Program: General/Intelligent Systems/Cybersecurity
Level: Third Term: Fall 2023/2024
Course Code: 02-24-01203 Course Title: Data Science Tools &Software
Time Allowed: 60minutes Total points: 20
Professor name: Dr. Mohamed Abd El-Hafeez

جامعة الإسكندرية
ALEXANDRIA UNIVERSITY
كلية الحاسبات وعلوم البيانات

Attempt ALL the following 53 questions

Choose the MOST APPROPRIATE answer for the following statements.

You may choose E (=ALL) if all answers (A, B, C and D) are correct or choose F (=NONE) if none of the answers fits.

Please write your answers on the ANSWER SHEET ONLY

In the designated answer sheet, mark your choice (ⓐ, ⓑ, ⓒ, ⓓ, ⓔ, or ⓕ) in front of the question number.

Be sure that you have filled the appropriate bubbles carefully as in the example below.

Example: if the choice for question 300 is "C" then your answer sheet should look like this:

300. ⓐ ⓑ ▣ ⓓ ⓔ ⓕ

1. Which of the following is not true regarding Data Science?
   a) Concerned only with big data
   b) Heavy focus on machine learning algorithms
   c) Concerned only with small data
   d) Concerned with theories in statistics
2. Which module in Python supports regular expressions?
   a) String b) re c) pyregex d) sklearn
3. What does the function "search" in the regular expressions package do?
   a) matches a pattern at the start of the string
   b) matches a pattern at any position in the string
   c) replace all matched
   d) delete all matched
4. Which of the following HTTP methods never modifies a server's state?
   a) response = requests.put(...) b) response = requests.post(...)
   c) response = requests.delete(...) d) response = requests.get(...)
5. Which module in Python supports parsing HTML and XML documents?
   a) BeautifulSoup b) numpy c) pandas d) sklearn
6. What is the library that corresponds to the alias "ps" in the following code

```
df = ps.DataFrame([(1, 'Kolter', 'Zico')])
```

   a) pandas
   b) panorama
   c) pymatplots
   d) scipy

Answer the following two questions regarding the state after the execution of the following code:-

```
df = DataFrame([(1, 'Kolter', 'Zico'),(2, 'Manek', 'Gaurav'), (3, 'Rice', 'Leslie')],
columns=["Person ID", "Last Name", "First Name"])
df.drop(1, inplace=True, axis=0)
df.drop(2, inplace=True, axis=1)
```

7. how many records (rows) will be in the dataframe df, after executing the above code?
   a) 1 b) 3 c) 2 d) 4
8. how many columns the dataframe df will have, after executing the above code?
   a) 6 b) 2 c) 3 d) 1
9. Which of the following is not an example of unordered data?

a) Employee records b) Documents c) Bank transactions d)Time Series

10. What is the primary purpose of the Request module?
    a) Send HTTP requests to a server and retrieve web page content
    b) Manage database connections for data storage
    c) Execute complex algorithms for data analysis
    d) Control graphical user interface interactions
11. What will be the output of the following Python code?

```
CarName = 'Porche'
WordName = 'World'
print('{0} is the fastest car in the {2}'.format(CarName, WordName))
```

    a) Porche is the fastest car in the World
    b) Porche is the fastest car in the
    c) Porche is the fastest car in the 2
    d) IndexError: tuple index out of range
12. What does the term "ACID" stand for in the context of databases?
    a) All-Comprehensive Isolation and Durability
    b) Atomicity, Consistency, Isolation, Durability
    c) Advanced Configuration for Isolated Databases
    d) Association of Concurrent Information and Data
13. How is the _id field automatically created if not provided in MongoDB?
    a) Integer b) Timestamp c) ObjectId d) AutoID
14. What is MongoDB?
    a) Relational database b) Document-oriented database c) NoSQL database d) Both B and C
15. In MongoDB, what is a document equivalent to in a SQL database?
    a) Table b) Record c) Field d) Column
16. Which method is used to find documents in a MongoDB collection based on a specific condition?
    a) get_one() b) search() c) find_one() d) query_one()
17. The hamming distance between two binary vectors is equivalent to :-
    a) Jaccard Index b) Euclidean Distance c) Squared Euclidean Distance d) cosine similarity
18. Question: What does setting New_max=1 and New_min=0 achieve in data normalization?
    a) Increases data complexity b) Reduces the impact of outliers
    c) Adds noise to the dataset d) Standardizes data within a specific range
19. Which of the following is common technique to replace missing data in a dataset?
    a) Mean b) Median c) Mode d) Random Value
20. What is the cosine similarity between the vectors (1, 0) and (0, 1)?
    a) 1 b) 0 c) 0.5 d) 2
21. What is the primary purpose of converting an image to grayscale in machine learning algorithms?
    a) To increase computational complexity
    b) To introduce color variations
    c) To reduce computational complexity
    d) To improve image resolution
22. In the context of image normalization, what is the benefit of scaling all images to a common range such as [0,1]?
    a) It increases computational complexity
    b) It ensures fairness across all images
    c) It introduces colour variations
    d) It reduces the need for data augmentation
23. What is data augmentation in the context of image processing?
    a) Increasing the size of an image dataset
    b) Making minor alterations to existing data to increase diversity
    c) Reducing the diversity of a dataset
    d) Converting images to grayscale
24. What is the assumed seasonality for a monthly time series?

