TRƯỜNG ĐẠI HỌC VĂN LANG
ĐƠN VỊ: KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ THI/ĐỀ BÀI, RUBRIC VÀ THANG ĐIỂM
THI KẾT THÚC HỌC PHẦN
Học kỳ 2, năm học 2023-2024

I. Thông tin chung

Tên học phần:	Nhập môn phân tích dữ liệu lớn			
Mã học phần:	72ITDS40303	Số tin chỉ:	3	
Mã nhóm lớp học phần:	232_72ITDS40303_01			
Hình thức thi: **Đồ án**		Thời gian làm bài:	**21**	Ngày
☐ Cá nhân		☒ Nhóm		
Quy cách đặt tên file		***Mã SV_Ho va ten SV_MaNhom***		

1. Format đề thi

- Font: Times New Roman
- Size: 13
- Quy ước đặt tên file đề thi/đề bài: 72ITDS40303_ Nhập môn phân tích dữ liệu lớn _72ITDS40303_DOAN_De1

2. Giao nhận đề thi

Sau khi kiểm duyệt đề thi, đáp án/rubric. **Trưởng Khoa/Bộ môn** gửi đề thi, đáp án/rubric về Trung tâm Khảo thí qua email: khaothivanlang@gmail.com bao gồm file word và file pdf (***nén lại và đặt mật khẩu file nén***) và nhắn tin + họ tên người gửi qua số điện thoại **0918.01.03.09** (Phan Nhất Linh).

II. Các yêu cầu của đề thi nhằm đáp ứng CLO

(Phần này phải phối hợp với thông tin từ đề cương chi tiết của học phần)

Ký hiệu CLO	Nội dung CLO	Hình thức đánh giá	Trọng số CLO trong thành phần đánh giá (%)	Câu hỏi thi số	Điểm số tối đa	Lấy dữ liệu đo lường mức đạt PLO/PI
(1)	(2)	(3)	(4)	(5)	(6)	(7)
CLO1	Cài đặt pyspark cho các ứng dụng bigdata.	Đồ án	20%	- Mô tả đồ án	10	
CLO2	Áp dụng các nguyên tắc và phương pháp phân tích, tính toán và trực quan hóa dữ liệu liệu lớn.	Đồ án	20%	- Cơ sơ lý thuyết	10	
CLO3	Thực hiện trách nhiệm cá nhân vào việc thực hiện đồ án nhóm và giải quyết các vấn đề liên quan đến nhóm.	Đồ án	40%	- Personal - Nộp đúng hạn - Khả năng trình bày - Trả lời câu hỏi - Ontime - Định dạng báo cáo	10	
CLO4	Sử dụng thành thạo pyspark trong phân tích đánh giá, trực quan bài toán dữ liệu lớn cụ thể.	Đồ án	20%	Kết quả thực nghiệm	10	

Chú thích các cột:

(1) Chỉ liệt kê các CLO được đánh giá bởi đề thi kết thúc học phần (tương ứng như đã mô tả trong đề cương chi tiết học phần). Lưu ý không đưa vào bảng này các CLO không dùng bài thi kết thúc học phần để đánh giá (có một số CLO được bố trí đánh giá bằng bài kiểm tra giữa kỳ, đánh giá qua dự án, đồ án trong quá trình học hay các hình thức đánh giá quá trình khác chứ không bố trí đánh giá bằng bài thi kết thúc học phần). Trường hợp một số CLO vừa được bố trí đánh giá quá trình hay giữa kỳ vừa được bố trí đánh giá kết thúc học phần thì vẫn đưa vào cột (1)

(2) Nêu nội dung của CLO tương ứng.

(3) Hình thức kiểm tra đánh giá có thể là: trắc nghiệm, tự luận, dự án, đồ án, vấn đáp, thực hành trên máy tính, thực hành phòng thí nghiệm, báo cáo, thuyết trình,..., phù hợp với nội dung của CLO và mô tả trong đề cương chi tiết học phần.

(4) Trọng số mức độ quan trọng của từng CLO trong đề thi kết thúc học phần do giảng viên ra đề thi quy định (mang tính tương đối) trên cơ sở mức độ quan trọng của từng CLO. Đây là cơ sở để phân phối tỷ lệ % số điểm tối đa cho các câu hỏi thi dùng để đánh giá các CLO tương ứng, bảo đảm CLO quan trọng hơn thì được đánh giá với điểm số tối đa lớn hơn. Cột (4) dùng để hỗ trợ cho cột (6).

(5) Liệt kê các câu hỏi thi số (câu hỏi số ... hoặc từ câu hỏi số... đến câu hỏi số...) dùng để kiểm tra người học đạt các CLO tương ứng.

(6) Ghi điểm số tối đa cho mỗi câu hỏi hoặc phần thi.

(7) Trong trường hợp đây là học phần cốt lõi - sử dụng kết quả đánh giá CLO của hàng tương ứng trong bảng để đo lường đánh giá mức độ người học đạt được PLO/PI - cần liệt kê ký hiệu PLO/PI có liên quan vào hàng tương ứng. Trong đề cương chi tiết học phần cũng cần mô tả rõ CLO tương ứng của học phần này sẽ được sử dụng làm dữ liệu để đo lường đánh giá các PLO/PI. Trường hợp học phần không có CLO nào phục vụ việc đo lường đánh giá mức đạt PLO/PI thì để trống cột này.

