Rapport Phase 3 du Projet de Parsing de CV

Titre : Extraction d'entités nommées dans des CV à laide de SpaCy

Objectif de la phase

Lobjectif principal de cette phase est didentifier automatiquement les entités clés dans les CV (telles que les noms, les dates, les entreprises, les lieux, etc.) et de les catégoriser dans des classes définies à laide des techniques de Reconnaissance dEntités Nommées (NER).

Travail effectué

1. Chargement du modèle pré-entraîné de SpaCy (en_core_web_sm) :
   - Permet de détecter des entités telles que PERSON, ORG, DATE, GPE, etc.
   - Modèle léger, adapté pour une première implémentation rapide et efficace.

2. Extraction de texte depuis les CV PDF :
   - À laide de PyPDF2, les fichiers PDF sont convertis en texte brut pour traitement.

3. Application du modèle NER :
   - Le texte est traité par le modèle SpaCy pour détecter les entités.
   - Les entités extraites sont regroupées et affichées par type.

4. Interface avec Streamlit :
   - Application web simple permettant duploader un ou plusieurs fichiers.
   - Affichage dynamique du texte extrait et des entités détectées.
   - Présentation des entités sous forme de tableau lisible.

5. Architecture modulaire :
   - app.py : interface utilisateur (Streamlit)

- utils.py : fonctions de traitement (extraction, NER, formatage)

Technologies et bibliothèques utilisées

- Python 3.12
- SpaCy 3.7.2
- Modèle SpaCy : en_core_web_sm
- PyPDF2 pour lextraction du texte
- Pandas pour la structuration des entités extraites
- Streamlit pour linterface utilisateur

Limites et améliorations futures

- Le modèle utilisé est généraliste : il nest pas optimisé spécifiquement pour les CV.
- Étape suivante : entraîner un modèle NER personnalisé sur des jeux de données annotés.
- Ajouter un système de métriques pour mesurer la performance (précision, rappel, F-score).

Livrables

- Code source disponible sur GitHub : https://github.com/ndialvaye/resume-parsing-phase3
- Application déployée sur Streamlit Cloud (si activation réussie).