a) 7 b) 12 c) 30 d) 365

25. Executing print((lambda x, y: x//y)(4, 3)) in Python produces
a) 0 b) 1 c) 4/3 d) 7
26. Executing print(map(lambda x: x**3 , [0,1,2])) in Python produces
a) [0,0,0] b) [0,1,2] c) [0,1,8] d) [0,1,3]
27. Executing print(list(filter(lambda x: x > 2 and x < 8, [-1,0,5,3]))) in Python produces
a) [5,3] b) [3,5] c) [-1,0,5,3] d) 8
28. Executing print(`functools`.reduce(lambda x, y: x+y, [1,2,3,4])) in Python produces
a) 24 b) 10 c) [1,3,6,10] d) 1
29. Which of the following database is not a relational database?
a) SQLite b) MySQL c) Oracle d) MS Access
30. Which of the following library is used for data visualization?
a) TensorFlow b) Scrapy c) Scikit Learn d) Matplotlib
31. Which of the following is not a tool for data processing, machine learning algorithm implementation, and visualization.
a) SAS b) Weka c) RapidMiner d) SAS and WEKA
32. Complex data streams can be analyzed and visualized dynamically using
a) Apache Spark b) Scrapy c) MS Excel d) MS Powerpoint
33. metrics.DistanceMetric.get_metric is a function defined in
a) Pandas b) Scrapy c) Sklearn d) Matplotlib
34. The output of the following code is

```
dist = get_metric('euclidean')
X = [[2, 3]]
Y = [[2, 2]]
dist.pairwise(X,Y)
```

a) 1 b)5 c) 9 d) 0

35. Let x={a,b,d} and y={b,c} then $S_{jaccard}.(x,y)$ is ---------
a) 0.25 b) 1 c) 0.5 d)
36. To normalize the following dataset
10,40,50,10,50, 70,90,30
a) Divide each value by the mean of the data
b) Replace each value x by (x - the mean of the data / the standard deviation of the data)
c) A followed by B
d) Replace each value x by (x - the standard deviation of the data/ the median of the data)
37. The missing value in the following data may be replaced by
10,40,50,10,50,?,10,60,10,30
a) the mean of the remaining data b) a value of 10
c) the median of the remaining data d) a value of 30
38. Which of the following distances has three components (distance due to span, content and position)
a) D("Hello","HALA") b) D({a,b,c,d},{a,d})
c) D(10:30,5:20) d) D(10101,01111)
39. Which of the following is an example of Application of Dimensionality Reduction
a) Microarray data analysis b) Protein classification
c) Face recognition d) Handwritten digit recognition
40. Which of the following is not a Challenge in the analysis of the data:-
a) Few samples b) Mixed data types and unbalanced data
c) Very high dimensionality d) Noise
41. To use PCA in python you should import it using " From ---------- import PCA"
a) sklearn b) sklearn.models c) sklearn.decomposition d) sklearn.metrics
42. MRMR ---------- the relevancy while ---------- the redundancy
a) maximizes /minimizes b) minimizes/ minimizes c) minimizes/ maximizes d) maximizes / maximizes
43. which of the following is true regarding the Wrapper model
a) Relying on a predetermined classification algorithm
b) Using predictive accuracy as goodness measure
c) High accuracy
d) computationally expensive

Regarding the following code, answer the following two questions:-

```
iris = datasets.load_iris()
df = pd.DataFrame(iris['data'], columns = iris['feature_names'])
scalar = StandardScaler()
scaled_data = pd.DataFrame(scalar.fit_transform(df))
pca = PCA(n_components = 2)
pca.fit(scaled_data)
data_pca = pca.transform(scaled_data)
```

44. What is the purpose of using StandardScaler() in the following code
    a) reduce the dimension b) fill missing data c) normalize the data d) remove noise
45. the number of columns of the data_pca
    a) 4 b) 2 c) 3 d) 1

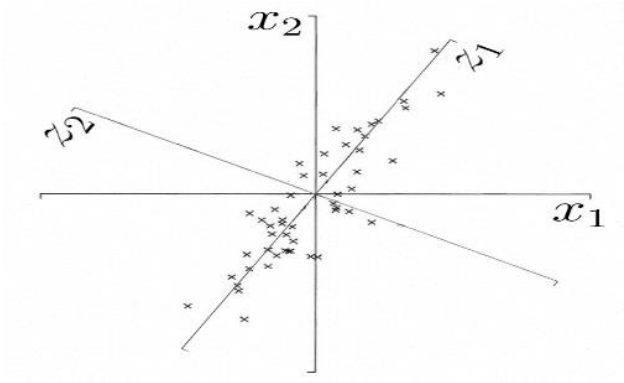
Given the following term frequencies in a corpus D that contains 3 documents D1..D3, answer the following questions:-

Document 1 (D1)	
Term	Term
Caw	2
Sudan	1
Camel	1

Document 2 (D2)	
Term	Term
Sudan	3
Caw	2
Nile	1

Document 3 (D3)	
Term	Term Count
Egypt	2
Nile	2
Caw	1

46. The resulting data matrix will be of size
    a) 3×5 b) 4 × 4 c) 5×5 d) 5×4
47. The normalized term frequency of tf ("camel",D1) is
    a) 0.20 b) 3 c) 4 d) 0.25
48. The inverse document frequency idf("Camel",D)
    a) 3 b) 1 c) 1/3 d) 0
49. what is the tflogidf("caw",D)
    a) 0 b) 1 c) 3 d) 5
50. The resulting distance matrix will be of size
    a) 3×5 b) 4 × 4 c) 5×5 d) 3×3
51. The corresponding feature vector of document D1 using binary term frequency is
    a) [1 1 1 0 0] b) [1 0 0 0 1] c) [1 0 1 1] d) [2 1 1]
52. The correlation between the data using the new axes z1,z2 is ---------- than the correlation between the same data with respect to the axes x1,x2
    a) Higher b) lower c) equals d) higher or equals
53. Which axis you may neglect to reduce the dimension
    a) z1 b) z2 c) z1 or z2 d) z1 and z2



Best Wishes