Trạng thái	Đã xong
Bắt đầu vào lúc	Thứ Ba, 9 tháng 12 2025, 8:36 AM
Kết thúc lúc	Thứ Ba, 9 tháng 12 2025, 8:57 AM
Thời gian thực hiện	21 phút 12 giây
Điểm	**26,00** trên 30,00 (**86,67**%)

Câu hỏi **1**

Đúng

Đạt điểm 1,00 trên 1,00

Với hai tập C1 và C2, xác suất để giá trị MinHash trùng nhau bằng:

Select one:

a. Jaccard(C1, C2) ✓ **Giải thích:** Xác suất minhash(C1) = minhash(C2) chính là độ tương đồng Jaccard giữa hai tập.

]]>

b. Cosine(C1, C2)

c. Khoảng cách Euclid(C1, C2)

d. 1 nếu |C1| = |C2|

Câu hỏi **2**

Đúng

Đạt điểm 1,00 trên 1,00

Phát hiện cộng đồng (community detection) trong đồ thị mạng xã hội:

a. Cộng đồng được phát hiện có thể có độ dẫn điện tối thiểu cục bộ tương ứng với các cụm tốt.

b. Tất cả những gì đã đề cập. ✓

c. PageRank được cá nhân hóa được sử dụng để tính toán Conductance cho một cộng đồng.

d. K-means là một giải pháp tốt.

Câu hỏi **3**

Đúng

Đạt điểm 1,00 trên 1,00

Giả sử dữ liệu từ Kafka đến Spark không theo thứ tự thời gian, bạn nên dùng cơ chế nào của Spark để đảm bảo tính đúng đắn khi tính toán theo cửa sổ thời gian (window)?

a. Partition Rebalance
b. Watermarking ✓
c. Replication
d. Trigger Once

Câu hỏi **4**

Đúng

Đạt điểm 1,00 trên 1,00

Ý tưởng cốt lõi đằng sau thuật toán PageRank là gì?

a. Đếm số lượng từ khóa xuất hiện trên trang
b. Dựa vào lưu lượng truy cập (traffic) hàng ngày của website
c. Dựa vào tốc độ tải trang của website
d. Xem các liên kết (hyperlinks) như những phiếu bầu độ tin cậy ✓

Câu hỏi **5**

Đúng

Đạt điểm 1,00 trên 1,00

Chọn câu trả lời đúng về PageRank:

a. Thuật toán PageRank chỉ dừng trong hai trường hợp: (1) các giá trị quan trọng giống nhau hoặc (2) tất cả bằng không.
b. Không có gì được đề cập.
c. Dịch chuyển tức thời (Teleportation) là một kỹ thuật để giải quyết các vấn đề của Dead End và Spider Trap. ✓
d. PageRank là lựa chọn tốt nhất để phát hiện cộng đồng trong biểu đồ.

Câu hỏi **6**

Đúng

Đạt điểm 1,00 trên 1,00

Chuỗi bước điển hình để phát hiện gần trùng lặp văn bản bằng LSH là gì?

Select one:

a. Tokenization → TF–IDF → K-means

b. PCA → k-d tree → Tính Jaccard chính xác

c. Loại bỏ stopword → Word2Vec → HNSW

d. Shingling → MinHash → LSH (banding) ✓ **Giải thích:** Quy trình chuẩn: (1) Shingling để biến văn bản thành tập, (2) MinHash nén tập thành chữ ký bảo toàn tương đồng Jaccard, (3) LSH (banding) sinh cặp ứng viên hiệu quả.

]]>

Câu hỏi **7**

Đúng

Đạt điểm 1,00 trên 1,00

Điều gì xảy ra nếu một trang có 3 liên kết ra ngoài (outbound links) thay vì 1?

a. Mỗi liên kết sẽ truyền đi 100% giá trị PageRank của trang gốc

b. Chỉ liên kết đầu tiên nhận được giá trị, 2 liên kết sau thì không

c. Giá trị PageRank truyền đi sẽ được chia đều cho 3 liên kết đó ✓

d. Trang gốc sẽ bị phạt vì có quá nhiều liên kết

Câu hỏi **8**

Sai

Đạt điểm 0,00 trên 1,00

Chọn câu trả lời đúng về K-means:

a. K-means là một thuật toán có chi phí tính toán nhỏ cho kích thước bài toán lớn.

b. K-means có thể là một giải pháp trong việc phân cụm mạng xã hội. ✗

c. Không có gì được đề cập.

d. Không thể sử dụng phép đo khoảng cách Jaccard trong K-means.

Câu hỏi **9**

Đúng

Đạt điểm 1,00 trên 1,00

Thuật toán nào sau đây KHÔNG PHẢI là một kỹ thuật giảm chiều dữ liệu dựa trên phép chiếu tuyến tính?

a. SVD
b. K-means ✓
c. PCA
d. CUR

Câu hỏi **10**

Đúng

Đạt điểm 1,00 trên 1,00

Chọn câu trả lời chính xác về trực quan hóa (visualization) dữ liệu lớn:

a. Thách thức là sự lộn xộn về thị giác (visual clutter), các vấn đề về hiệu suất, nhận thức hạn chế. ✓
b. Khai phá dữ liệu (Data mining), Mã hóa (Encoding) & Bố cục (Layout) và Kết xuất (Rendering) là vô dụng.
c. Cả A, B và C đều đúng.
d. Không có kỹ thuật trực quan nào hữu ích cho việc truyền dữ liệu.

