Prénom et nom : Alpha Bah

Matricule ULB :

TP1 :Web scraping et APIs

Dans le cadre de ce TP, j’ai procédé comme suit :

1. Utilisé mon compte GitHub et forké le repository contenant tout le code pour ce cours : <https://github.com/madewild/tac>
2. Mis en place d’un site HTML basique avec GitHub Pages sur la base des explications fournies en annexe.

Voir <https://AlphaBah93.github.io/>

1. Téléchargé sur l’UV le sous-ensemble du corpus CAMille et exploré au hasard quelques fichiers pour avoir une première impression de la langue utilisée, des thèmes abordes, de la longueur des textes, de la qualité de l’OCR, etc.
2. Choisis une requête SPARQL sur Wikidata qui puisse être utile pour les chercheurs en journalisme utilisant le corpus CAMille. Cette requête permet d’obtenir la liste des politiciens belges, de langue francophone et ayant obtenu le titre ou grade universitaire de type doctorat.

Voir <https://w.wiki/7FX2>

1. La requête a été implémenté dans un court script (sur la base de s2\_sparql.ipynb). Ensuite j’ai pushé le résultat mon repo GitHub dans un dossier nomme « tp1 »
2. Sur la base du catalogue des API sur <https://rapidapi.com/blog/most-popular-api/> j’ai choisis « **Text Entities API**», une API de reconnaissance d'entité qui renvoie les entités clés dans un texte, y compris les événements, le lieu, l'organisation, la personne, les titres, les quantités et bien plus encore. Voir <https://rapidapi.com/jjespinozala/api/text-entities-api/>

La reconnaissance d'entités nommées permet d'extraire automatiquement ces entités nommées d'un document texte. Cette tâche est utile pour une variété de tâches telles que la recherche d'informations, la réponse aux questions et la synthèse de texte.

1. Sur la base de s3\_api.ipynb, j’ai écrit un court script, qui devrait permette d’interroger cette API. Le script a été pushé sur mon repo GitHub dans le dossier tp1 et porte le nom de tp1\_api.ipynb