Classification Automatique des documents

Plan

1. Démarche de Travail
2. Périmètre Fonctionnel
3. Solutions possibles/ préconisées
4. Socle Technique & Architecture
5. Annexes:Concepts de base

Démarche de travail

Démarche de Travail

Périmètre Fonctionnel

Périmètre Fonctionnel

01

Classification par type de document

Détermination automatique du type du document selon les catégories déjà définies par domaine

02

Implémentation d'une API

Solutions

- Solutions possibles
- Fonctionnement des solutions possibles
- Solutions préconisées
- Processus détaillé des solutions préconisées

Classification par domaine / par Type de document

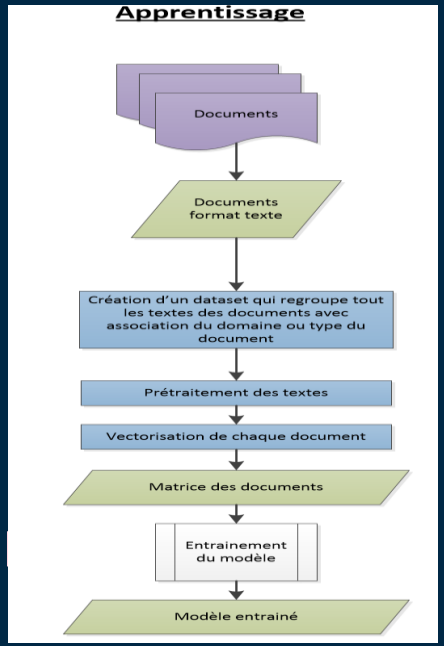
Solutions possibles

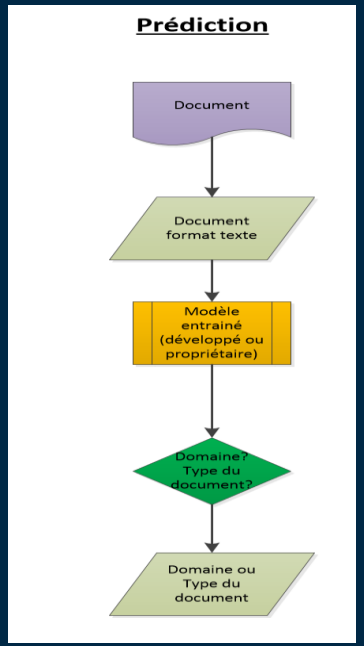
Solutions propriétaires ML & NLP	Développement des méthodes basées sur les règles	Développement des méthodes de NLP & ML	Développement des Méthodes de NLP& DL
• AutoML NLP de Google • Microsoft Cognitive services • Amazon SageMaker • Datakeen • Parashift's Classification system • IBM Watson™ Knowledge Studio	• Nécessité de définir les règles pour chaque type de document	• Nécessite une phase de prétraitement des données • Une bonne capacité d'apprentissage, d'où leur efficacité	• Phase de prétraitement moins sophistiquée • Capacité d'apprentissage plus puissante que celle de ML

Fonctionnement des Solutions Possibles:

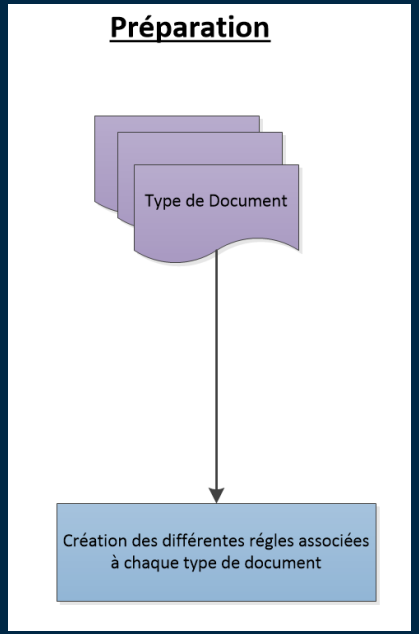
Classification par domaine / par Type de document

