TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
CHƯƠNG TRÌNH CHẤT LƯỢNG CAO

ĐỀ CƯƠNG MÔN HỌC

CTT305: KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

Học kỳ: **II / 2015-2016**

THÔNG TIN GIẢNG VIÊN

Họ và tên: Nguyễn Ngọc Thảo

Văn phòng làm việc: I81

Email: nnthao@fit.hcmus.edu.vn

Số điện thoại:

Thời gian tiếp sinh viên: 14h – 17h chiều thứ 4 hàng tuần

THÔNG TIN MÔN HỌC

Số tín chỉ: 3 tín chỉ

Điều kiện bắt buộc: Không có

Lớp: 13CLC

MỤC TIÊU MÔN HỌC

Để đạt môn học này, sinh viên cần:

- hiểu rõ khái niệm khai thác dữ liệu và vai trò của nó trong đời sống và lĩnh vực học thuật
- vận dụng thành thục các kỹ thuật tiền xử lý dữ liệu để chuẩn bị nguồn dữ liệu tốt cho các tác vụ phân tích ở mức độ cao hơn
- biết cách phân tích dữ liệu để tìm ra thông tin hữu ích tiềm ẩn trong dữ liệu thô thông qua việc áp dụng thuật toán khai thác mẫu phổ biến và luật kết hợp, phân lớp dữ liệu và gom nhóm dữ liệu,
- có khả năng tự triển khai cài đặt một số kỹ thuật khai thác dữ liệu phổ biến
- nhận biết các thách thức trong khai thác dữ liệu và xu hướng giải quyết vấn đề hiện nay

MÔ TẢ MÔN HỌC

Tìm kiếm thông tin có ích từ một khối lượng lớn dữ liệu thô luôn là vấn đề quan trọng thường gặp trong thực tế lẫn nghiên cứu học thuật. Môn học cung cấp cho sinh viên kiến thức cơ bản về khai thác dữ liệu để giải quyết vấn đề này. Nội dung của môn học gồm 4 chủ đề chính. 1) Tiền xử lý dữ liệu: giới thiệu đến sinh viên các kỹ thuật biến đổi trên dữ liệu thô nhằm đạt được nguồn dữ liệu ban đầu có chất lượng tốt đáp ứng yêu cầu của các tác vụ phân tích cấp cao hơn, ví dụ tìm giá trị thích hợp điền vào các ô dữ liệu thiếu, chuẩn hóa miền giá trị của trường dữ liệu,…2) Khai thác mẫu phổ biến và luật kết hợp tìm mọi quan hệ đồng xuất hiện, hay còn gọi là mối kết hợp, giữa các hạng mục dữ liệu. Một ứng dụng kinh điển của bài toán này là phân tích dữ liệu giỏ mua hàng nhằm tìm hiểu thói quen mua sắm của khách hàng, trong một cửa hàng hay siêu thị, thông qua mối liên kết giữa những sản phẩm được mua. 3) Phân lớp dữ liệu và 4) Gom nhóm dữ liệu sử dụng các kỹ thuật học máy để thực hiện phân hoạch dữ liệu theo nhu cầu, chẳng hạn như phân loại văn bản theo chủ đề, gom nhóm khách hàng theo sở thích, v.v. Sinh viên rèn luyện khả năng giải quyết vấn đề thông qua việc cài đặt hoặc sử dụng công cụ hỗ trợ, viết báo cáo về một số bài toán khai thác dữ liệu quy mô nhỏ.

TÀI LIỆU MÔN HỌC

Sách tham khảo

[1]. Jiawei Han, Micheline Kamber, and Jian Pei. 2011. Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[2]. Bing Liu. 2006. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Phần mềm

[1]. WEKA: http://www.cs.waikato.ac.nz/ml/weka/

[2]. Visual Studio C++/C#

Website môn học

[1]. Theo website chương trình/moodle

CHỦ ĐỀ MÔN HỌC

- Chủ đề 1: Tiền xử lý dữ liệu
- Chủ đề 2: Khai thác mẫu phổ biến và luật kết hợp
- Chủ đề 3: Phân lớp dữ liệu
- Chủ đề 4: Gom nhóm dữ liệu

YÊU CẦU MÔN HỌC

Bài tập thực hành	Thực hiện theo yêu cầu của trợ giảng, nội dung bài tập gồm câu hỏi báo cáo và câu hỏi cài đặt, tương ứng với kiến thức được học trong từng chủ đề.
Bài kiểm tra tại lớp	Thực hiện trong các buổi làm bài tập lý thuyết, nội dung tương ứng với kiến thức được học trong từng chủ đề
Kiểm tra giữa kỳ	Hoàn thành bài kiểm tra viết về chủ đề Tiền xử lý dữ liệu và Khai thác mẫu phổ biến và luật kết hợp
Kiểm tra cuối kỳ	Hoàn thành bài kiểm tra viết về chủ đề Phân lớp dữ liệu và Gom nhóm dữ liệu

THANG ĐIỂM

Thành phần môn học	Phần trăm
Bài tập thực hành	40 %
Bài kiểm tra tại lớp	10 %
Kiểm tra giữa kỳ	20 %
Kiểm tra cuối kỳ	30 %

QUI ĐỊNH VỀ ĐẠO ĐỨC VÀ TÍNH TRUNG THỰC

Sinh viên không được sao chép đáp án có sẵn (bài làm của người khác, tài liệu trên Internet,…). Nếu chỉ tham khảo lấy ý tưởng cũng cần phải ghi rõ trong báo cáo. Các trường hợp vi phạm đều bị 0 điểm bài làm.

