

UNIVERSITÉ LUMIÈRE LYON 2

Thème :

*Expérimentation et évaluation
d'outils d'OCR et d'OLR.*



Présenté par Edina ADJARO PATOUSSI

30 Aout 2024

Tuteur laboratoire:



Julien VELCIN

Ludovic MONCLA

Tuteur enseignant:

Valentin Lachand-Pascal

sommaire

- 
- 01 CONTEXTE ET PROBLÉMATIQUE
 - 02 PRÉSENTATION DE LA STRUCTURE D'ACCUEIL
 - 03 OBJECTIF
 - 04 SEGMENTATION OU OLR
 - 05 RECONNAISSANCE DE CARACTÈRE OU OCR
 - 06 EXPÉRIMENTATION DES OUTILS.
 - 07 PROPOSITION DE MÉTHODOLOGIE
 - 08 RÉSULTATS OBTENUS
 - 09 ÉVALUATION DES RÉSULTATS
 - 10 CONCLUSION ET PERSPECTIVES
- 

Contexte et problématique.



Le **Trévoux** est un dictionnaire français datant du 19^{ème} siècle, qui existe en plusieurs versions.

Étant une œuvre importante, il a suscité l'intérêt de **linguistes** et a été l'objet de diverses études. Cependant, ne disposant pas d'un format numérique (.txt) il est difficile d'en faire une analyse approfondie.

C'est dans cette optique que ce stage a été initié, afin d'expérimenter des outils d'**OCR** et d'**OLR** sur le Trévoux

Comment surmonter les difficultés liées à l'**OCRisation** du Trévoux tout en respectant la structure du document ?



Pr sentation de la structure d'accueil



Deux  quipes :

-  quipe **DMD** :
 - Intelligence artificielle
 - Machine learning
 - Data mining
-  quipe **SID** :
 - Big data management
 - Big data analytics
 - Analyse en ligne (OLAP)
 - S curit  des donn es

Douze  quipes de recherche structur es en six p les de comp tences:

- Donn es, Syst me et S curit 
- Informatique Graphique et G om trie
- Images, Vision et Apprentissage (** quipe IMAGINE**)
- Interactions et cognition
- Algorithmique et Combinatoire
- Simulation et Sciences du Vivant

Objectif

- Tester et comparer les outils d'OLR/OCR sur les pages de Trevoux de 1704 et 1743
- Constituer un jeu de données annoté pour l'entraînement de modèles d'OLR/OCR avec Label Studio ou Prodigy si nécessaire.
- Développer une méthode permettant de comparer la sortie de ces outils avec la vérité terrain.

Trevoux 1704

ACC.
ACCUSATEUR, ACCUSATRIX. Subj. masc. & fem. Celui ou celle qui accuse, ou qui poursuit quelqu'un en Justice criminellement. *Accusator, Accusatrix.* Par le Droit civil il n'y avoit point d'accusateur public. Chaque particulier, soit qu'il eût intérêt au crime public, ou non, pouvoit accuser, & conclure au châtiment de l'accusé. En France il n'y a que le Procureur Général, ou les Substituts préposés dans chaque Siège, qui se puissent constituer accusateurs; c'est à eux seuls à qui appartient la vengeance publique. La partie civile ne peut conclure qu'à la réparation, & aux intérêts, & non pas à la punition du criminel. En quelque lieu que se trouve un Parricide, il rencontre un Accusateur, un Juge, & un Bourreau. L. A. M. A. Y. Cette femme étoit dangereuse accusatrix. Son accusatrix étoit fort animée contre lui. Au dernier jour nos pechez se présenteront comme autant de cruels accusateurs. N. I. C. O. L.

ACH.
ACHALANDER. V. ach. Attirer les marchandises, & vendre. Toute une boutique, ou une maison en réputation d'avoir de bonne marchandise, & à bon prix. *Empyros alliter.* Toute la fortune d'un marchand consiste à bien achalander sa boutique. C'est un terme du peuple; ou tout au plus de la conversation.

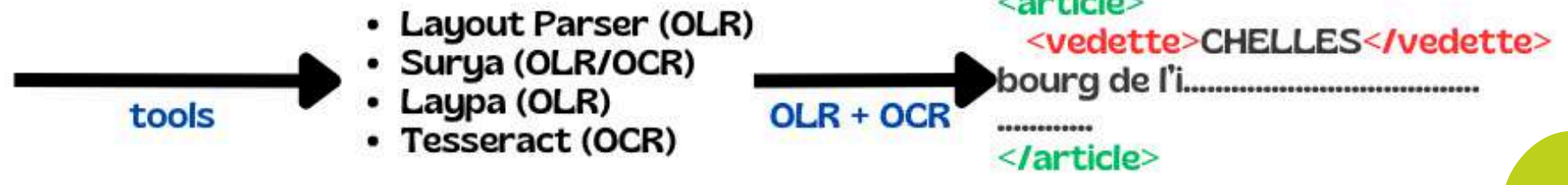
ACH.
ACHALANDER. V. ach. Attirer les marchandises, & vendre. Toute une boutique, ou une maison en réputation d'avoir de bonne marchandise, & à bon prix. *Empyros alliter.* Toute la fortune d'un marchand consiste à bien achalander sa boutique. C'est un terme du peuple; ou tout au plus de la conversation.

ACH.
ACHALANDER. V. ach. Attirer les marchandises, & vendre. Toute une boutique, ou une maison en réputation d'avoir de bonne marchandise, & à bon prix. *Empyros alliter.* Toute la fortune d'un marchand consiste à bien achalander sa boutique. C'est un terme du peuple; ou tout au plus de la conversation.

Trevoux 1743

CHE
CHELLES, Bourg de l'Île de France, à quatre lieues de Paris, sur la Seine. Sainte Baudour, femme de Clovis II. y fonda une Abbaye de Religieuses, dans laquelle elle se retira après la mort du Roi son mari; & où Clotaire II. son fils fut enterré. Le Roi Robert y avoit un Palais, qu'il appelle dans un Edit *Kala nostra Palatium*. Du CHESNE, *Arbor, des vill. de Fr. L. I. C. 30.*

CHE
CHELONITE. L. I. C. est une pierre qui se trouve au ventre des jeunes hirondelles, qu'on estime bonne pour le mal caduc. *Chelonia*. Il y a une autre chelonite qui se trouve aux tortues des Indes, qui a la vertu de résister au venin. Quel-



Segmentation ou OLR.

Definition:

La **segmentation d'image** est une **opération de traitement d'images** consistant à **détecter et rassembler les pixels** suivant **des critères**, notamment d'intensité.

Pourquoi :

- Bien **Extraire** la **Structure** des Dictionnaires
- Améliorer la **Précision** de l'**OCR**

Erreur courante pendant la segmentation

Vérité du terrain

Text line 1 Text line 2

Text line

Text line 1
Text line 2

Text1 Text3
Text2 Text4

Après segmentation

Text line1 || Text line 2

Text line

Text line1
Text line2

Text1 || Text 3
Text2 || Text 4

Reconnaissance de caractère ou OCR

Definition:

La Reconnaissance Optique de Caractères (OCR) est un processus qui permet de **convertir du texte imprimé** en **format numérique**

Erreur :

- **Délétion** : caractère manquant
word \rightarrow wrd
- **Substitution** : un caractère remplacé par un autre (voire plusieurs)
word \rightarrow w0rd word \rightarrow ivord
- **Insertion** : caractère inséré dans le texte
word \rightarrow wordsd

Expérimentation des outils

Layout Parser

Bibliothèque Python qui permet de faire de l'analyse d'image en détectant les layouts des différents éléments d'une image en utilisant des modèles entraînés avec Detectron.

