

빅 데이터 전문가

예상문제

빅데이터전문가 예상문제

1. 다음 중 빅데이터의 역사 과정이 아닌 것은? (3)

- ① PC시대
- ② 인터넷 시대
- ③ 전화 시대
- ④ 모바일 시대

해설) 빅데이터의 역사는 PC 시대, 인터넷 시대, 모바일 시대, 스마트 시대로 분류된다.

2. 다음 중 빅데이터의 3V가 아닌 것은? (4)

- ① 규모(Volume)
- ② 다양성(Variety)
- ③ 속도(Velocity)
- ④ 비용(Cost)

해설) 빅데이터의 3V는 규모(volume), 다양성(variety), 속도(velocity) 이다.

3. 빅데이터를 고객의 구매나 활동 기록을 분석하여 취향을 분석하거나, 개인화된 추천 시스템에 활용하는 분야는? (2)

- ① 정치
- ② 경제
- ③ 문화
- ④ 과학

해설) 빅데이터를 고객의 구매나 활동 기록을 분석하여 취향을 분석하거나, 개인화된 추천 시스템에 활용하는 분야는 경제 분야이다.

4. 언어, 문자 등 계량 가능한 형태이고, 정형이나 비정형의 데이터 형태인 데이터의 존재론적 특징에 따른 구분은? (1)

- ① 정량적 데이터
- ② 정성적 데이터
- ③ 정신적 데이터
- ④ 정수적 데이터

해설) 정량적 데이터는 언어, 문자 등 계량 가능한 형태이고, 정형이나 비정형의 데이터 형태인 데이터의 존재론적 특징에 따른 구분이다.

5. 수집 데이터의 선정에 있어서 고려할 필요가 없는 사항은? (4)

- ① 가능성
- ② 정확성
- ③ 난이도
- ④ 형태

해설) 수집 데이터의 형태는 선정 과정에서 고려할 필요가 없다.

6. 다음 중 빅데이터의 수집 사례가 아닌 것은? (4)

- ① JSON
- ② Flume
- ③ Chukwa
- ④ R

해설) R은 수집 사례가 아닌 분석 도구이자 시각화 도구이다. JSON, Flume, Chukwa는 수집 사례에 포함된다.

7. XML 유사 데이터 정형화 방식이며, 인터넷 상의 데이터 송수신 방식으로, 작은 용량과 빠른 변환 속도를 가지며 프로그래밍 언어 또는 플랫폼 독립적인 자바스크립트 기반 데이터 표현 방식은? (1)

- ① JSON
- ② Java
- ③ C++
- ④ PHP

해설) JSON은 JavaScript Object Notation이며, 플랫폼독립적인 인터넷 상의 데이터 교환 형식으로, 빅데이터 시대에 대중화되었다.

8. 다음 중 하둡 분산 시스템(HDFS)의 구성요소가 아닌 것은? (4)

- ① 네임노드 서버
- ② 보조 네임노드 서버
- ③ 데이터 노드 서버
- ④ 통신 노드 서버

해설) 하둡 분산 시스템(HDFS)의 구성요소는 네임노드 서버, 보조 네임노드 서버, 데이터 노드 서버 등이며, 통신 노드 서버는 포함되지 않거나 굳이 구분하지 않는다.

9. 다음 중 NoSQL의 데이터 모델별 분류가 아닌 것은? (3)

- ① 키밸류
- ② 순차적 키밸류
- ③ 행렬

④ 빅테이블

해설) 행렬은 NoSQL의 데이터 모델별 분류가 아니다. 키밸류, 순차적 키밸류, 빅테이블, 도큐먼트, 그래프 등이 모델별 분류에 속한다.

10. 다음 중 로컬 파일을 HDFS에 저장하기 위하여 사용하는 하둡 명령은? (2)

- ① get
- ② put
- ③ ls
- ④ cp

해설) get은 하둡으로부터 로컬로 파일을 받아오기 위한 명령이며, ls는 파일의 목록을 보기 위한 명령이다. Cp는 하둡 HDFS에서 파일을 복사하기 위한 명령이다.

