Ton Duc Thang University
Faculty of Information Technology

ĐỀ TÀI TÌM HIỂU

Môn: Xử lý dữ liệu lớn

I. Hình thức

- Đề tài được thực hiện theo nhóm **04 – 05** sinh viên.
- Nhóm sinh viên thực hiện các yêu cầu và nộp bài theo hướng dẫn chi tiết bên dưới.

II. Yêu cầu

Nhóm sinh viên thực hiện tìm hiểu ngôn ngữ Python và các thư viện hỗ trợ để thu thập và phân tích mã HTML của các trang web từ đó xây dựng đồ thị liên kết giữa các trang và tìm ra pagerank tương ứng.

- Thuyết trình lần 1: Nhóm sinh tìm hiểu và báo cáo các nội dung sau.
    - Thư viện download mã html của trang web theo URL.
    - Thư viện đọc và phân tích mã HTML.
    - Chương trình Python nhận vào một URL nguồn, một prefix và số lượng giới hạn tiến hành thu thập mã HTML để phân tích, kết xuất ra một DataFrame (pyspark.sql) với 2 cột [Page] và [Successor].
        - Mỗi dòng của DataFrame chứa hai URL a (Page) và b (Successor).
        - Trong mã nguồn của trang a, có xuất hiện b.
        - a, b phải chứa **domain** được nhập vào, ví dụ: tdtu.edu.vn
        - Số lượng dòng của DataFrame không vượt quá con số giới hạn nhập vào.
- Thuyết trình lần 2: Nhóm sinh viên viết chương trình Python thực hiện các yêu cầu sau dựa vào DataFrame đã có ở lần 1.
    - Thêm vào DataFrame cột [Out-degree] để thống kê số lượng URL đi ra từ mỗi page.
    - Thêm vào DataFrame cột [Dead-ends] với giá trị là 1 nếu page là dead-ends và 0 nếu không phải.
    - Thêm vào DataFrame cột [PageRank] với giá trị là pagerank của mỗi URL.

III. Hướng dẫn thuyết trình

Nhóm sinh viên làm bài thuyết trình trong thời lượng 10 phút (khoảng 15-20 slide) để trình bày.

IV. Hướng dẫn nộp bài

- Tạo thư mục với tên theo cú pháp <MSSV1>_<MSSV2>_... (danh sách mã số sinh viên), trong đó gồm:
  - <MSSV1>_<MSSV2>_….ipynb chứa mã nguồn của chương trình.
  - <MSSV1>_<MSSV2>_....pdf kết xuất pdf mã nguồn đồ án từ Google Colab.
  - present.pdf: bài thuyết trình báo cáo kết quả tìm hiệu, lập trình và thực thi của nhóm.
- Lưu ý giữ lại kết quả thực thi của các ô trong cả hai tập tin .ipynb và .pdf.
- Nén thư mục thành <MSSV1>_<MSSV2>…..zip và nộp theo deadline.

V. Thời gian nộp bài

- Thời hạn nộp bài và thuyết trình được thông báo trên hệ thống.
- Sinh viên nộp trễ hơn hạn trên đồng nghĩa với việc **0.0 điểm** giữa.

VI. Quy định

- **Mọi hành vi sao chép code trên mạng, chép bài bạn hoặc cho bạn chép bài nếu bị phát hiện đều sẽ bị điểm 0.0.**
- **Nếu bài làm của sinh viên có dấu hiệu sao chép trên mạng hoặc sao chép nhau, sinh viên sẽ được gọi lên phỏng vấn code để chứng minh bài làm là của mình.**

-- HẾT --