HJDataset

PubLayNet

PrimaLayout

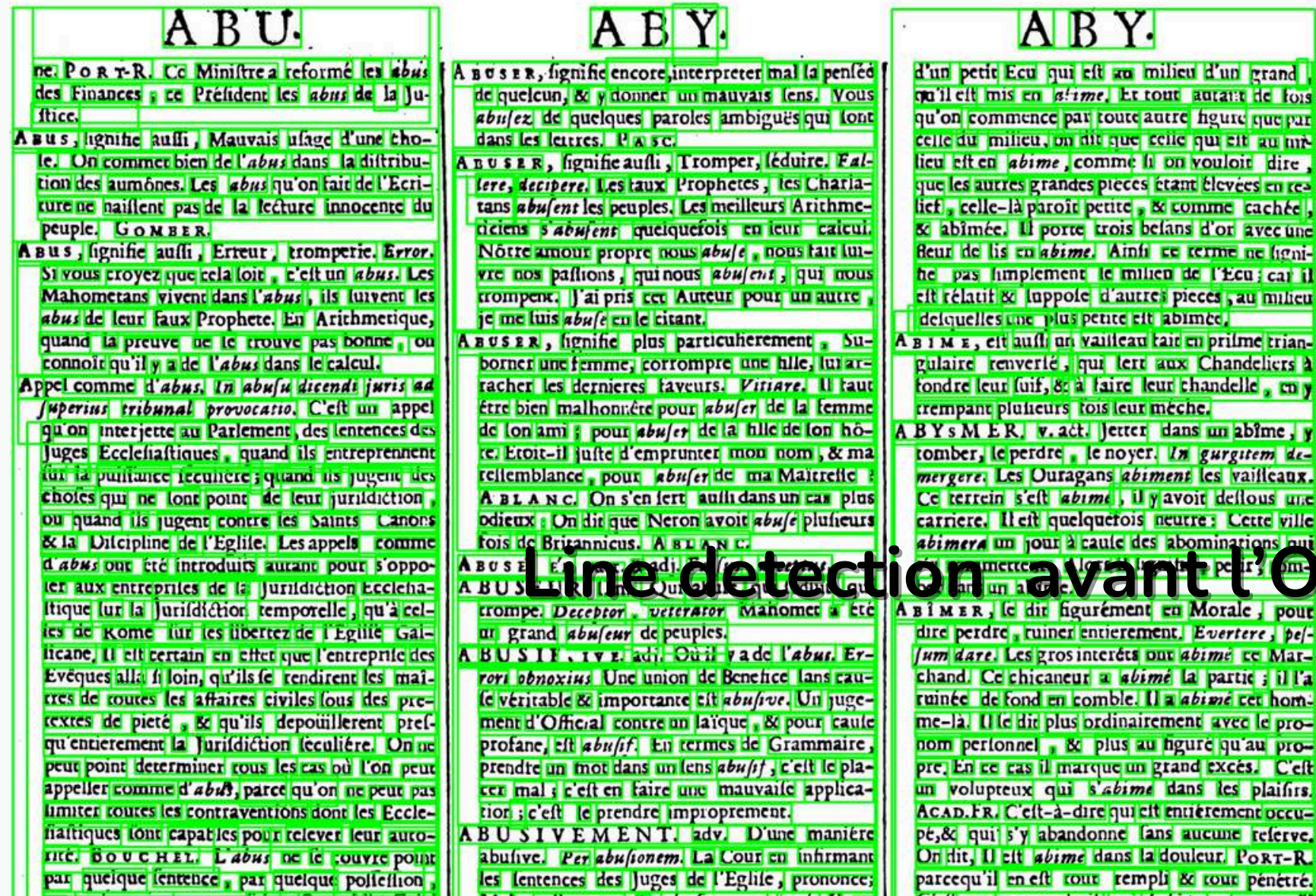


Expérimentation des outils

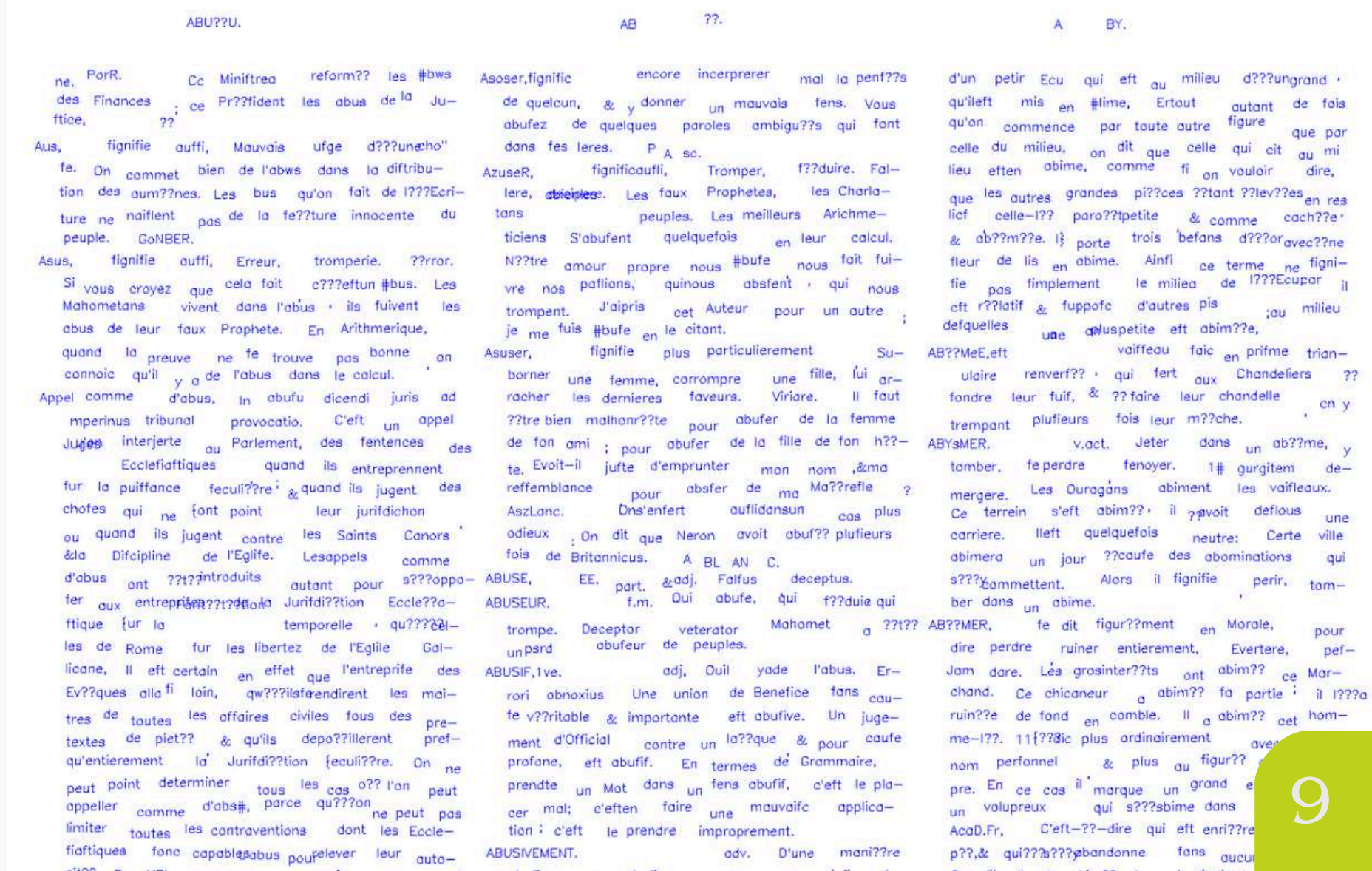
Tesseract

Librairie qui permet de faire de l'extraction de texte et de l'OCR. Pour l'utiliser, nous sommes passés par la bibliothèque pytesseract.

OCR de Tesseract



Line detection avant l'OCR



Expérimentation des outils

Résultat Laypa



Résultat surya

A B. A B A.

A B A.

che: être bien à cheval. 1. La posture & le geste: A genoux, à bras ouverts. 3. La distance: A vingt lieues de là. 4. La qualité: De l'or à tant de carats. 5. Le prix: A dix écus. 6. La quantité: L'eau est à la hauteur d'une toise. 7. La manière: Il est habillé à l'Espagnole. Il faut dire à coups de trait, à coups de canon: & non pas à coups de traits, & à coups de canons. MENAG. 8. La fin: Les fraudes à bonne intention ne manquent point d'approbateurs parmi les devoirs. PORT-R.