Câu hỏi **11**

Đúng

Đạt điểm 1,00 trên 1,00

Mô hình "Random Surfer" (người lướt web ngẫu nhiên) giả định điều gì?

a. Người dùng hoặc nhấp vào liên kết trên trang hiện tại, hoặc nhảy ngẫu nhiên đến một trang bất kỳ khác ✓
b. Người dùng luôn nhấp vào liên kết đầu tiên họ thấy
c. Người dùng không bao giờ quay lại trang trước
d. Người dùng chỉ truy cập các trang web có tên miền .com

Câu hỏi **12**

Đúng

Đạt điểm 1,00 trên 1,00

Trong thuật toán K-means, tham số "K" đại diện cho điều gì?

a. Số chiều của không gian dữ liệu sau khi giảm chiều
b. Khoảng cách tối đa cho phép giữa hai điểm dữ liệu
c. Số lần lặp tối đa của thuật toán
d. Số lượng cụm (clusters) mà người dùng cần xác định trước ✓

Câu hỏi **13**

Đúng

Đạt điểm 1,00 trên 1,00

Thuật toán Bloom filtering:

a. Cả A, B và C đều sai (Bloom Filter là cấu trúc dữ liệu xác suất để kiểm tra thành viên, dễ hiện thực phần cứng, độ chính xác phụ thuộc kích thước bit array và số hàm băm). ✓

b. Hiện thực thuật toán Bloom filtering vào phần cứng rất khó.

c. Nó là một giải thuật sampling nhanh.

d. Độ chính xác không tăng tuyến tính theo số hàm băm được dùng.

Câu hỏi **14**

Đúng

Đạt điểm 1,00 trên 1,00

Ưu điểm chính của phân rã CUR so với SVD và PCA là gì?

a. Không cần tính toán ma trận nghịch đảo

b. Tạo ra các đặc trưng trực giao hoàn toàn

c. Tính dễ giải thích (Interpretability) do sử dụng trực tiếp các cột và hàng của ma trận gốc ✓

d. Luôn có sai số tái tạo thấp hơn SVD

Câu hỏi **15**

Sai

Đạt điểm 0,00 trên 1,00

Trong hệ thống gợi ý với các mục (items) và người dùng (users):

a. Đề xuất dựa trên nội dung tốt hơn lọc cộng tác trong trường hợp người dùng mới. ✗

b. Lọc cộng tác là đề xuất dựa trên nội dung tốt hơn trong trường hợp các mục mới.

c. Không có gì được đề cập.

d. Jaccard hoặc Cosine có thể được sử dụng trong lọc cộng tác để đo lường sự tương đồng giữa những người dùng.

Câu hỏi **16**

Đúng

Đạt điểm 1,00 trên 1,00

Phát biểu nào sau đây là ĐÚNG về việc khởi tạo các tâm cụm (centroids) trong K-means?

a. Việc chọn vị trí khởi tạo ngẫu nhiên có thể dẫn đến các kết quả hội tụ khác nhau ✓

b. Vị trí khởi tạo không ảnh hưởng đến tốc độ hội tụ của thuật toán

c. Các tâm cụm bắt buộc phải được chọn từ các điểm dữ liệu có sẵn

d. Thuật toán luôn hội tụ về cùng một kết quả bất kể khởi tạo thế nào

Câu hỏi **17**

Đúng

Đạt điểm 1,00 trên 1,00

Trong LSH cho khoảng cách Euclid với phép chiếu lên trục ngẫu nhiên và lượng tử hóa theo bề rộng 'bucket' a, trực giác đúng là:

Select one:

a. a không ảnh hưởng đến xác suất đụng độ

b. Hai điểm rất xa hầu như luôn rơi vào cùng bucket

c. Nếu hai điểm cách nhau d ≪ a, chúng có xác suất cao rơi vào cùng bucket ✓ **Giải thích:** Khi bề rộng bucket a đủ lớn so với khoảng cách d, hai điểm gần nhau dễ lượng tử hóa vào cùng bucket; điểm xa thì ít trùng bucket hơn.

d. Chỉ phép OR dùng được; phép AND là không hợp lệ

Câu hỏi **18**

Sai

Đạt điểm 0,00 trên 1,00

Những đặc điểm nào KHÔNG phải trong data streaming:

a. Truy vấn liên tục nhưng không truy vấn chuyên biệt (ad-hoc queries).

b. Khái niệm trên đường chuyền và trôi dạt.

c. Sức mạnh tính toán hạn chế. ✗

d. Bộ nhớ hạn chế.

