**(44) KHAI PHÁ DỮ LIỆU VÀ ỨNG DỤNG**

**1. Tên học phần:** KHAI PHÁ DỮ LIỆU VÀ ỨNG DỤNG

**2. Mã học phần:** INT3012

**3. Số tín chỉ: 3** (2LT+1TH)  **Học phần:** *Bắt buộc*

* Lý thuyết : 30 tiết
* Thực hành (thảo luận, bài tập, kiểm tra): 30 tiết
* Tự học : 75 giờ

**4. Phân bố thời gian**

 Thời điểm thực hiện:

* Học kỳ : 8
* Số tiết/ tuần : 4 tiết/ tuần
* Tổng số tuần/ học kỳ: 15 tuần

**5. Bộ môn/ Khoa phụ trách:** Khoa Công nghệ thông tin

**6. Điều kiện ràng buộc:**

- Học phần tiên quyết : Cấu trúc dữ liệu và giải thuật

- Học phần học trước : Nhập môn trí tuệ nhân tạo

- Học phần song hành :

- Học phần thay thế (nếu có):

**7. Mô tả học phần**

- Học phần Cấu trúc dữ liệu và giải thuật là học phần bắt buộc, được bố trí vào học kỳ 6 của chương trình đào tạo (CTĐT).

- Học phần nằm trong khối kiến thức giáo dục chuyên nghiệp, phần kiến thức cơ sở của nhóm ngành.

- Học phần có 8 chương, cung cấp cho sinh viên các kiến thức và kỹ thuật khai phá dữ liệu để rút trích các tri thức quí báu từ các kho dữ liệu. Mối quan hệ giữa tri thức rút trích và tiến trình ra quyết định, hoạch địch chính sách sẽ được thảo luận với nhiều ứng dụng thực tế. Trong học phần này, sinh viên sẽ tìm hiểu các chủ đề:

- Vai trò của khai phá dữ liệu trong bối cảnh tràn ngập dữ liệu.

- Chuẩn bị dữ liệu như thế nào để đạt kết quả khai phá dữ liệu tốt.

- Các nhiệm vụ của khai phá dữ liệu: dự đoán hay mô tả dữ liệu và kèm theo các ứng dụng thực tế.

- Các vấn đề cần quan tâm, giải quyết trong lĩnh vực khai phá dữ liệu..

**8. Mục tiêu học phần**

Kết thúc học phần này, sinh viên đạt được những yêu cầu cơ bản sau:

*8.1. Về kiến thức*

 *-* Nêu được các kiến thức cơ bản về khai phá dữ liệu, khả năng ứng dụng và nâng cao kỹ năng giải quyết vấn đề.

*8.2. Về kỹ năng*

 - Có kỹ năng vận dụng khai phá dữ liệu, dự đoán, mô tả dữ liệu, các ứng dụng trong thực tế.

*8.3. Về thái độ*

*-* Có ý thức quan tâm tìm hiểu vấn đề khai phá dữ liệu và ứng dụng.

**9. Nhiệm vụ của sinh viên**

- Nhiệm vụ 1: Phải nghiên cứu trước giáo trình, chuẩn bị các ý kiến hỏi, đề xuất khi nghe giảng;

 - Nhiệm vụ 2: Sưu tầm, nghiên cứu các tài liệu có liên quan đến nội dung của từng phần, từng chương, mục hay chuyên đề theo sự hướng dẫn của giảng viên;

 - Tham gia đầy đủ các giờ giảng của giảng viên và các buổi tổ chức thảo luận dưới sự hướng dẫn và điều khiển của giảng viên theo quy chế.

**10. Tài liệu tham khảo**

**10.1. Giáo trình, tài liệu chính**

[1]. Pang-Ning Tan, Michaech Steinbach, Vipin Kumar, Introduction to Data Mining, Addison Wesley, 2006

**10.2. Giáo trình tài liệu tham khảo**

[2]. Đỗ Trung Tuấn, Các hệ trợ giúp quyết định, NXB Đại học Quốc gia, 2015.