A signifie successivement: Pas à pas. Il se sent mourir peu à peu. Il signifie avec: Je l'abandonne à regret. Les douleurs à grand bruit sont d'ordinaire suspectes d'affectation. M. SCUD. Ce poste a été emporté à la pointe de l'épée. Peindre à l'huile.

A est plus élégant que par dans certaines phrases. Il ne faut point se laisser prendre à l'apparence, ni à l'éclat trompeur des grandeurs humaines. FLICH. Ne vous laissez pas conduire à vos passions. A signifie selon: A mon avis.

A Cette lettre s'emploie aussi fort souvent pour marquer ce que l'on possède. C'est un homme à carolle, à équipage.

A se met quelquefois absolument devant l'infinitif de quelques verbes, sans être précédé d'aucun nom qui soit ou exprimé, ou sous-entendu, & alors il se peut résoudre par le gerondif. A voir les airs dedaigneux. A tout prendre, l'aillemblage de ses traits, qui sont beaux en détail, ne fait point une belle personne. FONTEN. C'est comme si l'on disoit, en prenant tous ses traits ensemble. Passer tranquillement la nuit à bien dormir, & le jour à rien faire. BOIL. Il y a aussi des occasions où il se peut résoudre par quand, ou lorsque. A ne prévoir rien on est surpris, & à prévoir trop on est misérable. S. EVR. A raconter les maux souvent on les soulage. CORN. Il se met aussi devant l'infinitif de quelques ver-

Il seroit difficile de déterminer tous les differens usages de la préposition ou de la particule à. On les remarquera dans la suite, il s'en présentera des exemples presque à toutes les pages.

A. A. A. Les Chymistes se servent de ce signe pour signifier, Amalgamer, Amalgamation, & Amalgame. Voyez Amalgamer.

A B A.

A B. cinquième mois des Hebreux, qui répond à notre mois de Juillet.

A B. en langue Syriaque, le dernier mois de l'Eté.

A B A, ou ANBA, Pere, titre que les Eglises Syriaques, Cophtes & Ethiopiennes donnent à leurs Evêques.

A B A C O. subst. masc. Abacus. Ce mot se trouve dans Rouillard pour signifier l'Arithmetique. Les Italiens disent aussi abaco pour exprimer la même chose. C'étoit une petite table polie, sur laquelle les Anciens traçoient des figures, ou des nombres. Elle seroit à apprendre les principes de l'Arithmetique. Ils l'appelloient Table de Pythagore.

A B A D A. f. m. Animal farouche du païs de Benguela, dans la ballé Ethiopie. Il ressemble à un cheval par la tête, & par le crin. Il est un peu moins grand. Sa queue est pareille à celle d'un bœuf, excepté qu'elle est moins longue. Ses pieds sont fendus comme ceux du cerf, & plus gros. Il a deux cornes, l'une sur le front, & l'autre sur la nuque. Les Negres tuent ces animaux à coups de flèche, pour en prendre la corne, dont ils font un remede.

A B A D I R. Terme de Mythologie. C'est le nom d'une pierre que Saturne dévora. Car soit parce que son frere Titanus ne lui avoit cédé l'empire du monde, qu'à condition qu'il n'éleveroit point d'enfant mâle; soit parce que les destinées portoient qu'il seroit un jour détrôné par un de ses enfans: il les faisoit tous perir. Euhu Cybele

ABAISSE, signifie aussi, Diminuer le prix. Minuer. Le bon ordre de la police a fait abaisser le prix du blé; c'est-à-dire, qu'il est diminué. Ce mot en ce sens n'est pas du bel usage; il faut dire rabaisser. Voyez RABAISSE.

On s'en sert figurément dans le même sens. L'envie abaisse par ses discours les vertus qu'elle peut imiter. S. EVR. Abaisser la Majesté du Prince. L'usage, comme la fortune, chacun dans leur juridiction élève ou abaisse qui bon lui semble. VAUG. Les grands noms abaissent, au lieu d'élever, ceux qui ne savent pas les soutenir, ROCHER.

ABAISSE, signifie aussi en Morale, Ravaler l'orgueil de quelqu'un, le mortifier. Abjicere, Reprimere, Contundere. Les Romains se van-toient d'abaisser les superbes, & de pardonner aux humbles. S. EVR. Abaisser l'orgueil de Carthage. VAUG. Il faut abaisser les esprits hautains. S. EVR. La crainte trouble & abaisse l'esprit M. SCUD. C'est-à-dire qu'elle le relâche & l'avilit.

En termes de Fauconnerie on dit, Abaisser l'oiseau, lors qu'ayant trop d'embonpoint, on lui ôte quelque chose de son pâr ordinaire, pour le mettre en état de bien voler.

ABAISSE, en termes de Jardinage, signifie, Couper une branche près du tronc.

ABAISSE, se dit aussi avec le pronom personnel, & signifie alors, s'humilier, se soumettre, se ravaler. Abjicere se. Il faut s'abaisser devant la Majesté divine. S'abaisser à des choses indignes. S'abaisser jusqu'aux plus lâches complaisances. L'humilité n'est souvent que le rifice de l'orgueil, qu'une s'abaisse que lever. ROCHER. On le dit encore de quelqu'une personne éminente en dignité, qu'elle semble rabatre de sa grandeur pendant jusqu'à des personnes fort basses. Le Prince s'est abaisse jusqu'à moi.

Expérimentation des outils

Tableau Récapitulatif :

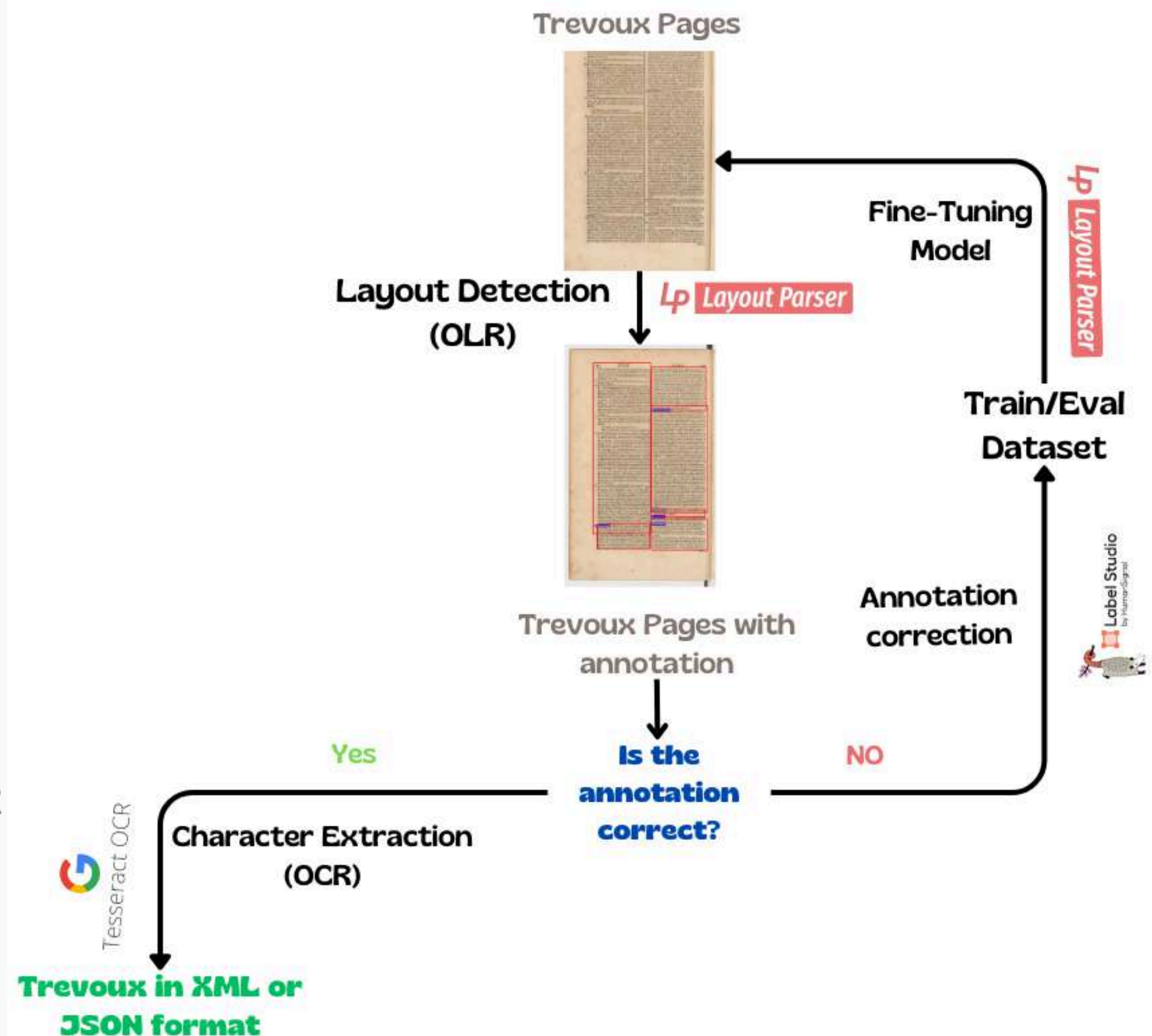
	Documentation	Type-Input	Type-Output	Open source	Fine -Tunning
Layout Parser	Exhaustive	Image, PDF	JSON , CSV	Oui	Oui
Laypa	Peux exhaustive	Image	XML	Oui	Possible mais difficile de constituer le dataset d'entraînement
Grobid	Exhaustive	PDF	JSON , XML	Oui	Possible mais difficile de constituer le dataset d'entraînement
Tesseract	Exhaustive	Image, PDF	Txt, HOOCR	Oui	Non
Surya	Presque inexistant	Image, PDF	JSON , HOOCR	Oui	Non

