Rapport – Phase 4 : Extraction de Relations

Objectif de la phase

L'objectif de cette phase est d'associer les entités extraites des CVs aux relations pertinentes entre elles, telles que les relations entre personnes, entreprises, dates, diplômes, ou rôles professionnels. Cela permet de structurer davantage les informations pour des traitements automatisés plus efficaces.

Approche technique

Cette étape repose sur les techniques de Named Entity Recognition (NER) et de dependency parsing fournies par SpaCy. Le modèle en_core_web_sm est utilisé pour détecter les entités, tandis que le parsing des dépendances grammaticales permet d'extraire les relations entre ces entités. Les entités sont associées si elles partagent une relation verbale ou structurelle (ex: 'travaille chez', 'étudie à', etc.).

Techniques et bibliothèques utilisées

- SpaCy pour NER et le dependency parsing

- Streamlit pour l'interface utilisateur

- PyPDF2 et python-docx pour lire les fichiers .pdf et .docx

- openpyxl pour l'export des résultats en Excel

Fonctionnalités de l'application

- Téléversement de fichiers PDF ou DOCX de CV
- Extraction des entités nommées (personnes, entreprises, institutions...)
- Extraction des relations entité-entité via parsing grammatical
- Affichage des relations extraites dans un tableau interactif
- Export des relations dans un fichier Excel (.xlsx)

Validation

Les relations extraites peuvent être comparées manuellement aux données de référence (ground truth) pour valider leur pertinence. L'utilisateur peut également affiner les règles pour améliorer la précision.