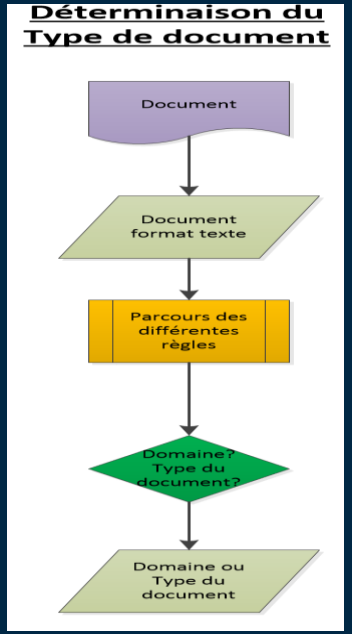
Fonctionnement des Méthodes NLP &ML | NLP &DL





Fonctionnement des méthodes à base des règles





Solution Préconisée :

Classification par domaine / par Type de document

Solution préconisée:
Développement des méthodes de ML & NLP

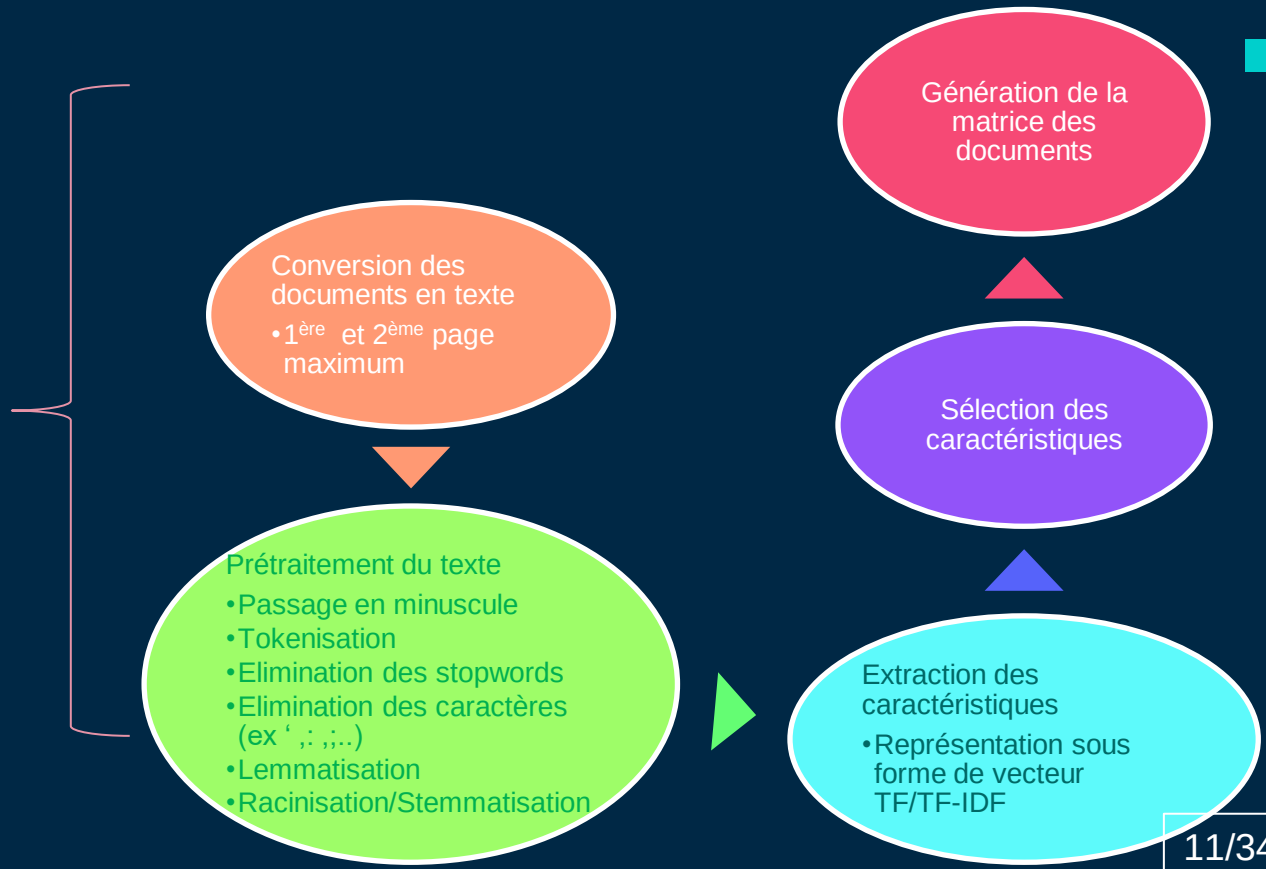
Argumentaire

- En raison de la multitude de domaines et des types de documents, l'implémentation des méthodes basées sur les règles peuvent s'avérer inefficace et peuvent être confronter à la difficulté de définir les règles.
- En raison du nombre limité de données dont on dispose, l'implémentation des méthodes de DL peuvent s'avérer inefficace
- Les solutions propriétaires nécessitent l'engagement d'un coût associé (Par exemple Entraînement : 3 $ par heure, Déploiement : 0,05 $ par heure, Prédiction : 5,00 $ par tranche de 1 000 enregistrements texte)