Proposition de méthodologie.

Fine-Tuner le modèle **primaLayout** basé sur **Detectron 2**, qui fonctionne assez bien sur les **pages** du Trévoux.

Pour ce faire, nous avons défini certaines étapes, qui sont les suivantes :

- Etape 1 : Préparation des données
- Etape 2 : Génération des annotations
- Etape 3 : Correction des annotations
- Etape 4 : Fine Tuning du model
- Etape 5 : Évaluation du nouveau model et OCR



Résultats obtenus

Avant

AUX-AUZ-AXE-AXI-AXU-AYA-AYE-AYN-AZE-AZI-AZO-AZU.

AUX **AUXILIAIRE**. adj. m. & f. Qui vient à secours. *Auxiliaris*. Un prince doit plus se fier à ses soldats, qu'à ses troupes auxiliaires. Outre les raisons principales, on se sert quelquefois heureusement de moyens auxiliaires, ou subsidiaires. Chez vous les injures sont les troupes auxiliaires de la raison. S. E. V. R.

AXU **AXONGE**. f. f. C'est une espèce de graisse la plus molle & la plus humide du corps des animaux, qui s'appelle autrement de l'oing. *Axungia*. Elle est différente du lard, qui est une graisse ferme; & du suif, qui est une graisse sèche. Les Latins font la même distinction de la graisse en *pinguedo*, qui est l'*axunge*, *lardum*, & *sebum*. On l'appelle aussi en Latin *axungia*, qu'on dit avoir été fait *ab axe rotarum qua unguuntur*. On se sert en Médecine de l'axonge d'oye, de canard, de vipère, & de plusieurs autres, même de celle de l'homme, qu'on estime beaucoup pour resoudre, & pour appaiser les douleurs.

AXONGE DE VERRE, qu'on appelle aussi *fiel*, ou *sel de verre*. C'est une écume séparée de dessus le *matras de verre*, comme on l'a vu ci-dessus.

AYA **AYANT**. Participe du verbe Avoir, qu'on rencontre presque par tout dans les Auteurs, qui s'exprime en latin par les adverb. *Cum*, *Postquam*, *Posteaquam*. *Ayant* dit cela, je m'en allai. *Ayant* fait beaucoup de plaintes, il se retirera. *Ayant* été dangereusement blessé, il fut emporté par des soldats. Pour dire, Après avoir dit, Après avoir fait, Après avoir été blessé.

AYE **AYEUL**. f. m. & f. Père, ou mère de ceux qui ont des enfans, à l'égard desquels on le nomme aussi *Grand-père*, ou *Grand-mère*. *Ayus*. Chaque enfant a un *ayeul* paternel, & un *ayeul* maternel. Il fait *ayeux* au pluriel, & non pas *ayeuls*. CORN.

Ce long amas d'ayeux que vous dissamiez, sont autans de temoins qui parlent contre vous. BOIL.

Quelque rang où jadis soient montez vos ayeux, RACIN.

Leur gloire de si loin n'ébloit point mes yeux. RACIN.

L'amour de vos ayeux passe en vous pour manie. BOIL.

Se base qui vaudra du nom de ses ayeux?

AZE **AZEROLIER**. f. m. Arbre sauvage, épineux, & de moyenne hauteur. Ses feuilles sont découpées comme celles du persil. Ses fleurs sont blanches & entassées en grappes. Il porte des fruits aigres & secs qu'on nomme *azerolles*, & qui sont rouges & gros comme des cerises. Ils sont assez agréables au goût étant meurs. C'est une espèce de néflier. En Latin *mespilus apii folio laciniato*, ou *mespilus Aronia*. On le greffe sur l'épine blanche, ou sur le sauvageon de poirier, & sur le cognacier. Il y en a un qui vient du Canada, dont les épines sont très-longues, & les feuilles très-grandes. Il y en a aussi un blanc qui vient de Florence, qu'on ne trouve qu'à Versailles, & qui ne diffère de l'autre que par la couleur de son fruit.

AZI **AZIMUTAL**. f. m. Terme d'Astronomie. C'est un grand cercle vertical qui passe par le zénith & le nadir, & qui coupe l'horizon à angles droits. *Verticalis circulus horizontem ad angulos rectos interfecans*. Or comme l'horizon est divisé par 360 degrés, il donne lieu à des *azimuts*. Ce mot est purement Arabe. Ces cercles sont les mêmes que les *rumbes* des Mariniers marquez sur la Carte. On commence à les compter depuis le point du vrai Orient ou de l'Orient Equinoctial, & on continue en allant vers le Midy jusqu'à 360. C'est dans ces cercles qu'on prend la hauteur des astres à toutes les heures.

AZO **ZOT**. f. m. Terme de Chymie. C'est ainsi que les Chymistes appellent la matière première des métaux.

AZI **ZOVALALA**. f. m. Petit fruit rouge de l'Isle de Madagascar. Il croît sur un petit arbrisseau comme les groseilles.

AZI **ZOUFA**. f. f. Bête du Royaume de Casubi. On en trouve aussi à Fez, & à Maroc. Ces animaux deservent les mors. & les dé-

Après

AUX-AUZ-AXE-AXI-AXU-AYA-AYE-AYN-AZE-AZI-AZO-AZU.

AUX **AUXILIAIRE**. adj. m. & f. Qui vient à secours. *Auxiliaris*. Un prince doit plus se fier à ses soldats, qu'à ses troupes auxiliaires. Outre les raisons principales, on se sert quelquefois heureusement de moyens auxiliaires, ou subsidiaires. Chez vous les injures sont les troupes auxiliaires de la raison. S. E. V. R.

AXU **AXONGE**. f. f. C'est une espèce de graisse la plus molle & la plus humide du corps des animaux, qui s'appelle autrement de l'oing. *Axungia*. Elle est différente du lard, qui est une graisse ferme; & du suif, qui est une graisse sèche. Les Latins font la même distinction de la graisse en *pinguedo*, qui est l'*axunge*, *lardum*, & *sebum*. On l'appelle aussi en Latin *axungia*, qu'on dit avoir été fait *ab axe rotarum qua unguuntur*. On se sert en Médecine de l'axonge d'oye, de canard, de vipère, & de plusieurs autres, même de celle de l'homme, qu'on estime beaucoup pour resoudre, & pour appaiser les douleurs.

