Introduction to Data Science

21KDL

Lab04 - Web Scraping & Regular Expression

Deadline: **23h59 - 27/05/2023**

Submitting via Google Form: https://forms.gle/8TwefFaesQwef6nw9

Tools and Language: Python / Jupyter Notebook

PROBLEM

Web Scraping:

Truy cập vào đường dẫn: https://arxiv.org/list/cs.AI/recent và sử dụng BeautifulSoup để thực hiện các yêu cầu sau đây:

1. Lấy các đường dẫn đến các bài báo.
2. Đi đến đường dẫn mỗi bài báo, trích xuất các thông tin như: tên bài báo, tên tác giả, abstract của bài báo, subjects, đường dẫn download bài báo.
3. Lưu tất cả các thông tin trên của tất cả bài báo thành định dạng file .csv, vơi các column là: Title, Authors, Abstract, Subjects, DownloadUrl

Regular Expression:

Sử dụng file văn bản pháp luật tại đường dẫn:
https://drive.google.com/file/d/1-6vz7gt0jJXrAJcfMq4Qxp3iuKY6NfIR/view?usp=share_link

Sử dụng Regex thực hiện các yêu cầu sau:

1. Trích xuất tất cả các định dạng ngày tháng năm có trong văn bản.
2. Split văn bản thành danh sách các điều luật.
   VD: [
   “Điều 1. Phạm vi điều chỉnh và đối tượng áp dụng
   1. Thông tư này quy định mã số, tiêu chuẩn chức danh nghề nghiệp và xếp lương viên chức chuyên ngành tuyên truyền viên văn hóa.
   2. Thông tư này áp dụng đối với viên chức tuyên truyền viên văn hóa làm việc trong các đơn vị sự nghiệp công lập và các tổ chức, cá nhân có liên quan.”

   “Điều 2. Mã số các chức danh nghề nghiệp viên chức chuyên ngành tuyên truyền viên văn hóa
   1. Tuyên truyền viên văn hóa chính Mã số: V.10.10.34
   2. Tuyên truyền viên văn hóa Mã số: V.10.10.35
   3. Tuyên truyền viên văn hóa trung cấp Mã số: V.10.10.36”,

   “Điều 3. …”
   ]

3. Tìm tất cả các mã luật có trong văn bản theo định dạng tương tự như: “204/2004/NĐ-CP”

NOTICE:

- Nộp bài tập thông qua Google Form.
- Nộp sai yêu cầu sau deadline xem như chưa nộp.

- Named **CORRECTLY** your notebook by the following pattern: DS2023_Lab<LabID>_<StudentID>_<StudentName>.ipynb.
    - Example:
      File name: DS2023_Lab01_21280075_NguyenVanA.ipynb

- Inside the coding file, there should be a brief introduction (as in the example below).

  """
  Introduction to Data Science
  Programming Exercise: 01
  Name: Nguyen Van A
  Student ID: 21280075
  """

There is NO acceptance for cheating or copying.