Processus de préparation du DataSet:

Solution préconisée : Développement des méthodes de ML & NLP

Phase de préparation des données à refaire en arrière plan pour augmenter la taille du Dataset afin d'améliorer les performances du modèle d'apprentissage

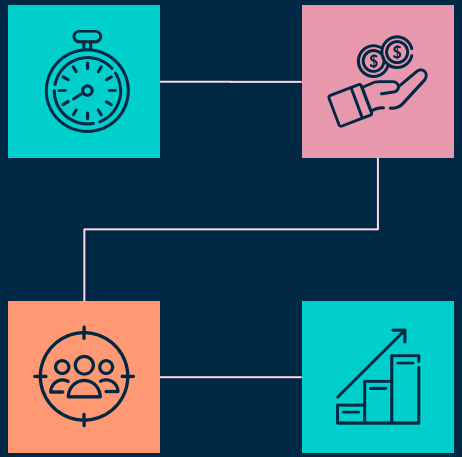


Processus de Construction du Modèle d'Apprentissage:

Solution préconisée : Développement des méthodes de ML & NLP

Définir un modèle propre d'apprentissage automatique

Dataset,ML, NLP



Entrainement du modèle

Application de plusieurs classifieurs(KNN,SVM, Naive Bayes,Random Forest,Logistic Regression, Decision Tree, SGD Classifier)

Evaluation du modèle

Choisir le classifieur le plus optimal en terme de précision et d'hyperparamètre

Validation du modèle

Jeu de test

L'entrainement est à refaire après chaque mise à jour du dataset

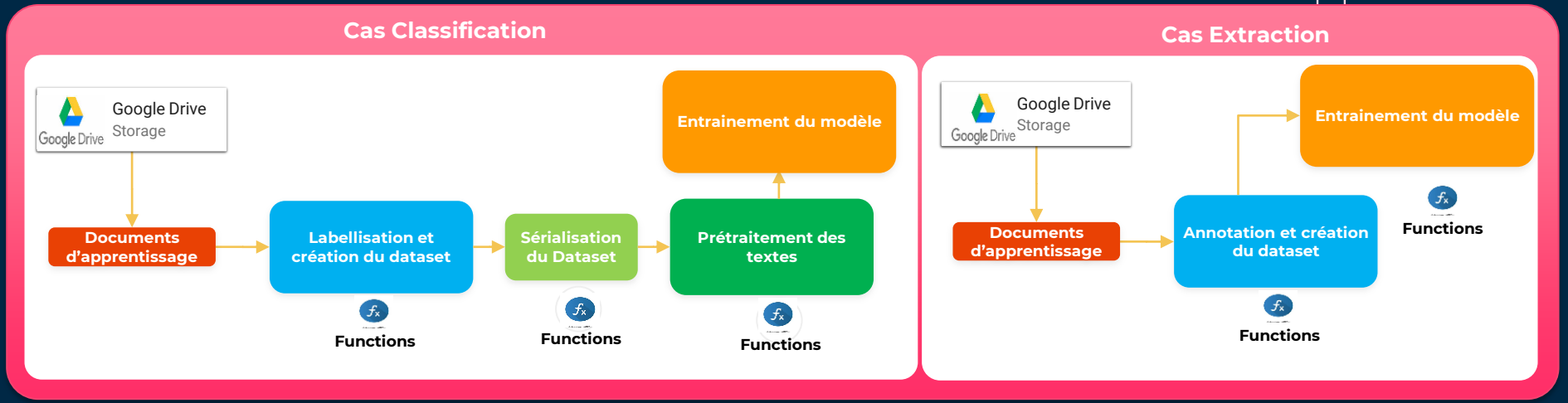
Solution Préconisée :

