Tổng kết và đồ án cuối kỳ

Trần Trung Kiên (ttkien@fit.hcmus.edu.vn)

Cập nhật lần cuối: June 12, 2021

fit@hcmus

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Nhìn lại: đã học được gì rồi

Môn này học về các công cụ hỗ trợ thực hiện qui trình Khoa Học Dữ Liệu (KHDL)

- Các câu lệnh Linux
- Git & Github
- Conda
- Jupyter Notebook
- Markdown
- Python
- Matplotlib
- Numpy
- Pandas

- Đưa ra câu hỏi có ý nghĩa cần trả lời
- Thu thập dữ liệu (phạm vi môn học: dùng dữ liệu có sẵn, hoặc có thể thu thập được một cách tương đối dễ dàng)
- Khám phá dữ liệu
- Tiền xử lý dữ liệu
- Phân tích dữ liệu (phạm vi môn học: phân tích đơn giản, tức là dùng tính toán và trực quan hóa để trả lời cho các câu hỏi về thông tin *nằm trong* dữ liệu mà ta có)

  $\longrightarrow$ câu trả lời
- Truyền thông kết quả / ra quyết định

Để thực hiện tốt qui trình này thì nhà KHDL cần:

- Sử dụng thành thạo các công cụ hỗ trợ trên máy tính
- *Luôn luôn bình tĩnh, khách quan, thành thật*

Q: Còn gì nữa là hoàn thành môn học?

A:

- HW3 (thời lượng ~ 2 tuần tính từ hôm nay)
- Đồ án cuối kỳ (thời lượng ~ 4 tuần tính từ hôm nay)

Đồ án cuối kỳ - đại ý

Tìm dữ liệu đã được public (ví dụ, trên Kaggle) về chủ đề mà nhóm bạn thấy hứng thú, khám phá dữ liệu (trong quá trình khám phá dữ liệu, có thể cần dừng lại để tiền xử lý rồi mới tiếp tục khám phá tiếp được), xác định các câu hỏi có thể được trả lời bằng dữ liệu, tiền xử lý (nếu cần) và phân tích để trả lời cho mỗi câu hỏi

Đồ án cuối kỳ - các nội dung cần trình bày trong file notebook

1. Thu thập dữ liệu

- Dữ liệu của bạn là về chủ đề gì và bạn lấy từ nguồn nào?
- Người ta có cho phép bạn dùng dữ liệu như này không? Ví dụ, bạn có thể xem thử license của dữ liệu là gì
- Người ta thu thập dữ liệu như thế nào?

2. Khám phá dữ liệu (thường đan xen với tiền xử lý)

- Dữ liệu có bao nhiêu dòng và bao nhiêu cột?
- Mỗi dòng có ý nghĩa gì? Có vấn đề các dòng có ý nghĩa khác nhau không?
- Dữ liệu có các dòng bị lặp không?
- Mỗi cột có ý nghĩa gì?
- Mỗi cột hiện đang có kiểu dữ liệu gì? Có cột nào có kiểu dữ liệu chưa phù hợp để có thể xử lý tiếp không?
- Với mỗi cột có kiểu dữ liệu dạng số (numeric), các giá trị được phân bố như thế nào?
    - Số-lượng/tỉ-lệ các giá trị thiếu?
    - Min? max? $\rightarrow$ Dựa vào min, max, và ý nghĩa của cột để phát hiện giá trị bất hợp lệ
- Với mỗi cột có kiểu dữ liệu dạng phân loại (categorical), các giá trị được phân bố như thế nào?
    - Số-lượng/tỉ-lệ các giá trị thiếu?
    - Số lượng các giá trị khác nhau? Show một vài giá trị

3. Đưa ra các câu hỏi có ý nghĩa cần trả lời

Nhóm bạn cần đưa ra $\geq$ số-lượng-thành-viên-của-nhóm câu hỏi mà có thể được trả lời bằng dữ liệu. Câu hỏi cần có ý nghĩa (nếu trả lời được câu hỏi thì sẽ có lợi ích gì) và không nên quá dễ trả lời (ví dụ, chỉ cần một dòng code là có thể ra được câu trả lời thì mình nghĩ là câu hỏi quá dễ). Nhóm bạn nên tập trung vào chất lượng câu hỏi hơn là số lượng câu hỏi

Trong file notebook, với mỗi câu hỏi, nhóm bạn cần trình bày:

- Câu hỏi là gì?
- Nếu trả lời được câu hỏi thì sẽ có lợi ích gì?

Tiền xử lý + phân tích dữ liệu để trả lời cho từng câu hỏi

Với mỗi câu hỏi:

- Có cần tiền xử lý dữ liệu không và nếu có thì nhóm bạn tiền xử lý như thế nào?
  - Text: vạch ra các bước thực hiện một cách rõ ràng và dễ hiểu sao cho nếu người đọc không đọc code thì vẫn có thể hiểu được cách bạn tiền xử lý
  - Code: cài đặt các bước đã vạch ra ở trên. Nhóm bạn cũng cố gắng viết code cho rõ ràng và dễ đọc (chọn tên biến gợi nhớ, comment ở những chỗ mà nên comment, không để một dòng code quá dài)
- Nhóm bạn phân tích dữ liệu như thế nào để ra được câu trả lời cho câu hỏi?
  - Text: tương tự như trên
  - Code: tương tự như trên

5. Nhìn lại quá trình làm đồ án

Sau bao ngày vất vả làm đồ án thì bây giờ đã kết thúc. Bây giờ là lúc để ngồi uống coffee và tĩnh tâm nhìn lại một xíu :-)

- Mỗi thành viên: Đã gặp những khó khăn gì? (Hay mọi chuyện đều thuận lợi)
- Mỗi thành viên: Có học được gì hữu ích? (Hay không học được gì)
- Nhóm: Nếu có thêm thời gian thì sẽ làm gì?