**11. Trang thiết bị dạy – học:** Máy tính, máy chiếu.

**12. Tiêu chuẩn đánh giá sinh viên**

Theo Điều 10, Điều 19, Điều 21, Điều 22 của Quy chế đào tạo đại học và cao đẳng chính quy theo hệ thống tín chỉ ban hành kèm theo Quyết định số 43/2007/QĐ- BGDĐT, ngày 15 tháng 8 năm 2007 của Bộ trưởng Bộ Giáo dục và Đào tạo, sinh viên tham dự học mỗi học phần được đánh giá loại đạt nếu:

- Có đăng ký học học phần đúng thời hạn, đảm bảo điều kiện tiên quyết vào đầu mỗi học kỳ với phòng Đào tạo nhà trường.

- Tích cực tham dự lớp học, hoàn thành đầy đủ các điểm đánh giá bộ phận (Ai) và điểm kết thúc học phần (ĐKTHP). Sinh viên vắng mặt không có lý do chính đáng trong buổi kiếm tra đánh giá điểm bộ phận hoặc thi kết thúc học phần sẽ nhận điểm 0.

- Có điểm học phần (ĐHP) đạt một trong các mức điểm A, B, C, D.

**13. Thang điểm**

Thang điểm tính theo Quy chế đào tạo đại học và cao đẳng chính quy theo hệ thống tín chỉ ban hành kèm theo Quyết định số 43/2007/QĐ- BGDĐT, ngày 15 tháng 8 năm 2007 của Bộ trưởng Bộ Giáo dục và Đào tạo và Thông tư số 57/2012/TT-BGDĐT, ngày 27/12/2012 của Bộ trưởng Bộ Giáo dục và Đào tạo sửa đổi, bổ sung một số điều của Quy chế đào tạo đại học và cao đẳng chính quy theo hệ thống tín chỉ ban hành kèm theo Quyết định số 43/2007/QĐ- BGDĐT, ngày 15 tháng 8 năm 2007 của Bộ trưởng Bộ Giáo dục và Đào tạo, cụ thể như sau:

| **Xếp loại** | **Thang điểm 10** | **Thang điểm chữ** | **Thang điểm 4** |
| --- | --- | --- | --- |
| Đạt(Tích lũy) | Giỏi | 9,0 – 10,0 | A+ | 4,0 |
| 8,5 – 8,9 | A | 3,8 |
| Khá | 8,0 – 8,4 | B+ | 3,5 |
| 7,0 – 7,9 | B | 3,0 |
| Trung bình | 6,5 – 6,9 | C+ | 2,5 |
| 5,5 – 6,4 | C | 2,0 |
| Trung bình yếu | 5,0 – 5,4 | D+ | 1,5 |
| 4,0 – 4,9 | D | 1,0 |
| Không đạt | Kém | < 4,0 | F | 0 |

