Sujet 1 – Extraction de graphes de connaissances à partir de texte

Idée naïve de résolution :

Extraction d'entités nommées (les sommets / ressources) :

D'abord, il nous faut prétraiter le ou les textes afin d'enlever toutes les informations superflues (titres, signature, formule de politesse, ponctuation, etc.). Donc on aura une fonction qui aura pour tâche d'analyser les textes et de supprimer ce qu'on veut.

Entrée : un document sous forme textuelle.
Sortie : le document en format raccourci rassemblant l'essentiel du document.

De plus, on découpe le texte en phrases contenant les entités.
Entrés : le texte complet
Sortie : une liste de phrase contenant les entités.

Ensuite, à l'aide d'une bibliothèque de traitement du langage NLP tels que SpaCy, Flair, NLTK, etc., on extrait des textes toutes les entités possibles (personne, organisation, date, lieu, etc.).

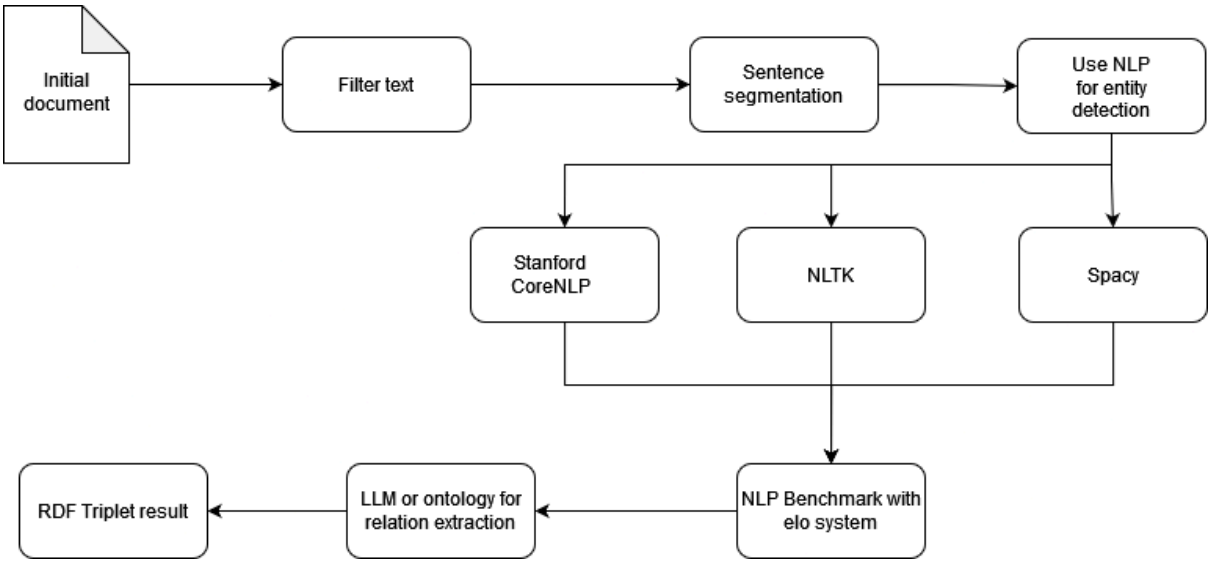
- Benchmark des différentes bibliothèques disponibles.
- Uniformisation des entités trouver (majuscule, dénomination, etc.). Rassembler les entités selon un certain delta de ressemblance.
- Faire une analyse croisée de toutes les bibliothèques pour trouver la meilleure de toutes.

Chaque entités extraites forment un sommet du graphes
Sortie : Un liste, tableau regroupant toutes les entités trouvées.

Extraction de liens entre ces entités (les arcs / propriétés)

On utilise une ontologie adaptée pour extraire les relations entre les entités.
Sinon on peut essayer de voir s'il existe des LLM pré-entraînés sur les sujets des textes.

Chaque relation obtenue forment une arête du graphe et relieront deux entités.
Entrée : La phrase contenant deux entités.
Sortie : Un triplet de la forme [ent1, rel, ent2]



Pour notre projet, les articles D1-2, D1-3, D2-2, D3-2 et D4-1 sont les plus pertinents.