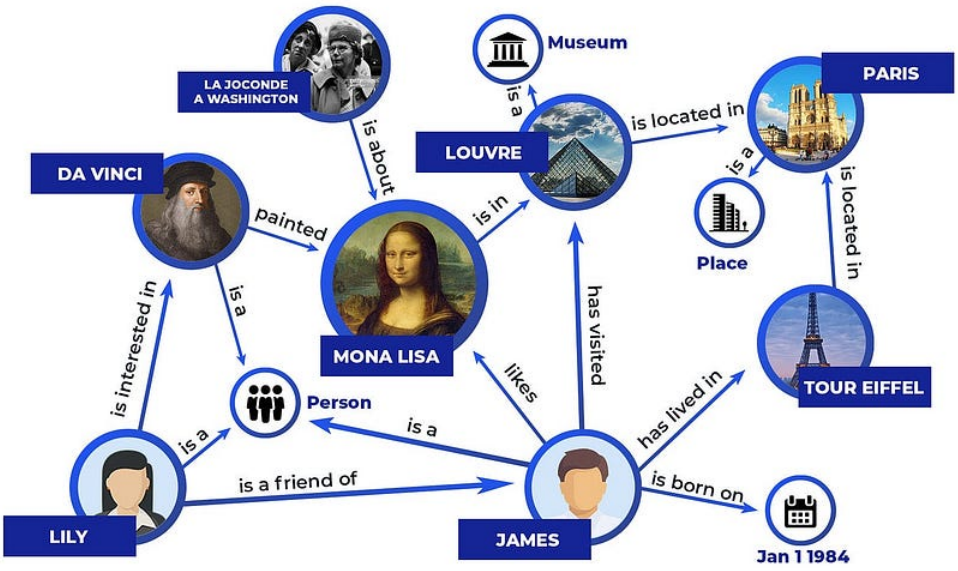
Contact : Olha NAHORNA, IR CNRS, BSE (Bordeaux Sciences Economiques)
olha.nahorna@u-bordeaux.fr

Transformer les articles scientifiques au graphe de connaissance

Extraire l'information pertinente dans le contenu textuelle est une tâche très importante pour mener différentes analyses dans les sciences économiques. Les liens sémantiques contiennent une information importante qui peut approfondir l'analyse et mieux comprendre les relations. Dans ce projet je vous propose d'analyser le contenu des articles scientifiques en économie dans l'objectif de représenter l'information le plus pertinent sous une forme de graphe de connaissance, c'est-à-dire sous une forme de représentation graphique basés sur les technologies du web-sémantique.



La tâche de ce projet est de convertir le texte non-structure vers les triplets <sujet-prédicat-objet> RDF (Resource Description Framework) en respectant les recommandations W3C. Il s'agit également de lier les entités entre eux et de les sauvegarder dans une base de données adaptée. Il existe différentes façons d'extraire et de normaliser les données textuelles, dans ce projet je vous propose de tester les dernières modèles LLM (large language models), de les finetuner si besoin, ou de créer vos propres modèles deep learning, ou pourquoi pas d'essayer le réseau de neurones graphe.

Dans le premier temps on va traiter les articles en anglais et en français, mais la solution doit être facilement adaptable au traitement des articles rédigés sur d'autres langues. La solution envisagée doit être en mesure de traiter plusieurs millions des articles, ainsi la question de la performance est très importante à prendre en compte. La solution proposée doit être adaptée à l'infrastructure universitaire (on peut exclure les solutions basées sur les infrastructures amazon, azure ou autres nécessitant le payement mensuel).

Langage de programmation : python

Quelques articles pour commencer :

https://medium.com/@jack16900/generative-knowledge-graph-construction-a-review-947521309ec5
https://app.ingemmet.gob.pe/biblioteca/pdf/SP-558-02.pdf
https://github.com/nicolas-hbt/pygraft
https://github.com/Fcabla/DBpedia-abstracts-to-RDF