11. 다음 중 MongoDB의 주요 개념이 아닌 것은? (4)

- ① 데이터베이스
- ② 컬렉션
- ③ 도큐먼트
- ④ 테이블

해설) 테이블은 MongoDB의 주요 개념이 아니다. 데이터베이스, 컬렉션, 도큐먼트, 익스텐트 등이 MongoDB의 주요 개념이다.

12. 다음 중 MongoDB에서 도큐먼트를 추가할 때 사용하는 명령은? (1)

- ① insert
- ② delete
- ③ remove
- ④ add

해설) MongoDB에서 도큐먼트를 추가할 때 insert 명령을 사용한다. 그 외의 명령은 도큐먼트를 추가하는 것과는 관계가 없는 명령이다.

13. 다음 중 R이 지원하지 않는 운영체제 환경은? (4)

- ① Windows
- ② Linux
- ③ Mac OS
- ④ MSX

해설) R은 Windows, Linux, Mac OS 등 최근의 대중적인 운영체제 환경을 모두 지원한다. MSX는 1980년대에 단종된 환경으로 현재는 널리 사용하지 않으며, R에서도 지원하지 않는다.

14. 다음 중 R에서 올바르지 않은 변수는? (4)

- ① test
- ② test2
- ③ test2.
- ④ _test

해설) R에서는 변수명의 첫 글자가 숫자나 언더 바(_)로 시작할 수 없다.

15. 다음 중 R에서 변수의 데이터형을 알아내기 위하여 사용하는 함수는? (1)

- ① class()
- ② view()
- ③ discover()
- ④ know()

해설) R에서 변수의 데이터형을 알아내기 위하여 사용하는 함수는 class() 함수이다.

16. R에서 사용하고 있는 변수 중 글자 v를 포함하는 변수를 출력하기 위하여 올바르게 사용한 명령은? (1)

- ① ls(pattern="v")
- ② ls(include="v")
- ③ print(pattern="v")
- ④ print(include="v")

해설) R에서 변수의 목록을 출력하려면 ls() 명령을 사용하며, 특정 글자를 포함하는지 여부를 확인하려면 인자에 pattern= 을 넣는다.

17. R에서 주석을 표현하기 위하여 명령의 맨 앞에 사용하는 기호는? (1)

- ① #
- ② \$
- ③ &
- ④ *

해설) R에서 주석을 표현하기 위하여 # 기호를 사용하여 주석을 표현할 수 있다.

18. R에서 실수(real number)를 표현하기 위한 데이터형은? (2)

- ① 논리형
- ② 숫자형
- ③ 정수형
- ④ 복소수형

해설) R에서 실수(real number)를 표현하기 위하여 적합한 데이터형은 숫자형(Numeric)이다.

19. 다음 중 R-객체가 아닌 것은? (4)

- ① 벡터(Vectors)
- ② 리스트(Lists)
- ③ 행렬(matrices)
- ④ 테이블(tables)

해설) 테이블은 데이터베이스의 개념이다. 데이터베이스에서의 테이블과 유사한 R-객체로는 데이터 프레임(data frames)이 있다.

20. 다음 대입 연산자 중 성격이 다른 하나는? (4)

- ① <-
- ② <<-
- ③ =
- ④ ->

해설) <-, <<-, = 는 오른쪽의 값을 왼쪽으로 대입하는 연산자임에 반해, ->는 왼쪽의 값을 오른쪽으로 대입하는 연산자이다.

21. R에서 벡터의 각 성분에 따라 반복하는 루프 구조로, 벡터 성분의 위치에 일일이 접근하지 않아 벡터 성분별 명령문을 실행할 때 유리한 루프 구조는? (1)

- ① for
- ② while
- ③ repeat
- ④ foreach

해설) for는 R의 대표적인 루프 구조 중 하나로, 벡터의 성분 위치에 일일이 접근하지 않아, 벡터 성분별 명령문을 실행할 때 유리한 루프 구조를 가진다.

22. R에서 루프 제어 명령문인 것은? (2)

- ① continue
- ② next
- ③ previous
- ④ forever

해설) next는 R에서 루프 제어 명령문의 하나로, 현재 실행중인 루프를 중단하고 맨 처음으로 돌아가 루프를 실행하는 특징을 가지고 있다.

