ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐỀ THI KẾT THÚC HỌC KỲ

Năm học: **2022 - 2023**; Học kỳ: **I**
Môn học: **Lập trình xử lý dữ liệu với Python** (AIT2023 1)

Hình thức thi: Tự luận (Thực hành)
Link nộp bài:

1. Nhóm thực hành 1: `https://forms.gle/C4kKwt7ib3aFuZ9d6`
2. Nhóm thực hành 2: `https://forms.gle/tskLnNkJ614A8v2P8`
3. Nhóm thực hành 3: `https://forms.gle/nLqcgNkJnyMAGWuR6`
4. Nhóm thực hành 4: `https://forms.gle/Z96QjcDiRHcZakG46`

Link chọn tài khoản Facebook được phân tích (không chọn trùng nhau):

- `https://bit.ly/3tzJXAC`

Điểm: 100/100.

Đề bài: Phân tích tương tác và nội dung của một/nhiều tài khoản Facebook

1 Mục tiêu:

Mục tiêu của nội dung kiểm tra này là giúp sinh viên làm việc thành thạo với các công cụ thu thập nội dung từ các trang mạng xã hội (Facebook), thu thập dữ liệu, và phân tích các dữ liệu thu thập được. Sau khi hoàn thành bài kiểm tra, sinh viên có thể thực hiện các công việc sau:

1. Sử dụng các công cụ thu thập nội dung bao gồm bài viết, tương tác, bình luận từ Facebook: Selenium, facebook-scraper, ...
2. Làm sạch, tiền xử lý các dữ liệu sau khi thu thập qua các công cụ kể trên.
3. Phân tích và trực quan hóa dữ liệu thu thập được sử dụng các phương pháp và công cụ thích hợp.
4. Giải thích kết quả và rút ra kết luận từ các phân tích được đưa ra.

Mô tả ngắn gọn: Công việc của sinh viên là phân tích các tương tác và nội dung của một/nhiều tài khoản Facebook bất kỳ do sinh viên chọn. Lưu ý, các sinh viên không chọn trùng tài khoản được phân tích với nhau, đăng ký và kiểm tra tài khoản được các bạn lựa chọn tại: `https://bit.ly/3tzJXAC`. Sinh viên thu thập dữ liệu từ một/nhiều tài khoản Facebook, nên là Fanpage, có thể bao gồm: bài đăng, bình luận, lượt tương tác (reactions, lượt bình luận, lượt chia sẻ, ...), v.v. Sau đó, sinh viên xử lý dữ liệu mình thu thập được, phân tích và rút ra các những xét của mình. Mục đích của bài kiểm tra này không chỉ để thể hiện kỹ năng sử dụng công cụ có sẵn mà còn thể hiện khả năng diễn giải và trình bày dữ liệu một cách có ý nghĩa.

2 Công việc cần thực hiện

1. **Chọn một hoặc nhiều Fanpage trên Facebook mà bạn quan tâm**. Lưu ý, các Fanpage này nên có trên 300.000 lượt thích để đảm bảo số lượng bài viết và tương tác (bình luận, chia sẻ, reactions) đủ và đa dạng cho việc phân tích dữ liệu ở các bước sau. **Yêu cầu cơ bản mỗi bạn sinh viên thu thập dữ liệu của chỉ một tài khoản Facebook. Trong trường hợp sinh viên muốn so sánh sự tương quan giữa các Fanpage khác nhau, có thể thu thập dữ liệu từ nhiều Fanpage.**

2. **Thu thập dữ liệu**: Sử dụng các công cụ có sẵn Selenium, facebook-scraper, hay bất kỳ công cụ nào bạn tìm thấy để thu thập nội dung từ Fanpage mà sinh viên quan tâm. Các thông tin được thu thập nên có tối thiểu các trường sau: Nội dung bài đăng, thời gian bài đăng, số lượt bình luận, số lượt reactions và từng loại reaction tương ứng, số lượt chia sẻ, một phần / tất cả bình luận trong từng bài đăng. **Lưu ý, các Fanpage lớn có lượng dữ liệu khổng lồ đặc biệt là số bình luận trong mỗi bài viết. Do vậy, việc thu thập toàn bộ dữ liệu từ khi Fanpage được lập ra là không cần thiết. Sinh viên chỉ cần thu thập các dữ liệu liên quan đến các bài đăng trong tối thiểu khoảng thời gian 01 tháng và số bài đăng tối thiểu 100 bài.**

3. **Làm sạch và tiền xử lý dữ liệu**:

   (a) **Làm sạch dữ liệu**: Dữ liệu thu thập ở bước 2 thường chứa nhiều khuyết thiếu đến từ lỗi của công cụ thu thập hay bản thân bài đăng trên Fanpage (chẳng hạn bài đăng không chứa nội dung, chưa có bình luận, hạn chế quyền truy cập bởi token, ...). Do vậy, sinh viên cần xem xét dữ liệu thô mình đã thu thập để loại bỏ đi các thông tin không liên quan hoặc không cần thiết trong bộ dữ liệu.