AXONGE DE VERRE, qu'on appelle aussi *fiel*, ou *sel de verre*. C'est une écume séparée de dessus la matière du verre, avant qu'elle se victrifie.

AYA **AYANT**. Participe du verbe Avoir, qu'on rencontre presque par tout dans les Auteurs, qui s'exprime en latin par les adverb. *Cum*, *Postquam*, *Posteaquam*. *Ayant* dit cela, je m'en allai. *Ayant* fait beaucoup de plaintes, il se retirera. *Ayant* été dangereusement blessé, il fut emporté par des soldats. Pour dire, Après avoir dit, Après avoir fait, Après avoir été blessé.

AYE **AYEUL**. f. m. & f. Père, ou mère de ceux qui ont des enfans, à l'égard desquels on le nomme aussi *Grand-père*, ou *Grand-mère*. *Ayus*. Chaque enfant a un *ayeul* paternel, & un *ayeul* maternel. Il fait *ayeux* au pluriel, & non pas *ayeuls*. CORN.

Ce long amas d'ayeux que vous dissamiez, sont autans de temoins qui parlent contre vous. BOIL.

Quelque rang où jadis soient montez vos ayeux, RACIN.

Leur gloire de si loin n'ébloit point mes yeux. RACIN.

L'amour de vos ayeux passe en vous pour manie. BOIL.

Se base qui vaudra du nom de ses ayeux?

AZE **AZEROLIER**. f. m. Arbre sauvage, épineux, & de moyenne hauteur. Ses feuilles sont découpées comme celles du persil. Ses fleurs sont blanches & entassées en grappes. Il porte des fruits aigres & secs qu'on nomme *azerolles*, & qui sont rouges & gros comme des cerises. Ils sont assez agréables au goût étant meurs. C'est une espèce de néflier. En Latin *mespilus apii folio laciniato*, ou *mespilus Aronia*. On le greffe sur l'épine blanche, ou sur le sauvageon de poirier, & sur le cognacier. Il y en a un qui vient du Canada, dont les épines sont très-longues, & les feuilles très-grandes. Il y en a aussi un blanc qui vient de Florence, qu'on ne trouve qu'à Versailles, & qui ne diffère de l'autre que par la couleur de son fruit.

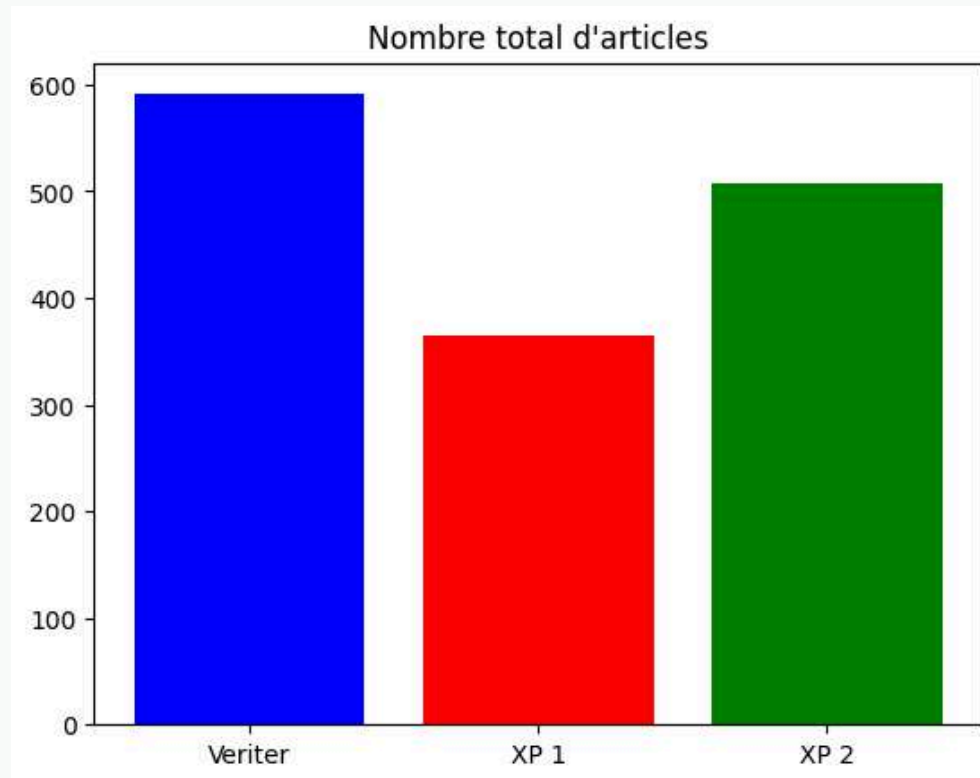
AZI **AZIMUTAL**. f. m. Terme d'Astronomie. C'est un grand cercle vertical qui passe par le zénith & le nadir, & qui coupe l'horizon à angles droits. *Verticalis circulus horizontem ad angulos rectos interfecans*. Or comme l'horizon est divisé par 360 degrés, il donne lieu à des *azimuts*. Ce mot est purement Arabe. Ces cercles sont les mêmes que les *rumbes* des Mariniers marquez sur la Carte. On commence à les compter depuis le point du vrai Orient ou de l'Orient Equinoctial, & on continue en allant vers le Midy jusqu'à 360. C'est dans ces cercles qu'on prend la hauteur des astres à toutes les heures.

AZO **ZOT**. f. m. Terme de Chymie. C'est ainsi que les Chymistes appellent la matière première des métaux.

AZI **ZOVALALA**. f. m. Petit fruit rouge de l'Isle de Madagascar. Il croît sur un petit arbrisseau comme les groseilles.

AZI **ZOUFA**. f. f. Bête du Royaume de Casubi. On en trouve aussi à Fez, & à Maroc. Ces animaux deservent les mors, & les dé-

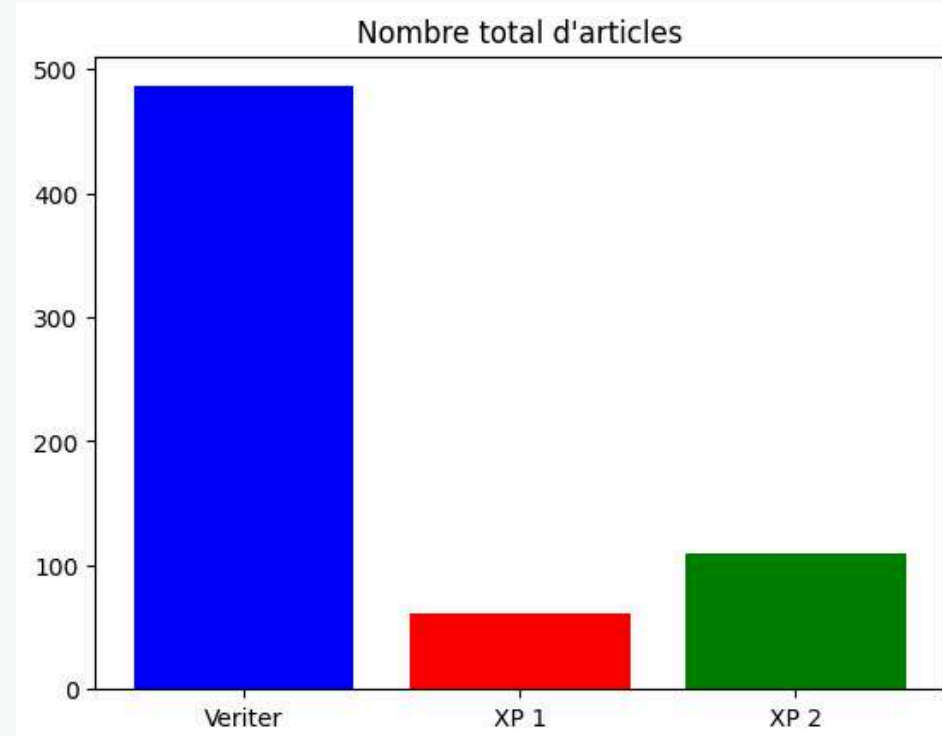
 valuation des r sultats



Interpr tation :

le mod le XP2 est plus performant que le mod le XP1 en termes de d tection d'articles.

