

Exercice Scientifique et technique

Amélioration et migration d'un outil d'extraction de données PDF par
IA

DOROTHEE Ethan

Sommaire

Contexte et problématique

Architecture

Amélioration

Validation

Décision

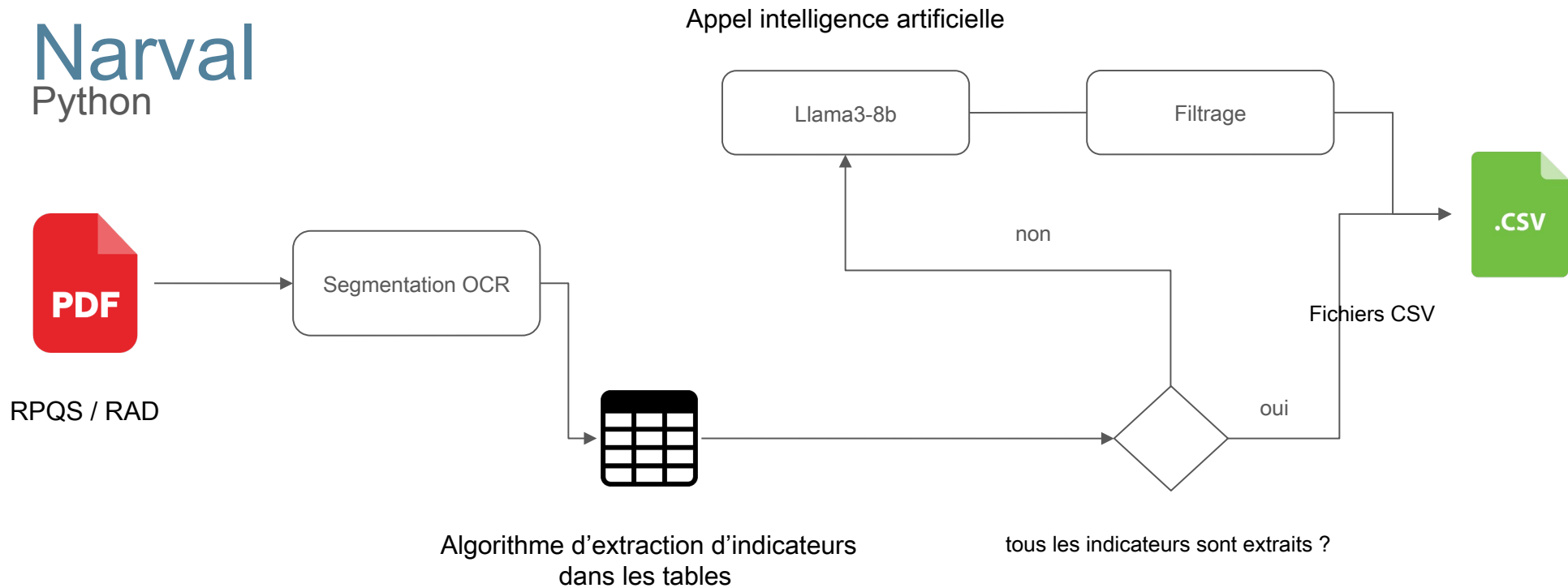
Bibliographie

Contexte et problématique



Contexte et problématique

Narval
Python



Contexte et problématique

Comment transformer l'amélioration de ce pipeline Python en une POC objective pour déterminer la nécessité d'un portage vers Java ?

