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

Tài liệu bài giảng:

KHAI THÁC DỮ LIỆU – IS252

Mở đầu:

GIỚI THIỆU MÔN HỌC

ThS. Dương Phi Long – Email: longdp@uit.edu.vn

NỘI DUNG

01 *Giới thiệu chung về môn học*

02 *Nội dung môn học*

03 *Đánh giá*

04 *Một số nguyên tắc và lưu ý*

Giới thiệu chung về môn học

- *Số tín chỉ - số buổi học:*
  - *Lý thuyết: 3 tín chỉ, 45 tiết, 11 buổi học*
  - *Thực hành: 1 tín chỉ, 30 tiết, 6 buổi học*
- *Tìm hiểu các nội dung:*
  - *Các khái niệm về KTDL, quá trình khám phá tri thức*
  - *Các giai đoạn chính của quá trình KTDL*
  - *Một số kỹ thuật KTDL*
  - *Các xu hướng, thách thức trong KTDL*

Mục tiêu môn học

- *Cung cấp những khái niệm, kiến thức cở bản về KTDL, về một số kỹ thuật KTDL.*
- *Trang bị và thực hành để hiểu rõ các kỹ thuật chính trong KTDL thông qua bài tập thực hành (cá nhân, nhóm) và đồ án môn học (nhóm)*

Nội dung môn học

- *Chương 1: Tổng quan về KTDL (buổi 1)*
- *Chương 2: Tiền xử lý dữ liệu (buổi 2)*
- *Chương 3: Tập phổ biến và Luật kết hợp (buổi 3)*
- *Chương 4: Dãy phổ biến (buổi 4)*
- *Chương 5: Tập thô (buổi 4, 5)*
- *Chương 6: Phân lớp dữ liệu (buổi 6, 7)*
- *Chương 7: Gom cụm dữ liệu (buổi 8, 9)*
- *Chương 8: Một số bài toán, xu hướng và thách thức (buổi 10)*
- *Ôn tập (buổi 11)*

Đánh giá

- *Thi lý thuyết cuối kỳ: 50%*
- *Thực hành: 50%*
  - *Đồ án môn học: 35% (4 sinh viên/ nhóm)*
    - *Đăng ký nhóm: buổi 1*
    - *Báo cáo giữa kỳ: buổi 3*
    - *Báo cáo cuối kỳ: buổi 6*
  - *Bài tập quá trình: 15%*
    - *Chuyên cần*
    - *Bài tập trên lớp, về nhà (cá nhân/ nhóm)*
  - *Điểm cộng: Ôn tập và Seminar, tối đa 1đ. Chủ đề: nội dung mở rộng trong 1 số buổi học.*

Đồ án môn học

- *Làm việc theo nhóm, mỗi nhóm 04 sinh viên.*
- *Mỗi nhóm chọn một vấn đề/chủ đề cần giải quyết, bộ dữ liệu sẽ được sử dụng, thuật toán trong ML/DM.*
- *Mỗi đồ án nên được mô tả chính xác*
  - *Vấn đề: mô tả ngắn, đầu vào, đầu ra, kiểu dữ liệu, ứng dụng trong tương lai,…*
  - *Các thuật toán hoặc công cụ, được lên kế hoạch sử dụng*
  - *Bộ dữ liệu được sử dụng*
- *Đăng ký nhóm và đồ án trên Moodle (trước buổi học thứ 3)*

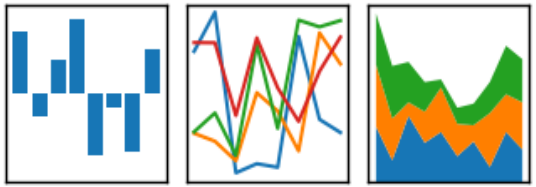
Đồ án môn học

- *Báo cáo vào những buổi học cuối. Yêu cầu các thành viên trong nhóm phải đóng góp và báo cáo đồ án.*
- *Báo cáo đồ án:*
  - *Source code: lưu mã của bạn vào một tệp zip*
  - *Readme.txt: mô tả cách setup, compile, run đồ án*
  - *File word và power point:*
    - *Giới thiệu vấn đề cần giải quyết, bộ dữ liệu đã sử dụng*
    - *Chi tiết về các phương pháp DL/ ML*
    - *Kết quả của các đánh giá khác nhau, kết luận/phát hiện mới, …*
    - *Tổng kết: khó khăn, giải pháp,…*

Đồ án môn học

- *Đánh giá dựa trên:*
  - *Khó khăn, thách thức khi thực hiện*
  - *Sự phù hợp & chất lượng của phương pháp/giải pháp đã chọn*
  - *Tính chặt chẽ của đánh giá thực nghiệm và đánh giá về phương pháp/giải pháp được lựa chọn*
  - *Chất lượng của bài thuyết trình và file word báo cáo*
- *Mỗi đồ án sẽ có 15' thuyết trình & demo*
- *Nếu sử dụng một số thư viện/package/source code: khai báo trong báo cáo trong file báo cáo và trình chiếu*

Công cụ thực hành



Một số nguyên tắc và lưu ý

- *Tham dự lớp học đầy đủ và đúng giờ. Lớp học bắt đầu lúc 7h45 (đối với lớp sáng), 13h15 (đối với lớp chiều).*
- *Theo dõi các thông báo trên Courses, trang Student.*
- *Tích cực tham gia trao đổi, thảo luận, làm bài tập trên lớp. Bài làm giống nhau chia n điểm (n bài giống nhau) hoặc 0 điểm.*
- *Làm bài tập về nhà.*
- *Chủ động đọc tài liệu, tìm hiểu thêm kiến thức.*
- *Báo cáo đồ án đầy đủ theo tiến độ.*

Một số tài liệu tham khảo

- **Tài liệu, sách:**

(1). Đỗ Phúc, Giáo trình Khai thác dữ liệu, NXB. ĐHQG-HCM, 2020

(2). Trần Minh Quang, Khai thác dữ liệu và Kỹ thuật phân lớp, NXB. ĐHQG-HCM, 2020

(3). Vũ Hữu Tiệp, Machine Learning cơ bản, NXB. Khoa học và Kỹ thuật, 2019

(4). Aurélien Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, 2019

(5). Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining Concepts and Techniques, 3rd edition, Morgan Kaufmann Publishers, Elsevier, 2012.

(6). Cao Thị Nhạn, Slide Khai thác dữ liệu, 2022

(7). Khoat Than, Slide Khai thác dữ liệu và Máy học, 2022

(8). Nguyễn Hoàng Tú Anh, Slide Khai phá dữ liệu

- **Data for experiments:**
  - *Kaggle: https://www.kaggle.com/*
  - *UCI repository: http://archive.ics.uci.edu/*

THANKS!

Any questions?