En effet, le mod le XP1 a d tect  364 articles sur 591 possibles, soit un ratio de 0,615, alors que XP2 a d tect  508 articles sur 591, soit un ratio de 0,859



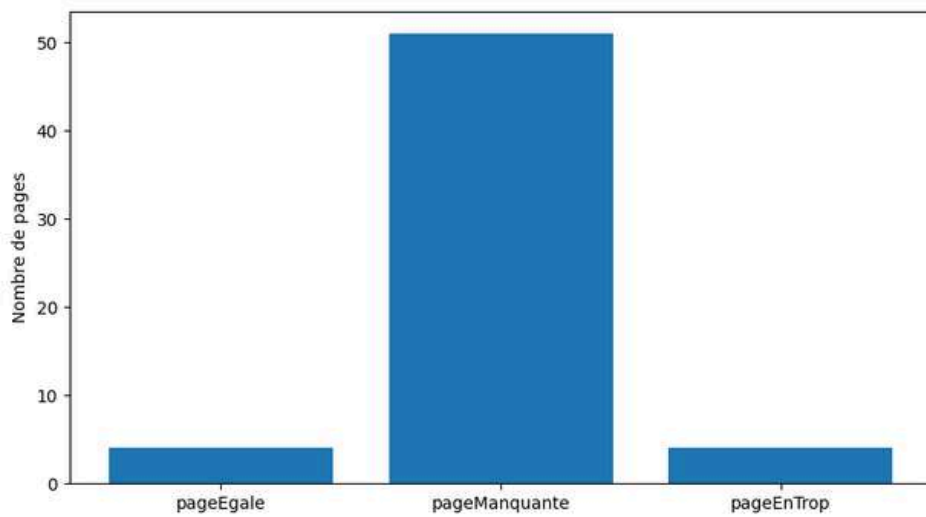
Interpr tation :

On remarque que les deux mod les sont relativement mauvais dans la d tection de vedettes.

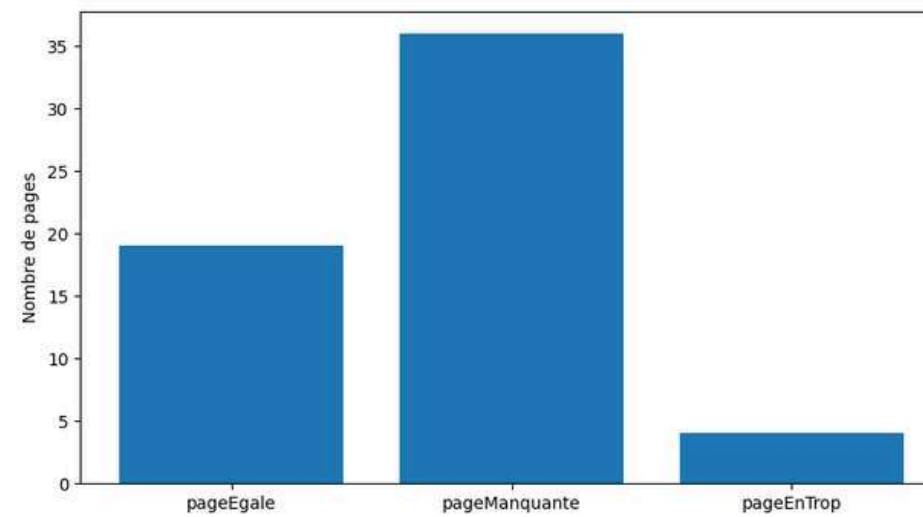
En effet, le mod le XP1 a d tect  61 vedettes sur 486 possibles, soit un ratio de 0,125, alors que XP2 a d tect  110 vedettes sur 486, soit un ratio de 0,22.

Évaluation des résultats

Evaluation des résultats de détection des articles

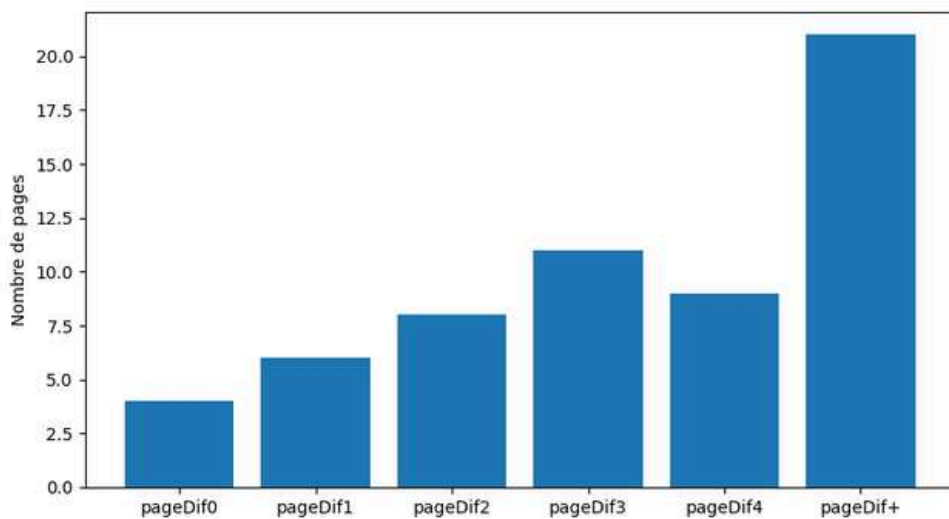


pageEgale:4
pageManquante:51

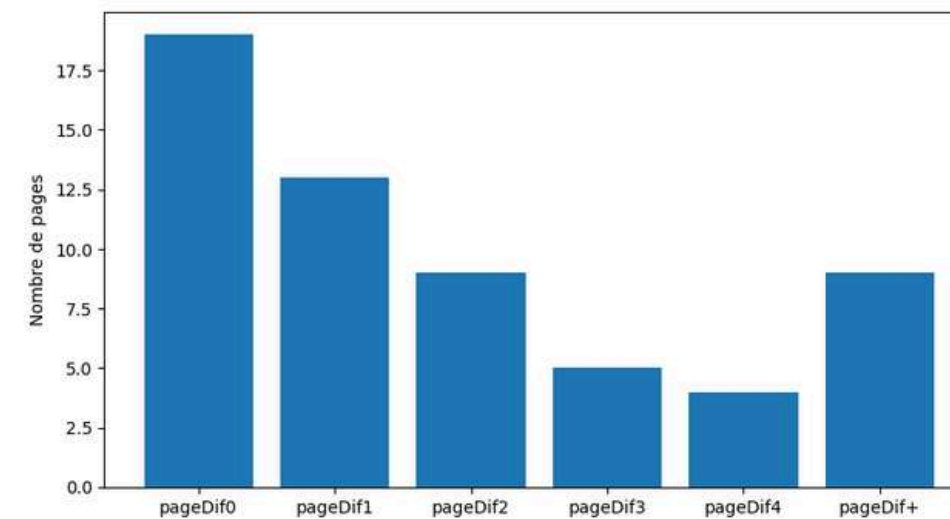


pageEgale:19
pageManquante:36

Evaluation des résultats de détection des articles



pageDif0:4
pageDif1:6
pageDif2:8
pageDif3:11
pageDif4:9
pageDif+:21



pageDif0:19
pageDif1:13
pageDif2:9
pageDif3:5
pageDif4:4
pageDif+:9

Interprétation :