23. 다음 중 R에서 함수를 정의할 때 필수요소가 아닌 것은? (4)

- ① 함수명

② function 키워드

③ 중괄호

④ 출력인자

해설) R에서 출력인자가 없는 함수를 만들 수도 있다. 입력인자, 출력인자, 입력인자명 등은 함수의 필수요소가 아닌 선택요소이다.

24. 다음 중 표본의 원천이 아닌 것은? (4)

① 정부, 공공기관 등의 데이터 수집

② 실험을 통한 데이터 획득

③ 사물, 객체, 현상의 관찰을 통한 데이터 획득

④ 개인정보 침해를 통한 데이터 획득

해설) 개인정보 침해를 통한 데이터 획득은 올바른 데이터 획득 방법이 아니므로, 유효한 표본이 될 수 없다.

25. 평균과 표준편차 기반의 연속 확률 분포의 일종으로, 중심극한정리에 근거하고 있는 확률분포는? (1)

① 정규분포

② 베타분포

③ 알파분포

④ 감마분포

해설) 정규분포는 평균과 표준편차 기반의 연속 확률 분포의 일종으로, 중심극한정리에 근거하고 있는 확률분포이다.

26. 다음 중 추정 방법론 내지는 추정량의 분류에 포함되지 않는 것은? (4)

① MLE

② MAP

③ MMSE

④ MS

해설) MLE(Maximum Likelihood Estimation), MAP(Maximum A Posteriori), MMSE(Minimum Mean Squared Error) 등은 빅데이터의 추정 방법론에서 자주 쓰이는 용어들이다. 기억해두도록 하자.

27. 다음 중 분산분석의 모형이 아닌 것은? (3)

① 고정효과 모형

② 무선효과 모형

③ 유선효과 모형

④ 혼합효과 모형

해설) 분산분석의 모형은 고정효과, 무선효과, 혼합효과 모형 등이 있다.

28. 다음 중 상관분석의 가정이 아닌 것은? (4)

- ① 선형성
- ② 동변량성
- ③ 정규분포성
- ④ 무작위성

해설) 상관분석은 선형성, 동변량성, 정규분포성, 무선폭립표본 등을 가정하고 있다.

29. 다음 중 상관분석의 방법이 아닌 것은? (4)

- ① 피어슨 상관계수
- ② 스피어만 상관계수
- ③ 크론바흐 알파 계수 신뢰도
- ④ 공분산분석

해설) 공분산분석은 상관분석이 아닌 분산분석의 한 방법이다.

30. 다음의 분석 방법론 중 과거의 데이터 뿐만 아니라 추세(momentum)까지 반영한 모델은? (4)

- ① AR
- ② MA
- ③ ARMA
- ④ ARIMA

해설) ARIMA 모델은 AR, MA 모델의 장점과 모멘텀을 고려한 통합 시계열 모델로, 과거의 데이터 뿐만 아니라 추세(momentum)까지 반영한다.

31. 다음 중 데이터 마이닝의 적용 분야가 아닌 것은? (4)

- ① 분류
- ② 군집화
- ③ 예측
- ④ 시각화

해설) 시각화는 빅데이터 방법론의 가장 마지막 단계에 속한다. 데이터 마이닝은 빅데이터 방법론의 분석 단계에 속하며, 분류, 군집화, 연관성, 연속성, 예측 등의 분야가 있다.

32. 다음 중 데이터 마이닝에서 학습 방법론이 아닌 것은? (4)

- ① 지도학습
- ② 반지도학습

③ 자율학습

④ 휴계학습

해설) 지도학습(supervised learning), 반지도학습(semi-supervised learning), 자율학습(unsupervised learning)은 데이터 마이닝의 학습 방법론의 대표적인 세 가지 철학이다. 반드시 기억해두도록 하자.

33. 유전자 정보에 대하여 데이터 마이닝을 적용하여 새로운 질병에 대한 빠른 진단이 가능해지고 신치료제 개발이 가능한 적용 분야는? (2)

① 공공서비스

② 보건/의료

③ 제조/물류

④ GPS

해설) 빅데이터 마이닝을 이용하여 유전자 정보에 대한 해독과 분석이 가능한 한편, 새로운 질병에 대한 대응이 가능해질 전망이다.