   (b) **Tiền xử lý dữ liệu**: Tổ chức lại bộ dữ liệu theo cách phù hợp cho bước phân tích

4. **Phân tích dữ liệu**: Một số câu hỏi tiềm năng để phân tích như: Đâu là bài viết có lượt tương tác lớn nhất trong bộ dữ liệu?; Các mốc thời gian trong ngày mà

Fanpage thường xuyên đăng bài viết?; Số lượt tương tác trong các bài đăng thay đổi như thế nào?; Đâu là các từ khóa xuất hiện nhiều nhất trong các bài đăng được thu thập?; Sự tương quan giữa số lượng reactions với các trường khác như số lượng bình luận, độ dài bài viết?; v.v. Ngoài ra, nếu có dữ liệu từ hai trang Fanpage mà giả sử là hai trang tin tức, sự tương quan giữa nội dung của hai trang Fanpage như thế nào?; v.v.

5. **Trực quan hóa dữ liệu (Visualization)**: Sử dụng các kỹ thuật trực quan hóa dữ liệu phù hợp để trình bày những phát hiện của sinh viên. Có thể sử dụng biểu đồ, đồ thị, bản đồ nhiệt, v.v.

6. **Diễn giải và Kết luận**: Dựa trên phân tích của kể trên, diễn giải những phát hiện của sinh viên và rút ra kết luận.

7. **Viết báo cáo**: Viết một báo cáo chi tiết về phương pháp, các phát hiện, và kết luận của sinh viên. Bao gồm các hình ảnh trực quan và tài liệu tham khảo thích hợp.

3 Sản phẩm cần bàn giao

Mỗi sinh viên sẽ upload các file dưới đây vào một repo cá nhân tương tự như các bài thực hành hàng tuần.

1. Sinh viên nộp lại code với Jupyter Notebook hoặc tương đương chứa các phần xử lý. Dữ liệu thô, dữ liệu đã làm sạch, và dữ liệu sau khi tiền xử lý cũng cần được nộp lại.

2. Trong trường hợp file dữ liệu quá lớn (lớn hơn dung lượng Github cho phép upload), có thể thêm file '`dataset.txt`' chứa link Google drive / Dropbox / v.v. trỏ đến nơi lưu trữ của file dữ liệu.

3. Báo cáo hoàn thiện.

4 Lưu ý và Gợi ý

1. Các công cụ thu thập dữ liệu Facebook sử dụng macro rất dễ bị Facebook quét và chặn quyền truy cập, nặng hơn là khóa tài khoản. Vì vậy, sinh viên tạo một tài khoản Facebook mới để lấy 'token' để thực hiện bài kiểm tra này.

2. Có thể sử dụng các công cụ học máy mà sinh viên tự mình tìm hiểu để phân loại, ngoại suy ra những kết luận (đi kèm visualizations) thú vị từ dữ liệu thu thập được.

3. Với trường `user_ID` mà sinh viên có thể thu thập trong phần bình luận của các bài viết. Sinh viên có thể thu thập thêm được các thông tin cá nhân (public) của họ nhằm rút ra một vài nét chung trong lượng người dùng thường xuyên tương tác với Fanpage được quan tâm.

5 Cách tính điểm

- Thu thập dữ liệu (**20 điểm**): Thu thập dữ liệu ít nhất đủ với các trường mà đề yêu cầu.
- Làm sạch và tiền xử lý dữ liệu (**20 điểm**): Làm sạch kỹ lưỡng và sắp xếp dữ liệu hợp lý để phân tích.
- Phân tích dữ liệu (**20 điểm**): Phân tích dữ liệu hiệu quả và trả lời các câu hỏi đề xuất.
- Trực quan hóa dữ liệu (**20 điểm**): Tạo hình ảnh trực quan rõ ràng, sâu sắc.
- Giải thích và Kết luận (**10 điểm**): Diễn giải chính xác các phát hiện và rút ra kết luận có ý nghĩa.
- Viết báo cáo (**10 điểm**): Viết báo cáo rõ ràng, kỹ lưỡng, và chuyên nghiệp.

Lưu ý: Điểm thưởng có thể được trao cho những phát hiện đặc biệt sâu sắc, kỹ thuật phân tích sáng tạo hoặc trình bày kết quả đặc biệt.

Vui lòng đảm bảo báo cáo cho bài kiểm tra này là do tự bạn thực hiện, được tham khảo chính xác và tuân thủ chính sách liêm chính trong học tập của trường đại học.

Chúc may mắn!