Projet de fin de semestre

Description générale

Ce projet consiste à donner des outils d'analyse sur un data set portant sur les articles publiés dans arxiv. Ce site web est une plateforme de publication d'article scientifique, les auteurs et les citations y sont référencé.

Provenance des données et méthodes d'obtention

Le site web arxiv.org fourni une api rest sur l'url http://export.arxiv.org/api
Le script *api_rest_xml_generator.py* est example de requête en script python.

Une documentation de l'api est accessible sur l'url : https://arxiv.org/help/api/index

Structures attendues

- Charger de l'ensemble des données en une base de données mongoDB.
- Créer d'une base de données néo4j avec comme nœuds :
  - l'id de l'article.
  - l'auteur.
  - Les affiliation (les laboratoires, ou institus des auteurs.)
  - Les categories
  - tout autre nœud que vous pourriez trouver utiles.

Livrables attendus

créer un attribut neo4j qui donne le nombre d'auteurs affilier à chaque établissement de recherche.

faire une fonction python qui affiche les statistiques d'un auteur :

- le nombre d'article publié,
- le ou les institus dont il dépend,
- tous les liens vers ses articles,
- tous les auteurs avec lesquels il a collaboré.

Créer une fonction qui pour deux auteurs affiche les éléments communs (si oui, un article sur lequel ils ont collaboré, si oui s'ils travaillent dans le même institut…)

Étendre la fonction à un deuxième niveau de profondeur.

N'hésitez pas à nous montrer toutes autres informations ou fonctions utiles sur ces données.

Documents attendu :

- Le code complet transférer sur le site github, celui-ci doit marcher en démarrant un docker-compose et en lançant un minimum de commande, voir aucune commande.

- Un Readme clair spécifiant :

    - La structure des bases de données.
    - L’utilité de chaque script et leur fonctionnement.
    - Les commandes à lancer pour exécuter les différentes fonctions pythons.

- Une présentation détaillant :
    - le set de données.
    - les bases de données et leur structure.
    - les réponses aux livrables.
    - d’autres choses intéressantes s’il y en a.