NHỮNG QUY ĐỊNH KHÁC

Quy định về thông tin, liên lạc qua máy tính

Moodle and e-mail sẽ được sử dụng để trao đổi với sinh viên trong suốt khóa học. Vì vậy, sinh viên nên kiểm tra e-mail mỗi ngày.

Khi gởi e-mail tới giảng viên, tiêu đề email bắt đầu: [CTT305-<Mã lớp>] <*Nội dung*>

LỊCH TRÌNH GIẢNG DẠY

(Gồm: chủ đề môn học, bài tập, các bài đọc liên quan, bài tập nhóm và kiểm tra)

Tuần	Thứ	Ngày	Nội dung	Bài đọc liên quan	Bài tập về nhà/Bài tập nhóm
1	2	11/01/2016	Giới thiệu môn học		
	6	15/01/2016	Bài 1 – Giới thiệu Khai thác dữ liệu	Chap 1 - J. Han	
2	2	18/01/2016	Bài 2 – Tìm hiểu dữ liệu	Chap 2 - J. Han	
	6	22/01/2016	Bài 3 – Tiền xử lý dữ liệu: Làm sạch dữ liệu, tích hợp dữ liệu	Chap 3 - J. Han: 3.2, 3.3	
3	2	25/01/2016	Bài 3 – Tiền xử lý dữ liệu (tt): Giảm chiều dữ liệu, biến đổi dữ liệu và rời rạc hóa dữ liệu	Chap 3 - J. Han: 3.4, 3.5	
	6	29/01/2016	Làm bài tập tiền xử lý dữ liệu		
4	2	22/02/2016	Giới thiệu phần mềm Weka		Bài tập thực hành 1: Tiền xử lý dữ liệu
	6	26/02/2016	Bài 4 – Khai thác mẫu phổ biến và luật kết hợp: Thuật toán Apriori	Chap 6 – J. Han Chap 2 – B. Liu: 2.2	
5	2	29/02/2016	Bài 4 – Khai thác mẫu phổ biến và luật kết hợp (tt): Thuật toán FP-Growth	Chap 6 – J. Han	
	6	04/03/2016	Bài 5 – Khai thác mẫu tuần tự: Thuật toán GSP và PrefixSpan	Chap 2 – B. Liu: 2.6 – 2.9	

6	2	07/03/2016	Làm bài tập khai thác mẫu phổ biến và luật kết hợp		Bài tập thực hành 2: Khai thác mẫu phổ biến
	6	11/03/2016	**Thi giữa kỳ**		
7	2	14/03/2016	Bài 6 – Phân lớp dữ liệu: Quy nạp cây quyết định và Phân lớp Naïve Bayes	Chap 8 – J. Han: 8.2, 8.3 Chap 3 – B. Liu: 3.2, 3.6	
	6	18/03/2016	Bài 6 – Phân lớp dữ liệu (tt): Phân lớp dựa trên luật và K-nearest neighbor	Chap 8, 9 – J. Han: 8.4, 9.5.1 Chap 3 – B. Liu: 3.9	
8	2	21/03/2016	Làm bài tập phân lớp dữ liệu		Bài tập thực hành 3: Phân lớp dữ liệu
	6	25/03/2016	Bài 7 – Gom nhóm dữ liệu: Phương pháp phân hoạch và phân cấp	Chap 10 – J. Han: 10.2, 10.3 Chap 4 – B. Liu: 4.2, 4.2	
9	2	28/03/2016	Bài 7 – Gom nhóm dữ liệu: Phương pháp dựa trên mật độ và dựa trên lưới	Chap 10 – J. Han: 10.4 – 10.6	
	6	01/04/2016	Làm bài tập gom nhóm dữ liệu		Bài tập thực hành 4: Phân lớp dữ liệu
10	2	04/04/2016	Bài 8 – Nhà kho dữ liệu: Khối dữ liệu, OLAP	Chap 4 – J. Han	
	6	08/04/2016	Bài 8 – Nhà kho dữ liệu (tt): Thiết kế, cài đặt và sử dụng	Chap 4 – J. Han	
11	2	11/04/2016	Bài 9 – Các xu hướng trong khai thác dữ liệu		
	6	15/04/2016	Tổng kết và Ôn tập		