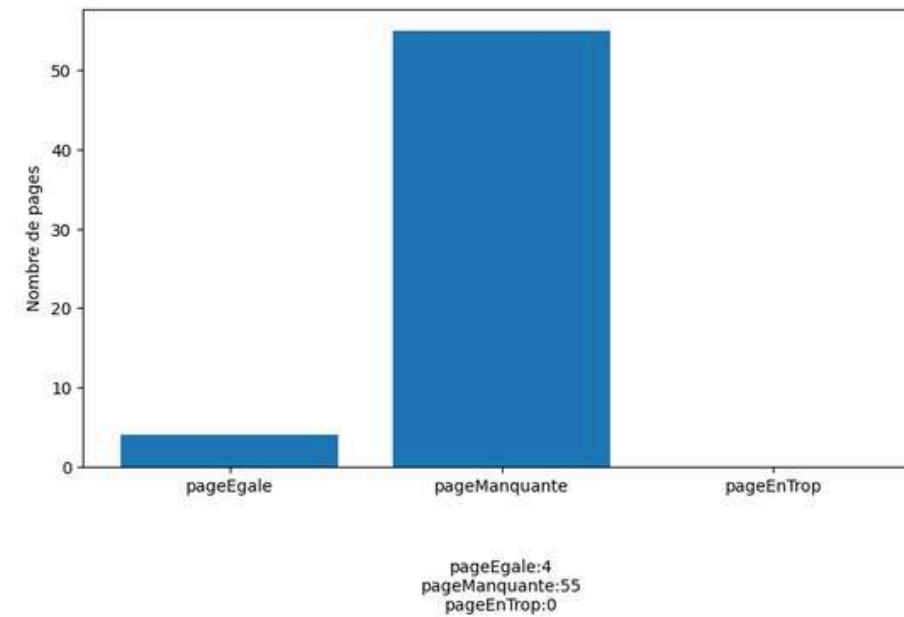
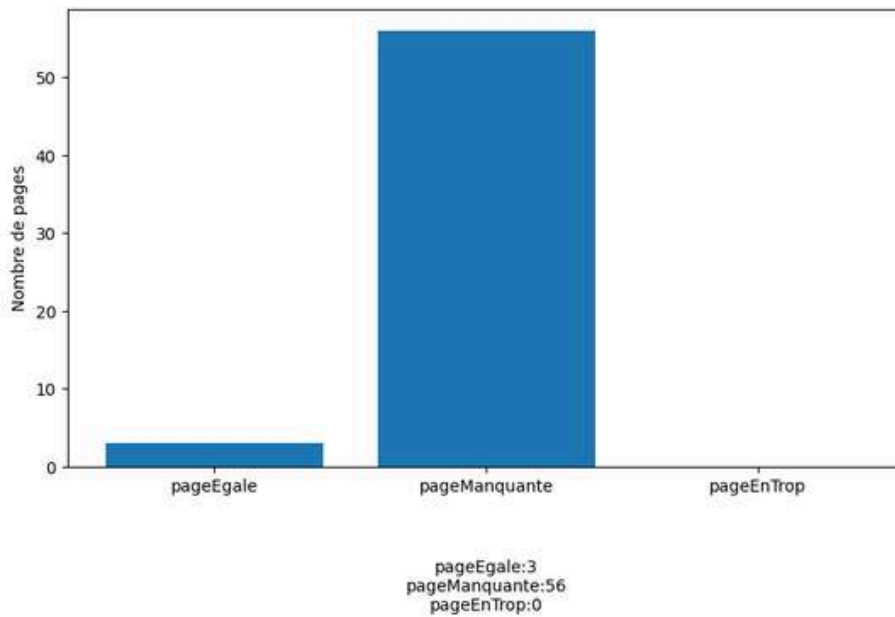
En zoomant sur le nombre de détections manquantes et en trop par page d'article, on remarque qu'il y a dans la plupart des cas une sous-détection, sauf dans le cas de 4 pages où les 2 modèles font de la surdétection

Interprétation :

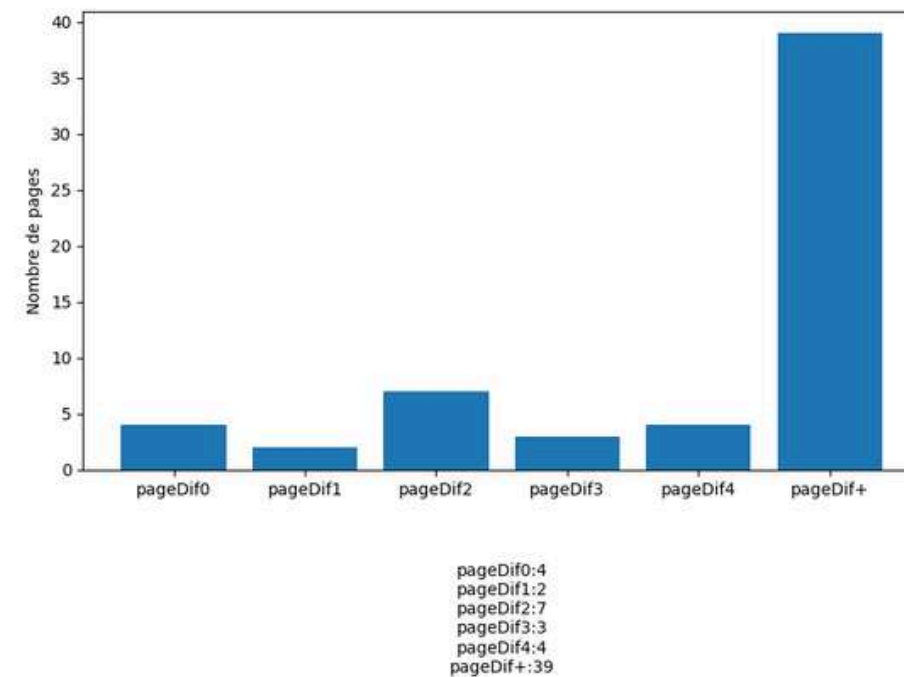
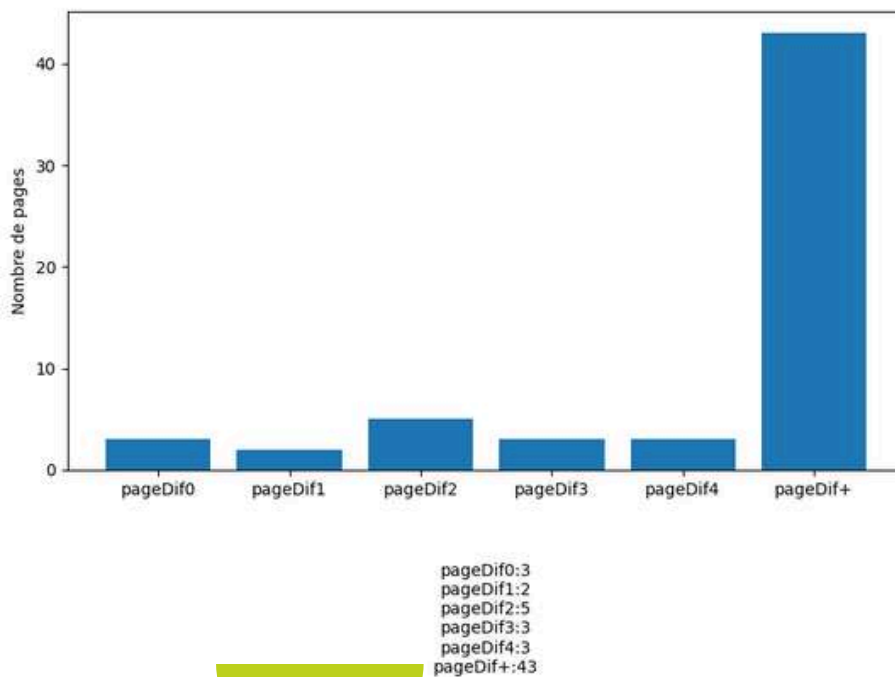
Ces graphiques montrent que le modèle XP2 (à droite) est beaucoup plus efficace dans la détection d'articles, car il fait le moins d'erreurs contrairement à XP1.