Câu hỏi **19**

Đúng

Đạt điểm 1,00 trên 1,00

Thách thức của việc data streaming trong dữ liệu lớn là:

a. Tất cả những gì đã đề cập.

b. Không thể sử dụng cấu trúc dữ liệu xác suất.

c. Không có thuật toán nào có thể giải quyết vấn đề này.

d. Dữ liệu đến liên tục, bộ nhớ kích thước cố định và thời gian xử lý hạn chế. ✓

Câu hỏi **20**

Đúng

Đạt điểm 1,00 trên 1,00

Mục tiêu chính của hàm mục tiêu trong thuật toán K-means là gì?

a. Tối thiểu hóa số lượng cụm được tạo ra
b. Tối thiểu hóa tổng bình phương khoảng cách giữa các điểm dữ liệu và tâm cụm của chúng ✓
c. Tối đa hóa khoảng cách giữa các tâm cụm
d. Tối đa hóa độ tương đồng giữa các điểm trong các cụm khác nhau

Câu hỏi **21**

Đúng

Đạt điểm 1,00 trên 1,00

Thiết kế pipeline đơn giản cho ứng dụng giám sát giao dịch ngân hàng theo thời gian thực:

a. Dữ liệu -> Hadoop -> Spark Batch
b. Dữ liệu -> Kafka -> Spark Streaming -> Dashboard ✓
c. Dữ liệu -> MySQL -> Spark
d. Dữ liệu -> Excel -> PowerBI

Câu hỏi **22**

Đúng

Đạt điểm 1,00 trên 1,00

Chọn câu trả lời đúng về Machine Learning (ML) trong dữ liệu lớn:

a. Cả A, B và C đều sai. ✓
b. Hiện tại không có phần cứng nào hỗ trợ ML được sử dụng trong các bài toán dữ liệu lớn.
c. ML quá phức tạp để áp dụng dữ liệu lớn.
d. ML không hiệu quả vì quá nhiều dữ liệu đầu vào.

Câu hỏi **23**

Đúng

Đạt điểm 1,00 trên 1,00

Trong việc tìm kiếm các mục tài liệu tương tự:

a. Băm nhạy cảm cục bộ (LSH) có thể đề xuất các mục ít nhất là ngưỡng tương tự do người dùng xác định.
b. Băm tối thiểu (Min-hashing) được sử dụng như một giải pháp giảm kích thước.
c. Một giải pháp kết hợp Min-hashing và LSH giúp giảm chi phí tính toán.
d. Tất cả những gì đã đề cập. ✓

Câu hỏi **24**

Đúng

Đạt điểm 1,00 trên 1,00

Chọn câu trả lời đúng về Hồ dữ liệu (Data lake) và Kho dữ liệu (Data warehouse):

a. Kho dữ liệu hỗ trợ dữ liệu có cấu trúc, còn Hồ dữ liệu hỗ trợ dữ liệu có cấu trúc và bán cấu trúc (và phi cấu trúc). ✓
b. Cả A, B và C đều sai.
c. Kho dữ liệu và Hồ dữ liệu chỉ hỗ trợ dữ liệu có cấu trúc.
d. Hồ dữ liệu là kho dữ liệu.

Câu hỏi **25**

Đúng

Đạt điểm 1,00 trên 1,00

So sánh SVD (Singular Value Decomposition) và CUR trong giảm chiều dữ liệu:

a. CUR có thời gian chạy ngắn và sử dụng ít bộ nhớ hơn SVD (đối với ma trận thưa lớn). ✓
b. Cả A, B và C đều sai.
c. SVD tốt hơn CUR vì hiệu quả và thời gian chạy ngắn.
d. CUR luôn cho kết quả tốt hơn SVD.

Câu hỏi **26**

Đúng

Đạt điểm 1,00 trên 1,00

Thuật toán PageRank được phát triển ban đầu bởi ai?

a. Larry Page và Sergey Brin ✓
b. Mark Zuckerberg và Eduardo Saverin
c. Bill Gates và Paul Allen
d. Steve Jobs và Steve Wozniak

Câu hỏi **27**

Đúng

Đạt điểm 1,00 trên 1,00

Trong pipeline "Kafka -> Spark -> Storage", vai trò của Kafka là gì?

a. Hiển thị kết quả trực quan
b. Cung cấp cơ chế publish-subscribe cho luồng dữ liệu đầu vào ✓
c. Lưu trữ dữ liệu đã xử lý
d. Thực hiện phân tích dữ liệu

Câu hỏi **28**

Đúng

Đạt điểm 1,00 trên 1,00

Nếu trang A có PageRank cao và liên kết đến trang B, điều gì sẽ xảy ra?

a. Trang B sẽ nhận được một phần giá trị PageRank lớn từ trang A ✓
b. PageRank của trang B không bị ảnh hưởng
c. PageRank của trang B sẽ giảm xuống
d. Trang A sẽ mất hết giá trị PageRank của mình

Câu hỏi **29**

Đúng

Đạt điểm 1,00 trên 1,00

Phân rã giá trị kỳ dị (SVD) phân tích một ma trận A kích thước $m \times n$ thành tích của ba ma trận nào?

a. $Q R$
b. $U \Sigma V^T$ ✓
c. $P D P^{-1
d. $L D U$

Câu hỏi **30**

Sai

Đạt điểm 0,00 trên 1,00

Select one:

a. ✗
b.
c.
d.