34. 데이터 마이닝의 데이터 축소, 분할 단계 중 데이터 집합의 종류에 속하지 않는 것은? (4)

① 학습용 데이터(training)

② 평가용 데이터(test, evaluation)

③ 검증용 데이터(verification)

④ 관찰용 데이터(observation)

해설) 관찰용 데이터는 시각화 단계에서 필요로 하는 데이터이다.

35. 인접한 k개의 이웃 정보를 토대로 데이터가 소속되는 그룹을 결정하는 분류 방법론은? (1)

① k-최근접 이웃

② k-평균 군집화

③ 인공 신경망

④ 지지 벡터 머신

해설) k-최근접 이웃(k-NN)은 인접한 k개의 이웃 정보를 토대로 데이터가 소속되는 그룹을 결정하는 분류 방법론이다.

36. 지도 분류 학습법에서 주로 사용하며, 랜덤 포레스트(random forest) 등의 방법을 이용하여 분류속도 향상을 꾀할 수 있는 결정 트리 학습법은? (1)

① 의사결정나무

② 베이지안 정리

③ 인공 신경망

④ 지지 벡터 머신

해설) 의사결정나무(decision tree)는 지도 분류 학습법에서 주로 사용하며, 랜덤 포레스트(random forest) 등의 방법을 이용하여 분류속도 향상을 꾀할 수 있는 결정 트리 학습법이다.

37. 각 계층에서 군집의 유사성을 쉽게 확인할 수 있는 군집 방법론으로, 흡수 과정(agglomerative)과 분리 과정(divisive)이 있는 방법론은? (1)

- ① 계통도
- ② k-평균 군집화
- ③ k-최근접 이웃
- ④ 베이지안 정리

해설) 계통도(dendrogram)는 각 계층에서 군집의 유사성을 쉽게 확인할 수 있는 군집 방법론으로, 흡수 과정(agglomerative)과 분리 과정(divisive)이 존재한다.

38. 다음 중 인간의 언어 현상을 컴퓨터에서 모사할 수 있도록 연구하고 구현하는 인공지능의 한 분야는? (1)

- ① NLP
- ② ALP
- ③ SLP
- ④ PLP

해설) NLP(Natural Language Processing)은 인간의 언어 현상을 컴퓨터에서 모사할 수 있도록 연구하고 구현하는 인공지능의 한 분야이다.

39. 다음 중 어휘 분석의 작업이 아닌 것은? (4)

- ① 문장 분리
- ② 토큰화
- ③ 형태소 분석
- ④ 문장 조립

해설) 어휘 분석(lexical analysis) 과정에서는 문장 분리, 형태소 분석, 토큰화, 포스 태깅(POS tagging) 등을 행한다.

40. 전처리 된 데이터에 대하여 인공신경망 등의 방법론을 적용하여 자율학습하여 특징을 추출하는 방법은? (2)

- ① 머신 러닝
- ② 딥 러닝
- ③ 멀티미디어 러닝
- ④ 추출 러닝

해설) 딥 러닝(Deep Learning)은 전처리 된 데이터에 대하여 인공신경망 등의 방법론을 적용하여 자율학습하여 특징을 추출하는 방법이다.

41. 사용자 간의 자유로운 의사소통과 정보 공유, 인맥 확대 등을 통하여 사회적 관계를 생성하고 강화하여 주는 온라인 플랫폼은? (3)

- ① 메신저
- ② 포털사이트
- ③ SNS
- ④ 인터넷뉴스

해설) 소셜 네트워킹 서비스(SNS)는 사용자 간의 자유로운 의사소통과 정보 공유, 인맥 확대 등을 통하여 사회적 관계를 생성하고 강화하여 주는 온라인 플랫폼이다.

42. 다음 중 시각화 절차가 아닌 것은? (4)

- ① 정보 조직화 단계
- ② 정보 시각화 단계
- ③ 상호작용 단계
- ④ 정보 분석 단계

해설) 정보 분석은 시각화 이전 단계에서 행해져야 한다. 정보 조직화 단계, 정보 시각화 단계, 상호작용 단계 등을 밟아야 한다.

