Project 3.A

I. Project Description

In this project, you will implement the database for LitCovid corpus in MongoDB using Python connector. LitCovid (https://www.ncbi.nlm.nih.gov/research/coronavirus/) is a comprehensive collections of scientific literature about COVID-19. It contains over 300k PubMed articles and is updated daily with new PubMed articles that are relevant to COVID-19. The data provided to you in project 3 is a small subset of this collection.

Specifically, you are asked to write scripts in Jupyter Notebooks (python 3, template is provided) for the following tasks.

1. ImportData [Points: 10]

Create the MongoDB database and import data. The data are provided as json files which you can download.

2. Query [Points: 90]

This script queries the following information

1) Count the number of documents in this corpus collection
2) List the fields for the first document in this corpus
3) Count the number of publications for each journal. Sort the result in descending order and print journals with more than 4 publications
4) Find all papers published in PLoS One journal. Print their years and titles
5) Count the number of publications for each author. Sort the results in descending order and return authors with 5 or more publications
6) Find the papers co-written by ‘Wang J’ and 'Zhang L', print the paper pmids, journal names and titles
7) Create text index on passages.text
8) count the number of publications that contains the phrase "COVID-19 Vaccine"
9) count the number of publications that contains the words "COVID-19" or "Sars-CoV-2"
10) count the number of publications that contains the words "COVID-19" and "Sars-CoV-2"

Submission Instruction

Rename project3.A template.ipynb to project3.A.ipynb. Submit this file to Gradescope.