pour améliorer la cohérence dans l'extraction des textes. Il croise, à la manière d'un cerveau humain, la vue (reconnaissance optique) et son savoir sur la prédiction du mot en fonction des mots précédents (connaissance du langage).

Base pratique en python :

<https://machinelearningmastery.com/text-generation-lstm-recurrent-neural-networks-python-keras/>


How Tesseract uses LSTMs...



Étape 2 : Table extraction

Nurminen paper : Algorithmic extraction of data in tables in PDF documents :
[see paper here](#)

Les données structurées en tableau sont difficiles à lire et à comprendre pour la machine, une première approche algorithmique a été recherchée.



| Etiology | Condition | Onset/ Duration | Symptoms |
|------------|----------------------|------------------------|---|
| Bacterial | Hyperacute bacterial | Acute | Purulent discharge, sometimes pain |
| | Acute bacterial | Acute | Tearing, lid crusting |
| | Chronic bacterial | Chronic | Lid crusting, foreign body sensation |
| Viral | Adenoviral | Acute | Tearing, lid crusting upon awakening |
| | Herpetic | Acute | Tearing |
| Allergic | Seasonal | Seasonal/ recurrent | Itching, tearing |
| | Vernal | Seasonal/ chronic | Itching, mucous discharge |
| | Giant papillary | Acute/ chronic | Itching, contact lens intolerance, mucous discharge |
| Chlamydial | Chlamydial | Acute/ Chronic | Tearing |

Étape 2 : Table extraction

La technique algorithmique utilise une projection de lignes imaginaires afin de constituer un tableau s'il n'est pas tracé.

Camelot emploie cette méthode, j'ai forké afin d'améliorer la détection.

<https://github.com/CartierPierre/camelot>

The diagram shows a table with the following structure and annotations:

- Table 1. Average grades** (Caption/legend)
- Term** (Row header)
- Assignments** (Superheader)
 - Ass1** (Subheader)
 - Ass2** (Subheader)
 - Ass3** (Subheader)
- Examinations** (Superheader)
 - Midterm** (Subheader)
 - Final** (Subheader)
- Final grade** (Subheader)

The table data is as follows:

| Term | Ass1 | Ass2 | Ass3 | Midterm | Final | Final grade |
|--------|------|------|------|---------|-------|-------------|
| 2012 | | | | | | |
| Winter | 85 | 80 | 75 | 60 | 75 | 75 |
| Spring | 80 | 65 | 75 | 60 | 70 | 70 |
| Fall | 80 | 85 | 75 | 55 | 80* | 75 |
| 2013 | | | | | | |
| Winter | 85 | 80 | 70 | 70 | 75* | 75 |
| Spring | 80 | 80 | 70 | 70 | 75 | 75 |
| Fall | 75 | 70 | 65 | 60 | 80 | 70 |

Annotations include: stub, title, subheaders, superheader, nested header, column, (table) header/column headers, boxhead, block, row header, caption/legend, elements, body/matrix, cell, and row.

Étape 2 : Table extraction

Nouvelles méthodes basées sur du DeepLearning pour la détection de tableau :

TableNet : Création de masques pour la détection des zones de tableau, lignes et colonnes.

<https://github.com/jainammm/TableNet>

DeepDeSTR : Détection de zones de tableau et détections des colonnes et lignes

<https://github.com/mawanda-jun/TableTrainNet>

Étape 3 : Géographie

Maintenant que nous avons un moyen d'extraire les données, donnons-leur un sens.

On peut regrouper les données les plus proches géographiquement en supposant qu'elles parlent de la même chose.

Ex : Destinataire d'une lettre, Mentions légales, En-tête d'entreprise, etc.

Étape 3 : Géographie

Joanna Binet

FACTURE

48 Coubertin
31400 Paris

Facturé à
Cendrillon Ayot
69 rue Nations
22000 Paris

Envoyé à
Cendrillon Ayot
46 Rue St Ferriol
92380 Île-de-France

Facture n° FR-001
Date 29/01/2019
Commande n° 1630/2019
Échéance 24/05/2019

| QTÉ | DÉSIGNATION | PRIX UNIT. HT | MONTANT HT |
|-----------|----------------------------------|---------------|------------|
| 1 | Grand brun escargot pour manger | 100.00 | 100.00 |
| 2 | Petit marinière uniforme en bleu | 15.00 | 30.00 |
| 3 | Facile à jouer accordéon | 5.00 | 15.00 |
| Total HT | | | 145.00 |
| TVA 20.0% | | | 29.00 |
| TOTAL | | | 174.00 € |

Joanna Binet

Conditions et modalités de paiement
Le paiement est dû dans 15 jours

Caisse d'Épargne
IBAN: FR12 1234 5678
SWIFT/BIC: ABCDFRP1XXX

Joanna Binet

FACTURE

48 Coubertin
31400 Paris

Facturé à
Cendrillon Ayot
69 rue Nations
22000 Paris

Envoyé à
Cendrillon Ayot
46 Rue St Ferriol
92380 Île-de-France

Facture n° FR-001
Date 29/01/2019
Commande n° 1630/2019
Échéance 24/05/2019

| QTÉ | DÉSIGNATION | PRIX UNIT. HT | MONTANT HT |
|-----------|----------------------------------|---------------|------------|
| 1 | Grand brun escargot pour manger | 100.00 | 100.00 |
| 2 | Petit marinière uniforme en bleu | 15.00 | 30.00 |
| 3 | Facile à jouer accordéon | 5.00 | 15.00 |
| Total HT | | | 145.00 |
| TVA 20.0% | | | 29.00 |
| TOTAL | | | 174.00 € |

Joanna Binet

Conditions et modalités de paiement
Le paiement est dû dans 15 jours

Caisse d'Épargne
IBAN: FR12 1234 5678
SWIFT/BIC: ABCDFRP1XXX

Étape 4 : NLP & Classification

Pour aller encore plus loin dans la compréhension, on peut utiliser du NLP pour comprendre le sens d'une phrase.

On spécialise un modèle pré-entraîné sur un modèle de langage, comme le français, avec Spacy.

On peut effectuer du NER pour récupérer les noms, lieux, adresses.

Étape 4 : NLP & Classification

Certaines informations fonctionnent à la manière d'une clé-valeur grâce au “:” comme :

Nom : Durand

Prénom : Michel

Ou par tableau :

| | |
|------------|---------|
| Prix total | 12345 € |
|------------|---------|

Pour aller plus loin & Sources

<https://nanonets.com/blog/ocr-with-tesseract/>

<https://github.com/tesseract-ocr/tesseract>

<https://nanonets.com/blog/named-entity-recognition-2020-guide/>

<https://nanonets.com/blog/key-value-pair-extraction-from-documents-using-ocr-and-deep-learning/>