43. 다음의 시각화 기술 중 그래픽 디자인을 위한 프로그래밍 언어는? (3)

- ① Tag Cloud
- ② Gephi
- ③ Processing
- ④ Tableau

해설) Processing은 그래픽 디자인을 위한 시각화 프로그래밍 언어 환경이다.

44. 다음 중 R에서 2차원 그래프를 표현하기 위하여 사용하는 함수의 이름은? (1)

- ① plot
- ② disp
- ③ print
- ④ representation

해설) R에서 2차원 그래프를 표현하기 위하여 사용하는 함수의 이름은 plot이다.

45. 다음 중 R에서 히스토그램을 표현하기 위하여 사용하는 함수의 이름은? (1)

- ① hist

- ② graph
- ③ express
- ④ function

해설) R에서 히스토그램을 표현하기 위하여 사용하는 함수의 이름은 hist이다.

46. 다음의 D3.js 출력 결과물 중에서, 데이터의 최대/최소 범위 뿐만 아니라 중간값 및 Q1, Q3를 함께 볼 수 있는 그래프는? (1)

- ① 박스 플롯
- ② 동시출현행렬
- ③ 버블 차트
- ④ 보로노이 다이어그램

해설) 데이터의 최대/최소 범위 뿐만 아니라 중간값 및 Q1, Q3를 함께 볼 수 있는 그래프는 박스 플롯이다.

47. 빅데이터에 대한 방법론을 적용하여 프로젝트의 완수를 위해 행하는 일체의 활동을 일컫는 용어는? (1)

- ① 빅데이터 프로젝트
- ② 빅데이터 활동
- ③ 빅데이터 방법론
- ④ 빅데이터 완수

해설) 빅데이터 프로젝트는 빅데이터에 대한 방법론을 적용하여 프로젝트의 완수를 위해 행하는 일체의 활동이다.

48. 빅데이터 프로젝트의 기획 단계에서 행하는 작업으로, 어떤 문제의 해결 목적과 이유에 대하여 정의하고 문제나 기회를 탐색하며 과제를 도출하는 단계는? (1)

- ① 분석 단계
- ② 계획 단계
- ③ 기회 단계
- ④ 도출 단계

해설) 분석 단계는 빅데이터 프로젝트의 기획 단계에서 행하는 작업으로, 어떤 문제의 해결 목적과 이유에 대하여 정의하고 문제나 기회를 탐색하며 과제를 도출하는 단계이다.

49. 다음 중 빅데이터 프로젝트 관리자의 자질이 아닌 것은? (4)

- ① 기업가 능력
- ② 관리 능력
- ③ 기술적 능력
- ④ 공공서비스 경험

해설) 빅데이터 프로젝트 관리자는 기업가 능력, 관리 능력, 지휘 능력, 의사소통 능력, 그리고 기술적 능력을 겸비하여야 한다.

50. 프로젝트 추진 일정 관리 단계에서 활용할 수 있는 가장 효과적인 프로젝트 관리 도구는? (1)

- ① Gantt 차트
- ② 엑셀
- ③ 파워포인트
- ④ R

해설) Gantt 차트를 이용하면 프로젝트 추진 일정 관리를 함에 있어서 일정 단계 및 활동에 따라 소요예상기간을 산정하여 일정 계획을 수립하는 것이 가능하다.

51. 다음 중 유권자의 성향 정보를 분석하여 맞춤형 선거 전략을 수립하고 시행하는 빅데이터의 활용 분야는? (1)

- ① 정치
- ② 경제
- ③ 사회
- ④ 의료

해설) 정치 분야에서는 유권자의 성향 정보를 분석하여 맞춤형 선거 전략을 수립하고 시행하는 빅데이터의 활용하고 있다.

52. 다음 중 빅데이터의 성공조건이 아닌 것은? (3)

- ⑤ 데이터 기반 문화 형성 및 적용
- ⑥ 분석지능의 극대화
- ⑦ 지배구조 개선 및 대기업화
- ⑧ 분석결과와 혁신의 연결 및 지속

해설) 빅데이터의 성공 조건으로는 데이터 기반 문화 형성 및 적용, 분석지능의 극대화, 분석결과와 혁신의 연결 및 지속 등이 있다.