Phần này có sao thì bạn nói vậy thôi, chứ không phải là viết cho có, hoặc tự chế ra để nghe cho hay

6. Tài liệu tham khảo

Để hoàn thành đồ án thì nhóm bạn đã tham khảo những tài liệu nào?

Một số ý nói chung:

- Trong file notebook, một việc quan trọng mà nhóm bạn cần cố gắng rèn luyện là tổ chức các ý và viết-lách/code một cách ngăn nắp, rõ ràng, dễ đọc. Nhóm bạn nên dùng các heading để tổ chức các ý, và nên dùng các extension của jupyter notebook/lab để có thể xem cũng như là nhanh chóng đi đến các heading (giống như bookmark trong file pdf). Trong quá trình viết-lách/code, cố gắng giữ một tinh thần tốt, nghĩ cho người đọc
- Trong các bước thực hiện đã nói, mình nghĩ phần chiếm nhiều thời gian nhất là có lẽ là từ bước tìm dữ liệu cho đến bước đưa ra được câu hỏi

Đồ án cuối kỳ - làm việc nhóm

Khi làm việc nhóm, nên có một file kế hoạch (ví dụ, file Google Sheets) về các công việc cần làm, lượng thời gian cho mỗi công việc, phân công ai làm công việc nào (file này có thể sẽ cần điều chỉnh nhiều lần trong suốt quá trình làm việc)

Nhóm bạn sẽ dùng Git và Github để hỗ trợ quản lý phiên bản và làm việc nhóm

Nhóm bạn cần đảm bảo:

- Khối lượng công việc của mỗi thành viên trong nhóm là khá tương đương nhau (được thể hiện bằng lịch sử commit trên Github)
- Mỗi thành viên trong nhóm đều hiểu phần làm của nhau

Để đạt được yêu cầu làm việc nhóm này, với mỗi bước trong qui trình KHDL:

- Các thành viên trong nhóm sẽ cùng làm bước này và làm độc lập với nhau $\leftrightarrow$ trong lịch sử trên Github: mỗi thành viên sẽ tạo một branch và làm ở đó (trước đó, một thành viên trong nhóm sẽ soạn file notebook khung chứa các đề mục và push lên Github cho cả nhóm)
- Sau đó nhóm sẽ họp để hiểu file notebook của mỗi thành viên và đưa ra phiên bản sau cùng của file notebook $\leftrightarrow$ trong lịch sử trên Github: merge các branch lại với nhau

Đối với bước tiền xử lý + phân tích dữ liệu để trả lời cho mỗi câu hỏi thì mình cho phép các thành viên làm các câu hỏi khác nhau, nhưng cần đảm bảo mọi người đều hiểu phần làm của nhau

Điểm đồ án là chung cho cả nhóm và sẽ xấp xỉ bằng trung bình điểm của các thành viên; nếu trong nhóm có bạn nào làm ít và/hoặc không hiểu file notebook của nhóm thì điểm chung sẽ bị kéo xuống

$\rightarrow$ Mỗi thành viên trong nhóm cần phải biết nghĩ cho cả nhóm:

- Các bạn học tốt cần cố gắng hỗ trợ (hỗ trợ chứ không phải làm thay) cho các bạn học yếu
- Các bạn học yếu cần cố gắng để không làm ảnh hưởng tới các bạn học tốt

Đồ án cuối kỳ - góp ý cho đồ án của nhóm khác

Gọi x là ngày vấn đáp (dự kiến: x = 10/7)

Trước 9h sáng ngày x-2: tất cả các nhóm sẽ upload đồ án lên Github, đồng thời mở một "issue" ở Github để thông báo là đã hoàn thành và nhờ các bạn góp ý

Trong ngày x-2: mỗi nhóm (tất cả thành viên chứ không phải là một người đại diện) sẽ đọc và góp ý cho một nhóm khác (mình sẽ phân công sau) bằng chức năng "issue" của Github

Mục tiêu của việc góp ý?
Để giúp nhóm khác hoàn thiện đồ án hơn trước buổi báo cáo, bằng cách cung cấp các góc nhìn của người ngoài cuộc mà người trong cuộc có thể không thấy được. Thông qua quá trình đọc và góp ý cho nhóm khác thì thường bạn cũng sẽ học hỏi được một số điều hữu ích

Đồ án cuối kỳ - nộp bài trên moodle và vấn đáp

Ngày x-1: các nhóm hoàn thiện đồ án dựa trên góp ý của nhóm khác, cập nhật trên Github và nộp bài trên Moodle. Nhóm bạn sẽ cần nộp: file kế hoạch trong quá trình làm việc nhóm (ở đầu ghi link tới thùng chứa Github), file notebook, file dữ liệu (nếu nặng thì có thể up lên đâu đó và để link trong file txt)

Ngày x: vấn đáp online (lịch cụ thể của từng nhóm sẽ được thông báo sau). Mỗi nhóm sẽ có 15 phút để trình bày trên file notebook và 5 phút để hỏi đáp (ai trình bày phần nào là do mình chỉ định) $\rightarrow$ cần tập luyện trình bày trước, khi trình bày trên file notebook thì hạn chế tối đa việc giải thích chi tiết code (ví dụ, bạn có thể dùng chức năng ẩn các code cell của Jupyter Lab, khi Thầy hỏi chi tiết thì mới show)

Slide cuối

Thank you :-)