**Exercice noté RI : Récupération et annotation de documents**

**Rapport**

Pour procéder à la récupération des documents et à leur annotation, voici les différentes étapes suivies :

Sur Europresse la requête choisie pour obtenir les documents était : *TEXT= "changement climatique" & impact* permettant ainsi d’avoir 1 000 articles sur un sujet d’actualité concernant l’écologie.

Une fois les documents récupérés au format .html et déposés sur le dépôt Git, leur traitement a entièrement été réalisé avec le script suivant examen\_1\_p1\_NERC.ipynb.

Dans un premier temps, le fichier .html a été converti en fichiers individuels .xml pour chaque article. Le script utilisé est celui fournit sur le dépôt Git du cours, à savoir le script euro2solr.py.

Ensuite, les EN ont été annoté avec NERC-fr à l’aide du script tag\_entities.py disponible également via le dépôt Git du cours. Un test, non présent dans les documents fournis, a été effectué avec SpaCy mais les résultats de NERC-fr semblaient plus concluants c’est pourquoi ce sont les résultats conservés.

Enfin, les différents dossiers générés par ces étapes sont téléchargés.

Il n’y a pas d’étape d’annotation des documents avec des expressions clés car je n’ai pas réussi à avoir un code fonctionnel permettant d’appliquer les données fournies dans le dépôt Git du cours. Je n’arrivais pas à créer de script permettant de gérer l’entièreté des documents .xml.

L’étape de nettoyage a été réalisée manuellement, c’est pourquoi il reste forcément des erreurs d’annotations des EN dans les documents finaux. Néanmoins, quelques grosses erreurs récurrentes ont pu être nettoyées comme la présence de caractère en trop (comme des parenthèses, des points, des crochets) ou encore de mots supplémentaires ou de mauvaise annotation. Au vu de la grande quantité de ressources à traiter, je n’ai pas pu regarder en détail chaque entité pour vérifier l’exactitude de l’annotation.