Évaluation des résultats

Evaluation des résultats de détection des articles



Evaluation des résultats de détection des articles



Interprétation :

En zoomant sur le nombre de détections manquantes et en trop par page de vedette, on remarque qu'il y a dans la plupart des cas une sous-détection donc nos 2 modèles ne sont pas assez bonne pour détecter les vedettes

Interprétation :

Dans la détection de vedettes, il n'y a pas de véritable tendance qui se dégage, car les deux modèles semblent être aussi mauvais l'un que l'autre.

 valuation des r sultats

Pour effectuer une analyse de l'OCR, nous avons suivi les  tapes pr liminaires suivantes :

- **Pr diction avec notre mod le** : R aliser une pr diction sur un  chantillon du Tr voux.
- **Extraction du texte** : Utiliser l'OCR Tesseract pour extraire les textes.
- **Reconstitution des pr dictions** : Reconstituer les pr dictions pour chaque page.
- **Pr paration de la v rit  terrain** : Reconstituer l' chantillon de v rit  terrain par page   l'aide des marqueurs disponibles dans nos fichiers.
- **Jointure des donn es** : Effectuer une jointure entre les pr dictions et la v rit  terrain en utilisant la cl  des pages.
- **Alignement des donn es** : R aliser un alignement gr ce   l'algorithme RETAS.
- **Affichage et comparaison des r sultats** : Afficher et comparer les r sultats obtenus.

Évaluation des résultats

25 donnera sera clair, ou obscur, ou obtus, bref ou long, selon les différentes	25 d'un mot fa prononciation chan- gera, & le fon qu'on lut donnera fera clair
26 consonnes qui le suivront, comme on vient de l'expliquer. A s. m. C'est le nom	26, ou obfcur , ou obtus, bref ou long , felon les ditférentes confonnes qui le
27 de cette lettre, ou du caractère que nous appellons a. Un grand a, un petit a,	27 {uivront, comme on vient de @expliquer. . A £, m. C@eft@le nom de cette letere
28 un a bien formé. Ce nom est du genre masculin, comme celui de toutes les	28, ou du caract@ère que nous appelions 4, Un@grand 4, un pe@nt@a, un # bien
29 voyelles Françaises. Cette lettre sert de corps à un Rebus en cette manière :	29 form@é. Ce nom eft du genre@mafculin, comme celui de toutes les voyel- ies
30 On range plusieurs A de suite jusqu'à un tombeau, & ces paroles font l'ame	30 Fran@goites. Ceste lettre fert de corps@d@un Rebus en cette mani@ère : On range
31 du Rebus, Amis jusqu'au tombeau. Cette lettre A étoit aussi chez les Anciens	31 pluficurs 4 de funite jufqu'd un tombeau, & ces paroles font @'ame du Rebus,
32 une lettre numérale qui signifioit 500. comme on le voit dans Valerius Probus.	32 Amis jufqu'an tonbean. Cetze lettre # @éroir aufli chez les Anciens une le@wre
33 Voyez sur ces prétendues lettres numériques ce qu'on en a remarqué sur la	33 num@érale qui {ignifioit \$o0, comme on le voit dans Valerius Probus. Voyez fur
34 lettre e. Il y a des vers anciens rapportés par Baronius, qui marquent les	34 ces@pr@étendues le@twes num@érales ce qu@on@en@a@re- marqu@é @(i la le@rre@e@ @e
35 lettres significatives des nombres, dont le premier est : Possidet A numeros	35 y a des vers anciens rapport@és par Baronius, qui marquent lgs lerrres
36 quingentos ordine recto Quand on mettoit un titre ou une ligne droite au-dessus	36 fignificatives des nom- bres, dont le premier eft : Poffider A numeros
37 de l'A, il signifioit cinq mille. Les Romains l'appelloient lettre salutare,	37 quingentos ordine reclo Quand on mettoit un titre ou une ligne droite au-deffus
38 parce qu'on s'en servoit pour déclarer innocent celui qui étoit accusé. A	38 de @T4, il fig@mf@oir cinq mille. Les Romains l'a@pfinlbi@cm lettre Salutare ,
39 vouloit dire absolvo, je l'absous. Cette lettre a diverses significations.	39 parce qu@on s'en fervoit pour d@éc@innocent * "celui qui @éroit accu@lé. A
40 Cependant il en faut éviter la rencontre trop fréquente dans une même	40 vouloit dire abfolve, je @Iablous. Cetre le@tre 2 diverfes fignifications.
41 période. Quelquefois cette répétition rend le discours rude & moins	41 Cependant il en faut @évi- « ter la rencontre grop fr@équente dans une m@ême
42 agréable. C'est quelquefois un substantif masculin. Cet A est mal formé. On	42 p@ériode. . Quel _{flxcfois cexce r@ép@@éution rend le difcours rude &@@@@ moin:
43 dit par une façon de parler proverbiale : il n'a pas fait une panse d'a, pour	43 agr@éable. = .Cclbt quickquefois un fubftantif mafculin. Cet A cft mal formé "
44 dire, il n'a pas formé une seule lettre, & figurément, il n'a fait quoi	44 On dit par nne fa@gon de parler proverbiale : il n'a pas fai@ une pan@c diur,
45 que ce soit. On dit aussi dans la conversation familière : Il ne sait ni A ni	45 pour dire , il n'a pas form@é une feule lettre: &@@@@ figurément , il r'a
46 B, pour exprimer un ignorant. Ci-dessous gât Mr l'Abbé, Qui ne savoit ni A ni	46 fair@quoi que ce foit. On dit aufli dans la converfarion famili@ère : Il ne fai@
47 B. Mâgnag. C'est aussi la troisième personne du verbe auxiliaire avoir. Il a	47 ni@4 ni B , pour expri . Aner un ignorank, S U Ci-deffous g@it Mr I Abb@é, e
48 fait de l'Acclat mal-à-propos. L'imagination du Poète n'a pu vous peindre si	48 Qai@fic@[avf]t@@idA@ni@B. 'M@finAC. i C@eft aulli la tro@@ifi@m% perfonne du verbe
49 belle que vous êtes. Voit. La vérité, qui a des bornes, a dit pour vous tout	49 auxiliaire avoir. 11 . "~ faic de l'@éclair mal-a@-propos. L'imagination du
50 ce que le mensonge, qui n'en connoît point, a inventé pour les autres. S. Evr.	50 Po@cte n'a pt wous peindre fi belle que: vous @&tes. Vorr. La v@érité , qu * a
	51 des bornes , « dit pour vous rout ce que le menfonge, qu _ n@en conno@it point,

Évaluation des résultats

L'analyse des résultats obtenus révèle que :

- Il y a une amélioration significative dans la détection des contours.
- Les contours prédits par nos modèles ne sont pas toujours parfaits, ce qui entraîne des parties de texte manquantes.
- Notre OCR Tesseract a tendance à commettre de nombreuses erreurs, notamment des confusions avec les "s" longs.
- De plus, notre OCR Tesseract a également tendance à halluciner, c'est-à-dire à reconnaître des éléments inexistantes.

Conclusion et perspectives

Conclusion

L'expérimentation des outils qui nous ont été proposés, tels que Tesseract, Layout Parser, Laypa, Grobid et Surya, nous a permis de découvrir un outil intéressant : Layout Parser avec Detectron2, qui semble, dans un premier temps, être le plus efficace sur Trevoux nous a inspirer une méthode.

Après l'expérimentation de cette méthodologie, nous avons constaté une nette amélioration des résultats de notre outil, et la marge de progrès semble significative au vu de la quantité de nos échantillons.

De plus, ce stage m'a permis de m'initier au monde de la recherche, mais aussi de développer différentes compétences, parmi lesquelles la prise en main rapide d'un outil grâce à sa documentation, ainsi que l'apprentissage de la bonne documentation d'un travail.

Conclusion et perspectives

Perspectives:

- **Amélioration des échantillons d'entraînement:** Continuer à enrichir et perfectionner la qualité et la quantité de nos échantillons d'entraînement afin d'optimiser les performances du modèle.
- **Entraînement de sous-modèles :** Pour améliorer nos détections, l'entraînement de plusieurs sous-modèles, à l'instar du modèle XP2, pourrait nous permettre d'obtenir des résultats plus précis et fiables.
- **Exploration d'autres OCR :** Explorer d'autres OCR spécifiquement adaptés aux textes en français ancien pourrait également constituer une piste prometteuse pour améliorer la précision des reconnaissances.

Merci!



Présenté par Edina ADJARO PATOUSSI
30 Août 2024

Tuteur laboratoire:
Julien VELCIN
Ludovic MONCLA

Tuteur enseignant:
Valentin Lachand-Pascal