Réalisation d'une API de classification et d'extraction

L'API va être utilisée par l'application GED pour :

- Faire la préparation du Dataset
- Lancement de l'apprentissage
- Prédiction du domaine d'un document
- Prédiction du type d'un document
- Extraction des données d'un document

Préparation du Dataset / Entrainement du modèle



Classification d'un document

Architecture Technique

- Socle Technique
- Diagramme d'architecture

Socle technique

Modèles Machines Learning

- K-nearest Neighbors (KNN)
- Support Vector Machines (SVM)
- Naïve Bayes
- RandomForest
- Logistic Regression

Conversion & Annotation des documents

Tesseract OCR

PRP VBP RB Web Anno

Librairies Python ML & NLP & NER

pandas

scikit learn Machine Learning with Scikit-Learn

RegEx Regular Expression

PyTesseract OCR WITH TESSERACT, OPENCV AND PYTHON

NumPy

python Natural Language Analyses with NLTK

PDFTRON PDF2Image

OpenCV

python Pickle

XLRD

spaCy

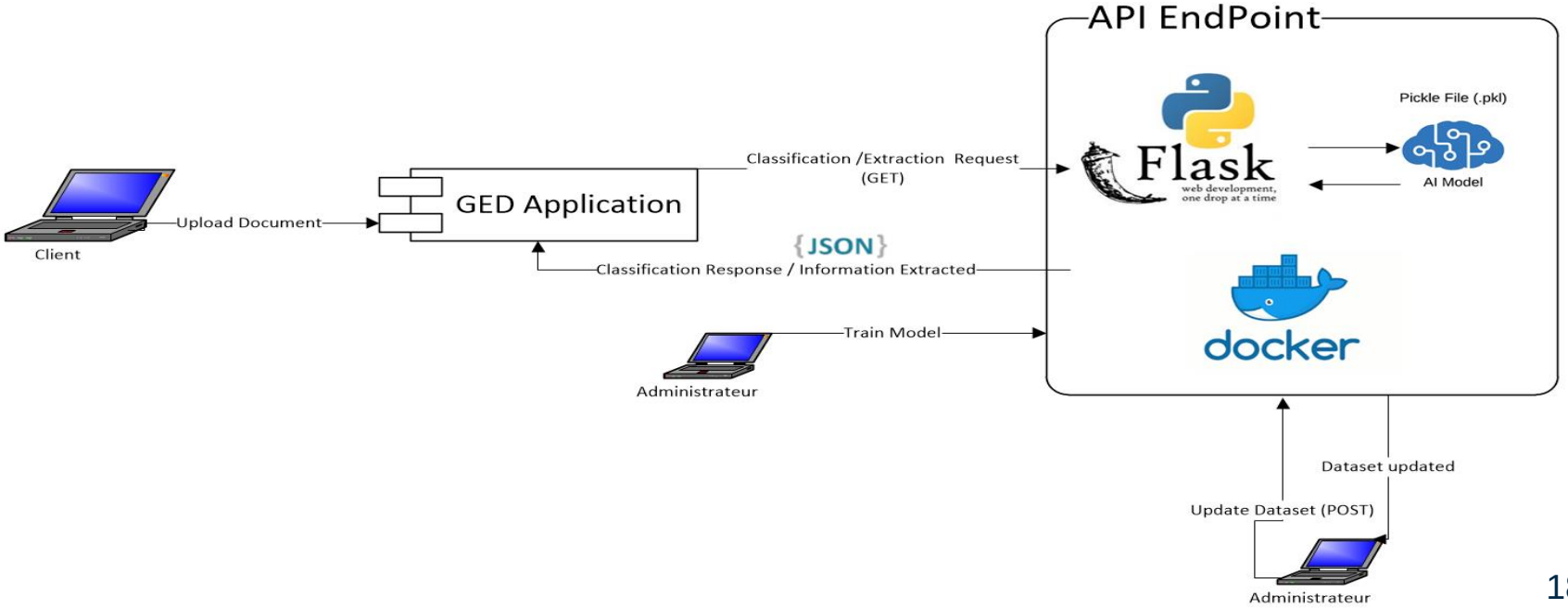
pillow

Framework de développement

Flask

Python

Diagramme d'Architecture



Annexes
Concepts de base

- Machine Learning (ML) / Apprentissage automatique
- Deep Learning (DL) / Apprentissage profond
- Critères de choix DP vs ML
- Natural Language processing (NLP) / Traitement Automatique du Langage Naturel (TALN)
- Named Entity Recognition (NER) / Reconnaissance des entités nommées
- Règles de classification

Machine Learning (ML) / Apprentissage automatique

Un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques et qui se base sur l'apprentissage à partir des exemples et expériences passées pour résoudre des problèmes.

Deep Learning (DL) / Apprentissage profond

Un champ d'étude de l'intelligence artificielle qui adopte une approche de réseau neuronal pour rechercher des modèles et des corrélations, qui sont moins dirigés et plus gourmand en terme des données.

Critères de choix DP vs ML

	Maching Learning	Deep Learning
Volume de données optimal	1000 enregistrements	1M enregistrements
Résultat escompté	Résultat numérique, comme une classification ou un score	Tout types de valeurs y compris du texte en langage naturel pour sous-titrer une image ou un son ajouté à un film muet