Architecture



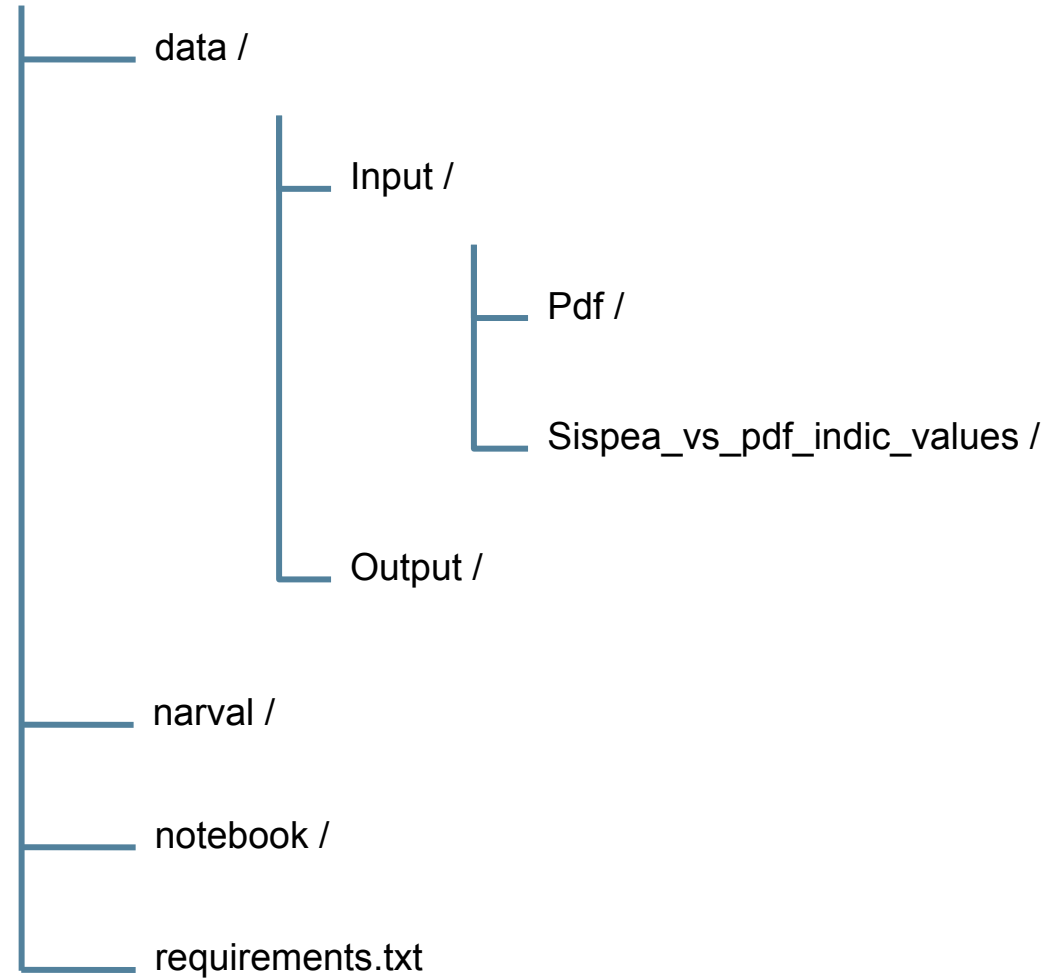
Hugging Face



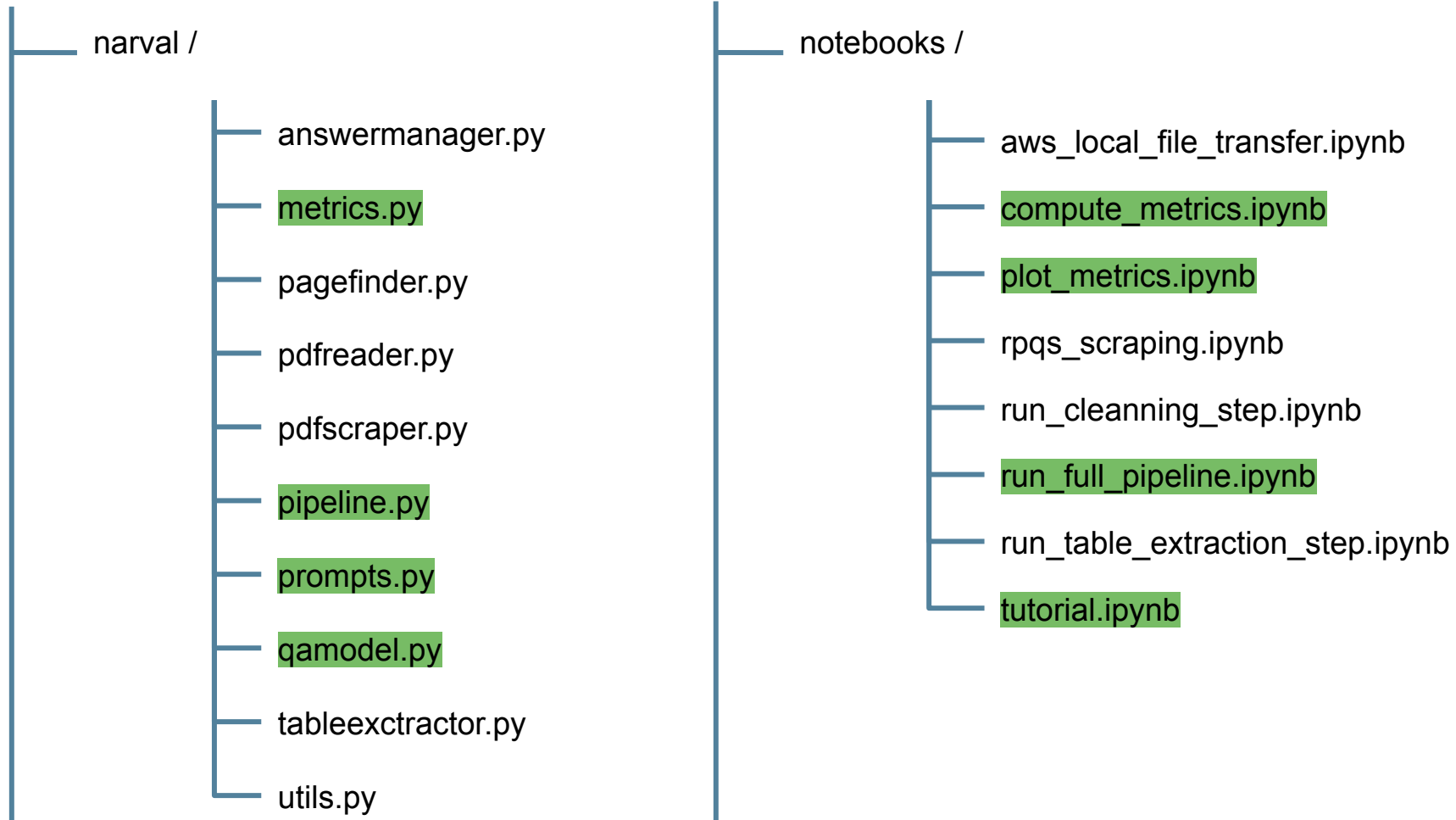
SSP Cloud
Onyxia

PyPDF2

PyPDF2 + pdfPlumber

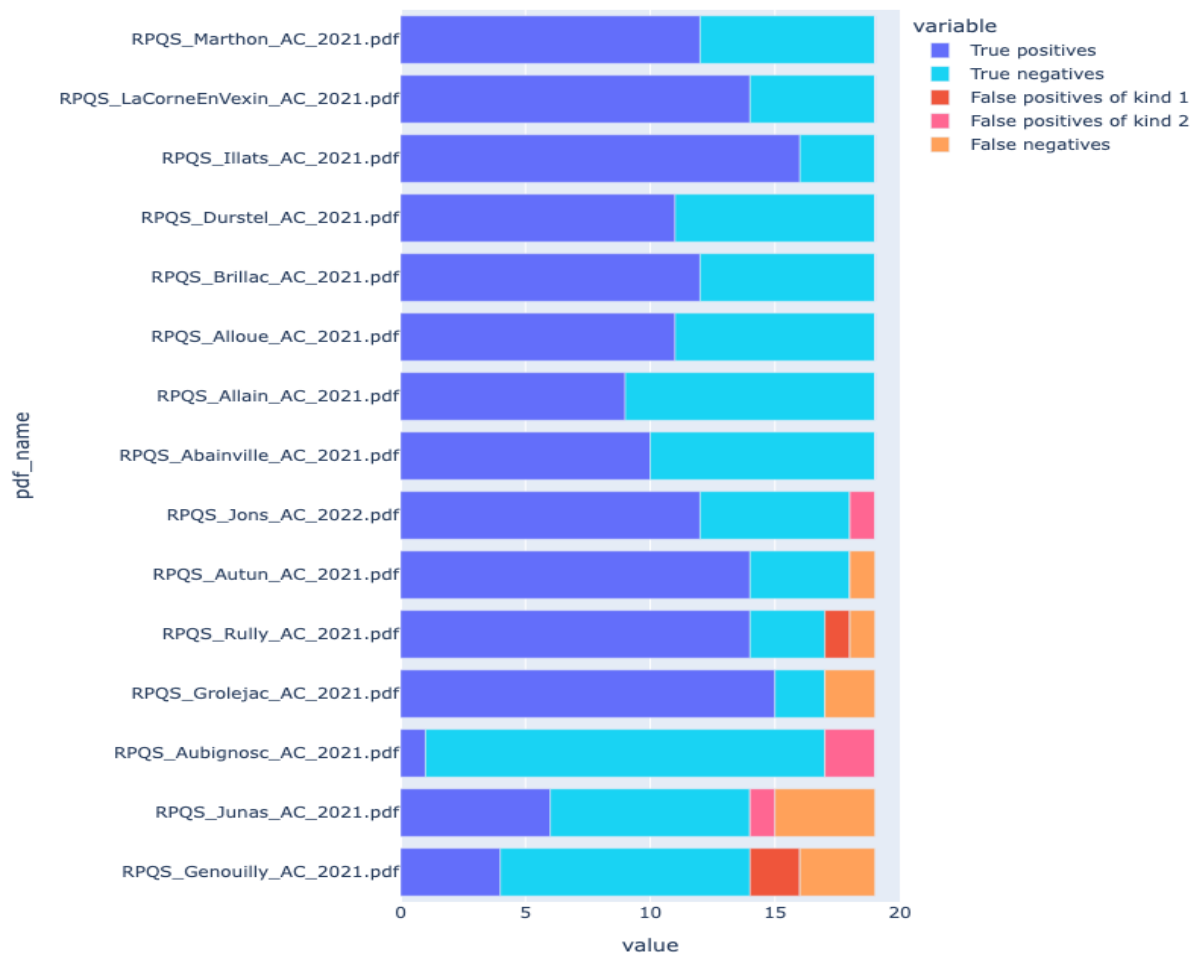


Architecture

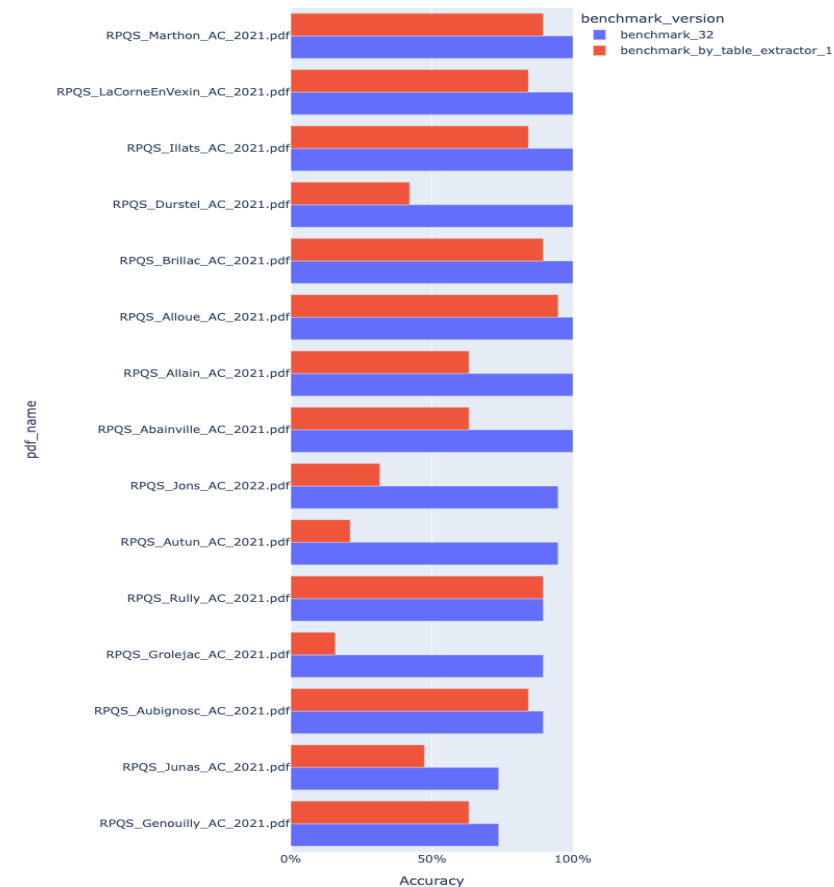


Metrique référence

Metrics per pdf for benchmark_32



Accuracy per pdf (vs_pdf)



Axe de qualité

IA et prompts

Librairie OCR

Axe d'extension des capacités

Traitement des images

Parallélisme : Comités de juges

Défis et limites



Problème des données de test

Type de documents	Avec tableaux	Sans tableau
RPQS (données saisi dans SISPEA)	11 fichiers	4 fichiers
RPQS ou RAD (données non saisi dans SISPEA)	15 fichiers	15 fichiers

