DỮ LIỆU LỚN

YÊU CẦU BÀI TẬP LỚN

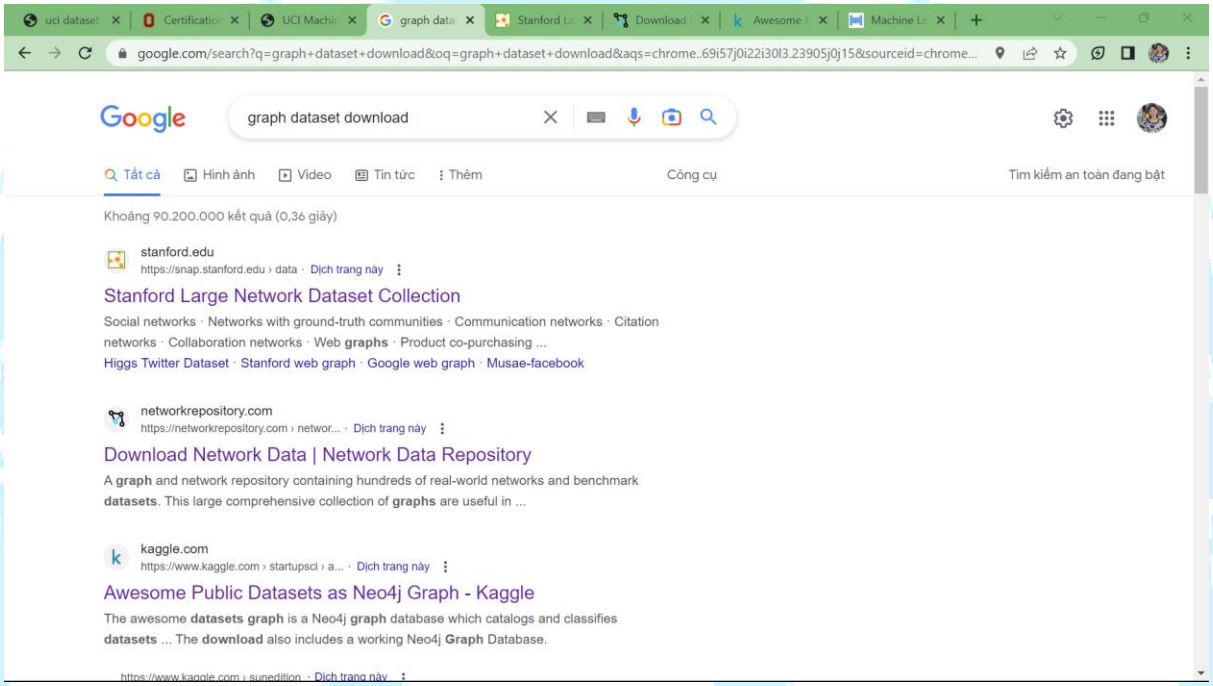
Mỗi nhóm 2 sinh viên thực hiện các yêu cầu sau:

1. Tìm hiểu và lựa chọn một bài toán thực tế, thu thập dữ liệu và viết mô tả về bài toán và tập dữ liệu.
2. Thiết kế cấu trúc cơ sở dữ liệu đồ thị cho tập dữ liệu đã thu thập được và cài đặt vào phần mềm Neo4j. (*Chú ý: Đồ thị có ít nhất 100 đỉnh và có ít nhất 1 loại quan hệ giữa các đỉnh*)
3. Phân tích được ít nhất 6 thông tin hoặc mẫu dữ liệu hữu ích từ CSDL đã xây dựng được ở câu 2. Các nhóm có thể viết một ứng dụng nhỏ về khai thác thông tin của CSDL đã cài đặt.
4. Viết và báo cáo kết quả thực hiện các yêu cầu trên trong buổi thi cuối kì.

Chú ý: Trong buổi thi cuối kì, ngoài việc báo cáo bài tập lớn, sinh viên sẽ phải trả lời thêm một số câu hỏi lý thuyết.

Hướng dẫn tìm nguồn dữ liệu về đồ thị

Tìm kiếm trên google search với cum từ: Graph dataset download



Một số link có nhiều nguồn dữ liệu:

https://snap.stanford.edu/data/

https://networkrepository.com/network-data.php

https://www.kaggle.com/datasets/startupsci/awesome-datasets-graph

https://paperswithcode.com/datasets?mod=graphs

Tài liệu tham khảo về phân tích đồ thị trên Neo4j

Một số chủ đề phân tích:

- Community detection
- Node embedding
- Node classification pipeline
- Link prediction

1.
https://neo4j.com/docs/graph-data-science/current/

2.
Từ trang https://neo4j.com/fr/, chọn Data scientists, chọn graphAcadamy for Data Science, đăng kí học miễn phí 3 khóa học (ngắn) về phân tích dữ liệuđồ thị trên No4j như trang bên dưới: