TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN

TÀI LIỆU HƯỚNG DẪN THỰC HÀNH
Môn: Khai thác dữ liệu

YÊU CẦU ĐỒ ÁN MÔN HỌC

I. Mục tiêu đồ án

1. Nhận diện và phát biểu được bài toán khai thác dữ liệu.
2. Hiểu được các đặc điểm và áp dụng các kỹ thuật tiền xử lý trên dữ liệu.
3. Mô tả và áp dụng các giải thuật khai thác dữ liệu để giải quyết bài toán đã trình bày.
4. Đánh giá, so sánh được kết quả giữa các phương pháp đã áp dụng.

II. Yêu cầu về dữ liệu và bài toán

- Sinh viên chọn nguồn dữ liệu có tối thiểu 15 thuộc tính x 10.000 dòng (hoặc 500 MB hình ảnh/âm thanh, 200 MB văn bản) và phải ghi rõ nguồn gốc. Có thể tham khảo một số kho dữ liệu trong phần Tài liệu tham khảo dưới đây. Báo cáo phải mô tả được các thành phần của dữ liệu, giá trị cao/thấp nhất, phổ biến/hiếm nhất, giá trị trung bình, trung vị…
- Đồ án môn học phải có bài toán cụ thể, được phát biểu rõ ràng. Sinh viên dựa trên bài toán này đưa ra các phương pháp giải quyết từ đó áp dụng các giải thuật khai thác dữ liệu.

III. Yêu cầu tiền xử lý dữ liệu

Dựa trên đặc điểm của dữ liệu, bài toán và giải thuật đã chọn, sinh viên phải lựa chọn các kỹ thuật tiền xử lý dữ liệu dự định sẽ áp dụng. Trong bản báo cáo phải giải thích được lý do lựa chọn và trình bày chi tiết về cách thức áp dụng, kết quả của từng kỹ thuật.

IV. Yêu cầu sử dụng giải thuật khai thác dữ liệu

- Sinh viên phải mô tả chi tiết trong báo cáo và giải thích được lý do chọn lựa phương pháp khai thác đáp ứng được bài toán và phù hợp với dữ liệu.
- Sinh viên được khuyến khích lập trình để hiểu rõ hơn về các giải thuật hoặc sử dụng các công cụ có sẵn để thực hiện việc khai thác dữ liệu (trừ Weka).

V. Yêu cầu đánh giá kết quả

- Kết quả của các giải thuật phải được đánh giá, so sánh bằng những độ đo phù hợp.
- Trong bản báo cáo phải trình bày cụ thể về nội dung, công thức, ý nghĩa và lý do lựa chọn từng độ đo.
- Sau khi đánh giá, so sánh kết quả, sinh viên phải nêu ra được những ưu điểm/hạn chế, kết luận và hướng phát triển mà đồ án cần hướng tới.
- Sinh viên được khuyến khích viết một ứng dụng sử dụng kết quả của việc khai thác dữ liệu để làm rõ được áp dụng của bài toán trong thực tế.

VI. Yêu cầu về hình thức

Gợi ý bố cục của một bản báo cáo:

- ✓ Lý do chọn đề tài
- ✓ Dữ liệu
- ✓ Mô tả bài toán
- ✓ Kỹ thuật tiền xử lý dữ liệu được lựa chọn
- ✓ Thuật toán khai thác dữ liệu được lựa chọn
- ✓ Kết quả đạt được
  - Phát biểu kết quả
  - So sánh, đánh giá
- ✓ Chương trình ứng dụng sử dụng kết quả (nếu có)
- ✓ Kết luận
  - Ưu điểm
  - Hạn chế
  - Hướng phát triển
- ✓ Bảng phân công công việc của các thành viên trong nhóm
- ✓ Bảng đánh giá chéo các thành viên trong nhóm (thang điểm 10)
- ✓ Tài liệu tham khảo

VII. Yêu cầu khác

Trước thời điểm báo cáo cuối kỳ, các nhóm chuẩn bị 01 file báo cáo, slide trình chiếu, mã nguồn, dữ liệu, hướng dẫn cài đặt.

Sau khi báo cáo xong, dựa theo sự góp ý của GVHD, nhóm sinh viên tiến hành chỉnh sửa và nộp lại đầy đủ những nội dung sau qua đường dẫn do GV cung cấp:

- ✓ File báo cáo (định dạng PDF và MS Word hoặc LaTeX…)
- ✓ Dữ liệu và mã nguồn phần mềm (nếu có).
- ✓ Hướng dẫn cài đặt môi trường và phần mềm (nếu có).

VIII. Tài liệu tham khảo

1. https://open.data.gov.vn/
2. https://opendata.hcmgis.vn/
3. http://ana.cachopo.org/datasets-for-single-label-text-categorization
4. http://archive.ics.uci.edu/ml/datasets.html
5. http://deeplearning.net/datasets/
6. http://lib.stat.cmu.edu/datasets/
7. http://neuroph.sourceforge.net/sample_projects.html
8. http://qwone.com/~jason/20Newsgroups/
9. http://snap.stanford.edu/data/#reviews
10. http://us1.campaignarchive1.com/?u=e4c8fb8b43860678deab268e5&id=5ce588387d&e=bf605e17d4
11. http://www.databaseanswers.org/data_models/
12. http://www.inf.ed.ac.uk/teaching/courses/dme/2014/datasets.html
13. http://yann.lecun.com/exdb/mnist/
14. https://archive.ics.uci.edu/ml/datasets.html
15. https://cs.joensuu.fi/sipu/datasets/
16. https://data.cityofnewyork.us/
17. https://data.gov.uk/dataset/agricultural_market_reports
18. https://learnersdesk.weebly.com/weka-tutorials.html
19. https://onlinecourses.science.psu.edu/stat857/node/215
20. https://opendata.cityofnewyork.us/
21. https://people.sc.fsu.edu/~jburkardt/datasets/datasets.html
22. https://stackoverflow.com/questions/2674421/free-large-datasets-to-experimentwith-hadoop
23. https://vincentarelbundock.github.io/Rdatasets/datasets.html

24. https://wiki.csc.calpoly.edu/datasets/wiki/Houses
25. https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
26. https://www.google.com/publicdata/directory#
27. https://www.kaggle.com/datasets
28. https://www.quora.com/Data/Where-can-I-find-large-datasets-open-to-the-public
29. https://www.researchgate.net/post/Where_can_I_find_Credit_Card_fraud_detection_data_set
30. http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.page
31. http://vincentarelbundock.github.io/Rdatasets/datasets.html

~ HẾT ~