Comment ça fonctionne	Utilise différents algorithmes d'apprentissage afin de prédire des futures actions sur la base des précédentes données	Utilise un réseau de neurones qui transmet les données à de nombreuses couches de traitement pour interpréter les caractéristiques et les relations entre les données
Comment c'est géré	Les algorithmes sont dirigés par un dataAnalyst pour examiner des variables spécifiques dans le dataset	Les algorithmes sont autogérés sur l'analyse des données

Natural Language processing (NLP) / Traitement Automatique du Langage Naturel (TALN)

C'est un domaine à l'intersection du Machine Learning et de la linguistique.
Il a pour but d'extraire des informations et une signification d'un contenu textuel.

Le NLP recouvre plusieurs champs d'application à savoir:

- Traduction/Correction/Résumé automatique
- Sentiment analysis
- Marketing(Recherche des prospects)
- ChatBot
- Classification

Le NLP est composé de trois étapes:

- Prétraitement
- Représentation du texte sous forme de vecteur
- Classification en cas de besoin

Named Entity Recognition (NER)

C'est une application du NLP qui consiste à reconnaître des entités nommées (Named Entities) dans un corpus (ensemble de textes) et de leur attribuer une étiquette telle que "nom", "lieu", "date", "email", etc. .

le NER a de nombreuses applications outre l'indexation de documents. Il est notamment utilisé dans les systèmes de questions-réponses (Q&A) qui consistent à répondre à une question posée en langage naturel en recherchant la réponse dans une collection de documents ou une base de connaissance.

Pour cet usage, le NER peut s'avérer utile pour déterminer le type de réponse que le système Q&A doit retourner en se basant sur des entités retrouvées dans la question (ex. un lieu, une date).

Règles de classification

Avec la classification basée sur des règles, le contenu est classé selon des règles prédéfinies.
Il peut s'avérer difficile dans certains cas de définir des critères spécifiques