**14. Nội dung học phần**

**Chương 1: TỔNG QUAN VỀ KHAI THÁC DỮ LIỆU**

*Tổng số: 4 tiết, trong đó Lý thuyết: 4 tiết, Thực hành 0 tiết; Tự học: 12 giờ*

**1.1. Tại sao cần Khai thác dữ liệu ?**

**1.2. Khai thác dữ liệu là gì ?**

**1.3. Quá trình khám phá tri thức từ dữ liệu**

**1.4. Các nhiệm vụ và lĩnh vực ứng dụng của Khai thác dữ liệu**

 - Nhiệm vụ dự đoán

 - Nhiệm vụ mô tả

**1.5. Các kỹ thuật của Khai thác dữ liệu**

**1.6. Các thách thức**

**Chương 2: QUY TRÌNH CHUẨN BỊ DỮ LIỆU**

*Tổng số: 12 tiết, trong đó Lý thuyết: 6 tiết, Thực hành 6 tiết; Tự học: 12 giờ*

**2.1. Tại sao cần chuẩn bị dữ liệu ?**

**2.2. Các bước trong qui trình chuẩn bị dữ liệu**

 - Làm sạch dữ liệu

 - Tích hợp và biến đổi dữ liệu

 - Giảm chiều dữ liệu

**2.3. Các kỹ thuật làm làm sạch dữ liệu**

 - Điền dữ liệu thiếu

 - Khử nhiễu dữ liệu

 - Giải quyết dữ liệu mâu thuẫn

**2.4. Các kỹ thuật tích hợp và biến đổi dữ liệu**

 - Tổ hợp dữ liệu

 - Tổng quát hoá dữ liệu

 - Chuẩn hoá dữ liệu

**2.5. Các kỹ thuật rút gọn dữ liệu**

 - Phương pháp chia giỏ (biểu đồ) dữ liệu

 - Phương pháp lấy mẫu

**Chương 3: KHAI THÁC TẬP PHỔ BIẾN VÀ LUẬT KẾT HỢP**

*Tổng số: 12 tiết, trong đó Lý thuyết: 6 tiết, Thực hành 6 tiết; Tự học: 12 giờ*

**3.1. Khái niệm cơ bản**

**3.2. Các phương pháp khai thác tập phổ biến**

 - Thuật toán Apriori

 - Các cải tiến của thuật toán Apriori

 - Thuật toán FP-Growth

**3.3. Luật kết hợp**

 - Luật kết hợp đa chiều

 - Luật kết hợp đa cấp

 - Luật kết hợp trên thuộc tính số

 - Độ đo tính lý thú, quan trọng của luật kết hợp

 - Ứng dụng luật kết hợp : phân tích web log, luật phân lớp

**Chương 4: PHÂN CỤM VÀ PHÂN LỚP DỮ LIỆU**

*Tổng số: 20 tiết, trong đó Lý thuyết: 8 tiết, Thực hành 12 tiết; Tự học: 12 giờ*

**4.1. Khái niệm cơ bản**

**4.2. Các phương pháp gom nhóm**

 - Phương pháp phân hoạch

 - Phương pháp phân cấp

 - Phương pháp dựa trên mật độ

 - Phương pháp dựa trên mô hình

**4.3. Các phương pháp đánh giá phân lớp**

 - Độ đo đánh giá

 - Đánh giá mô hình

**4.4. Các phương pháp phân lớp dữ liệu**

 - Phương pháp dựa trên cây quyết định

 - Phương pháp dựa trên luật

 - Phương pháp Naïve Bayes

 - Phương pháp máy hỗ trợ vectơ (SVM)

 - Phương pháp K- láng giểng gần nhất (k-NN)

**Chương 5: KHAI THỨC DỮ LIỆU PHỨC TẠP**

*Tổng số: 12 tiết, trong đó Lý thuyết: 6 tiết, Thực hành 6 tiết; Tự học: 12 giờ*

**5.1. Khai thác dữ liệu vản bản**

 - Phân tích dữ liệu văn bản và truy vấn thông tin

 - Phân tích mối liên kết dựa trên từ khoá

 - Phân loại tài liệu

 - Phân tích cụm văn bản

**5.2. Khai thác dữ liệu Web**

 - Khai thác cấu trúc liên kết web

 - Phân loại tự động tài liệu trên web

 - Khai thác mạng xã hội

 - Khai thác hành vi sử dụng web

**15. Phương pháp đánh giá học phần**

Quy định số lần kiểm tra bài tập hoặc tiểu luận, thi, số bài thực hành, trọng số của mỗi lần đánh giá:

|  |  |  |  |
| --- | --- | --- | --- |
| TP | Chuyên cần | Định kỳ | Thi |
| Trọng số 10% | 30% | 60% |
| TC | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| SL | HS | SL | HS | SL | HS | SL | HS | SL | HS | SL | HS | SL | HS |
|  | 1 | 1 |  |  |  |  |  |  |  |  | 2 | 1 | 1 | 1 |

(1) Điểm chuyên cần (vắng học 2% tổng số tiết trừ 1 điểm, tính theo thang điểm 10)

(2) Điểm kiểm tra thường xuyên trong quá trình học tập;

(3) Điểm đánh giá nhận thức và thái độ tham gia thảo luận;

(4) Điểm đánh giá thực hiện bài tập, thực hành;

(5) Điểm thi giữa kỳ;

(6) Điểm đánh giá định kỳ;

(7) Thi kết thúc học phần hoặc Điểm tiểu luận.

 Điểm thi kết thúc học phần có trọng số 60%. Hình thức thi: Thi viết

**16. Phương pháp dạy và học:** Giảng dạy lý thuyết kết hợp với thực hành.

|  |  |  |  |
| --- | --- | --- | --- |
| **Ban Giám hiệu** | **Trưởng khoa** | **Trưởng bộ môn** | **Người soạn** |