

Practice Exam Questions

DELLTechnologies



Specialist - Data Scientist Advanced Analytics



EXAMKILLER

Help Pass Your Exam At First Try

Total Question: 66 QAs

Question No: 1

Given an input vector of features, a Random Forests model performs a classification task and ends in a tie. How does the model handle this outcome?

- A. The model will be rebuilt
- B. A winner is chosen at random
- C. The tree that caused the tie is discarded
- D. One more tree is added to the forest

Answer: B

Question No: 2

Which HDFS feature protects against user errors causing accidental loss of data?

- A. Encryption
- B. Replication
- C. Name node federation
- D. Snapshots

Answer: B

Question No: 3

What process must address acoustic ambiguity in NLP?

- A. Part-of-speech tagging
- B. Word sense disambiguation
- C. Speech recognition
- D. Discourse

Answer: C

Question No: 4

A data engineer is asked to process several large datasets using MapReduce. Upon initial inspection the engineer realizes that there are complex interdependencies between the datasets.

Why is this a problem?

- A. MapReduce works best on unstructured data
- B. There is no problem; MapReduce accommodates all the data
- C. MapReduce can only parse one file at a time.
- D. MapReduce is not ideal when the processing of one dataset depends on another.

Answer: D

Question No: 5

What is a characteristic of stop words?

- A. Used in term frequency analysis
- B. Include words such as "a", "an", and "the"
- C. Meaningful words requiring a parser to stop and examine them
- D. Don't occur often in text

Answer: B

Question No: 6

Which is NOT a tenet of the Apache Pig Philosophy?

- A. It must be easily commanded
- B. Any type of data can be processed
- C. Hadoop is required
- D. Data should be processed quickly

Answer: D

Question No: 7

What is a property of a good color model for ordinal data?

- A. Uses a rainbow-like color map for distinction of categories
- B. Uses a rainbow-like color map for ease of display and printing
- C. Uses perceptually ordinal colors with just-noticeable increments
- D. Uses perceptually ordinal colors with linear, perceptual increments

Answer: D

Question No: 8

What is the most likely reason for an HBase table to contain millions of columns?

- A. Data is imported from a relational database table
- B. Data is stored in the column qualifier
- C. There are thousands of columns families
- D. The column names are randomly generated

Answer: B

Question No: 9

Which scenario would be ideal for processing Hadoop data with Hive?

- A. Structured data, real-time processing
- B. Unstructured data; batch processing
- C. Unstructured data; real-time processing
- D. Structured data; batch processing

Answer: B

Question No: 10

The naive Bayer classifier is trained over 1600 movie reviews and then tested over 400 reviews.

Here is the resulting confusion matrix:

190 (TP) 10(FN)

80 (FP) 120(TN)

What are the precision, recall, and the F1-score values?

- A. Precision 0.95; Recall: 0.704; F1-score: 0.809
- B. Precision 0.613, Recall: 0.95, F1-score: 0.745
- C. Precision 0.704, Recall: 0.95; F1-score: 0.809
- D. Precision 0.95; Recall: 0.613; F1-score: 0.745

Answer: C

Question No: 11

Why would a company decide to use HBase to replace an existing relational database?

- A. It is required for performing ad-hoc queries.
- B. Varying formats of input data requires columns to be added in real time.
- C. The company's employees are already fluent in SQL.
- D. Existing SQL code will run unchanged on HBase.

Answer: A

Question No: 12

Which metric would be most helpful in identifying a node that may cause network disruption if the node were removed?

- A. Degree
- B. Closeness
- C. Betweenness
- D. PageRank

Answer: A

Question No: 13

Which graph structure would best model the relationship between job seekers and employers?

- A. Bipartite
- B. Weighted
- C. Directed acyclic
- D. Ranked

Answer: A

Question No: 14

What is an ideal use case for HDFS?

- A. Storing files that are updated frequently
- B. Storing files that are written once and read many times
- C. Storing results between Map steps and Reduce steps
- D. Storing application files in memory

Answer: B

Question No: 15

A marketing team creates a graph using a square for each data point, where the length of each side is set to the data value. The data values are 10 and 20. What is the lie factor of the graph?

- A. 1
- B. 2
- C. 3
- D. 6

Answer: B

Question No: 16

How does Latent Dirichlet Allocation (LDA) interpret a document?

- A. As a single-predefined topic
- B. As a mixture of pre-defined topics
- C. As having a mixture of sentiments
- D. As having a single pre-defined sentiment

Answer: B

Question No: 17

What is a key beneficial characteristic of the Random Forest algorithm?

- A. Provides an explanatory model
- B. Distinguishes categorical from continuous variables
- C. Support for unstructured data
- D. Resiliency to complex, non-linear variable interactions

Answer: D

Question No: 18

What advantage does replication provide while storing a file in HDFS?

- A. Data protection and scheduling flexibility
- B. Elimination of requirement for a combiner process
- C. Elimination of requirement for Shuffle and Sort process
- D. Memory optimization and minimizing tasks to run

Answer: A

Question No: 19

Which library is NOT part of the Apache Spark distribution?

- A. MLib
- B. NLTK
- C. GraphX
- D. Spark SQL

Answer: B

Question No: 20

In which step in the visualization lifecycle would you determine how the raw data is stored?

- A. Visualization Planning
- B. Data Preparation
- C. Visualization Building
- D. Discovery

Answer: B

Question No: 21

What runs more efficiently because of Apache Tez?

- A. Pig and Hive
- B. Hive and HBase