III. Nội dung đề bài

1. Đề bài

Sinh viên chọn 1 trong các nội dung sau

1. **Xử lý và Phân tích Dữ liệu Lớn:**
    - Sử dụng PySpark để xử lý và phân tích dữ liệu lớn từ các nguồn khác nhau như logs, tệp CSV, hoặc cơ sở dữ liệu.
    - Thực hiện các phép biến đổi và tính toán phức tạp sử dụng các chức năng PySpark.
2. **Dự đoán và Phân loại:**
    - Xây dựng mô hình dự đoán hoặc phân loại sử dụng các thuật toán học máy trên dữ liệu lớn bằng PySpark MLlib.
    - So sánh hiệu suất giữa các thuật toán khác nhau và tối ưu hóa mô hình.
3. **Xây dựng Hệ Thống Gợi Ý:**
    - Sử dụng PySpark để xây dựng hệ thống gợi ý dựa trên lịch sử hành vi người dùng.
    - Sử dụng ALS (Alternating Least Squares) để xây dựng mô hình gợi ý.
4. **Phân tích Dữ liệu Đồng thời (Streaming):**
    - Xử lý và phân tích dữ liệu đồng thời sử dụng PySpark Streaming.
    - Xây dựng ứng dụng theo dõi và phản hồi liên tục từ dữ liệu đến hệ thống.
5. **Tối ưu hóa Hiệu suất:**
    - Nghiên cứu và triển khai các kỹ thuật tối ưu hóa hiệu suất cho các công việc PySpark.
    - Sử dụng công cụ như Tương Tác PySpark để theo dõi và điều chỉnh hiệu suất.
6. **Hệ Thống Được Phân tán:**
    - Xây dựng hệ thống được phân tán sử dụng PySpark trên môi trường đám mây hoặc các nút cụm.
7. **Phân tích Đồ thị Xã hội:**
    - Sử dụng PySpark để phân tích đồ thị xã hội từ dữ liệu mạng xã hội hoặc tương tác trực tuyến.

8. **Kết hợp PySpark với Hadoop Ecosystem:**
    - Tìm hiểu cách kết hợp PySpark với các công nghệ khác như Hadoop, Hive, HBase để xử lý và lưu trữ dữ liệu lớn.
9. **Xử lý Dữ liệu Văn bản và Ngôn ngữ Tự nhiên:**
    - Sử dụng PySpark để xử lý và phân tích dữ liệu văn bản lớn, và thực hiện phân tích ngôn ngữ tự nhiên.
10. **Tự chọn tương tự.**

2. Hướng dẫn thể thức trình bày đề bài

1. Quy định định dạng trang

- Khổ trang: A4.
- Canh lề trái: 3,5 cm; Canh lề phải, đầu trang và cuối trang 2 cm.
- Font chữ: Time New Roman, cỡ chữ 13.
- Cách dòng: Line Space: 1.2 -1.5.
- *Các đoạn văn cách nhau 1 dấu Enter.*

2. Đánh số trang

- Từ "Mở đầu" đến phần "Tài liệu tham khảo" đánh theo số (1,2,3…), canh giữa ở đầu trang.

3. Đánh số các đề mục

Đánh theo số thứ tự của chương và số thứ tự của đề mục cấp trên:

CHƯƠNG 1……

1.1……..

1.1.1……..

1.1.2 ……

1.2. ……

CHƯƠNG 2……..

2.1……

2.1.1……..

2.1.2 …..

……

4. Đánh số bảng, đồ thị, hình, sơ đồ

Mỗi loại công cụ minh họa (bảng, đồ thị, hình, sơ đồ…) được đặt tên và đánh số thứ tự

trong mỗi chương có sử dụng bảng, đồ thị, hình, sơ đồ … để minh họa. Số đầu là số chương, sau đó là số thứ tự của công cụ minh họa trong chương đó.

Ví dụ:

Bảng 2.6. Qui mô và cơ cấu khách đến Việt Nam phân theo phương tiện

	2000		2002		2005		2007	
	Ngàn lượt	Tỷ trọng (%)	Ngàn lượt	Tỷ trọng (%)	Ngàn lượt	Tỷ trọng (%)	Ngàn lượt	Tỷ trọng (%)
Đường không	1113,1	52,0	1540,3	58,6	2335,2	67,2	3261,9	78,2

TÀI LIỆU THAM KHẢO

Trịnh Lê A và Giang Xuân H (2003), "*Tiếp cận loại hình du lịch thể thao – mạo hiểm",* **Tạp chí Du lịch Việt Nam**, số 5.

3. Rubric và thang điểm

Criteria	Weight (%)	Excellent From 8p – 10p	Good From 6p – less than 8p	Average From 4p – less than 6p	Fail under 4p
Teamwork	10%	The team worked well together to achieve objectives.	A few occurrences of communication breakdown	Some members would work independently, without regard to objectives or priorities.	Team did not collaborate or communicate.
Content and creativity	50%	The deliverable demonstrated knowledge of the course content with evidence of extensive research	The deliverable demonstrated knowledge of the course content with evidence of limited research	The deliverable demonstrated knowledge of the course content without evidence.	The deliverable did not demonstrate knowledge of the course content
Coherence and Organization	30%	The essay were clearly stated and examples were appropriate. Slides were error-free and logically presented.	The essay were clearly stated; but not all examples were appropriate. Slides were error-free and logically presented.	The essay were not clearly stated. The conclusion was unclear. Slides contained error and a lack of logical progression.	No submission
Speaking skills and participant	20%	Assign work for the team and have good coordination	There are assignments but the coordination is not good	There are assignments but not reasonable	No assignments before practice
	100%				

TP. Hồ Chí Minh, ngày 09 tháng 04 năm 2024

Người duyệt đề

(Đã duyệt)

TS. Bùi Minh Phụng

Giảng viên ra đề

Nguyễn Tất Bảo Thiện