Défis et limites

Limites techniques actuelles



Maintenabilité du code



SSP Cloud
Onyxia

Difficulté de validation statistique

Test de validation

Classification

Catégorie	Définition
Vrai positif (VP)	Narval renvoie une valeur identique à celle du pdf
Vrai négatif (VN)	Narval ne renvoie pas de valeur et l'indicateur est absent du PDF
Faux positif de type 1 (FP1)	Narval renvoie un nombre alors que l'indicateur est absent
Faux positif de type 2 (FP2)	Narval renvoie une valeur incorrecte (présent dans le pdf mais pas la bonne)
Faux Négatif	Narval ne trouve pas la valeur alors qu'elle est présente dans le PDF

Test de validation

Formules des métriques

Métrique	Formule de calcul
Justesse Global (Accuracy)	$(VP + VN) / (VP + VN + FP1 + FP2 + FN)$
Précision	$VP / (VP + FP1 + FP2)$
Rappel	$VP / (VP + FN + FP2)$

Test de validation

Critère évalué	référence	seuil	Mesuré
Justesse Global (Accuracy)	89,9%	90 %	[A mesurer]
Précision	90,3%	95 %	[A mesurer]
Rappel	83,9%	80%	[A mesurer]

Bibliographie

- Github du projet initial (Geneviève Fleury):
https://github.com/gefleury/ofb_narval/blob/main/rapport_narval.pdf
- Site de diffusion de SISPEA :
<https://www.services.eaufrance.fr/>
- Site de la librairie Hugging face :
<https://huggingface.co/>
- Documentation du SSP cloud :
<https://www.sspcloud.fr/catalog?path=SSPCloud%E2%90%A3Documentation>