17 ต.ค. 61

Name_____ ID_____ Seat _____

King Mongkut's University of Technology Thonburi

Department of Computer Engineering

Midterm Examination 1/2018

CPE 325 Big Data
Date: October 2nd, 2018 **Time: 13:00 – 16:00**

Instructions:

1. Carefully read the explanation in each problem and then answer each question.
2. **Do not take the examination sheets out** of the examination room.
3. Write your name, ID, and seat number on every page of examination sheets.
4. Books and notes are **not** allowed to the exam room.
5. University-certified calculator is allowed.
6. Other electronic equipment is **not** allowed.
7. This examination has 2 sections, 7 pages: (Section 1) 5 pages, 20 questions (20 points) and (Section 2) 2 pages, 2 questions (20 points).

Section 1 (20 points)

Instruction: read the following question carefully, then select **one** answer from given choices.

1.1 Why data is important in decision making?
- a. It is generated from the business process.
- b. It is stored in the data warehouse.
- c. Its size is very large and thus need a special technique in processing.
- d. It contains both numerical and categorical variables.

1.2 What is data science?
- a. A method for high performance processing.
- b. A principle to extract knowledge from data.
- c. A tool to create visualization.
- d. All are correct.

1.3 What is the type of analysis in this situation? The company wants to predict whether the customer will buy the product or not.
- a. Descriptive analytics.
- b. Predictive analytics.
- c. Prescriptive analytics.
- d. None of the above.

Name_____ ID_____ Seat _____

1.4 What is the type of analysis in this situation? Managements want to understand the current proportion of customers using different products.
   a. Descriptive analytics.
   b. Predictive analytics.
   c. Prescriptive analytics.
   d. None of the above.

1.5 What is the type of analysis in this situation? The company want to find what products they should recommend to customers to purchase.
   a. Descriptive analytics.
   b. Predictive analytics.
   c. Prescriptive analytics.
   d. None of the above.

1.6 How does big data systems help data analytics?
   a. It provides a means of analyzing data from multiple sources together.
   b. It provides a faster way to process a large file.
   c. It allows visualization software to work interactive with a big data.
   d. All are correct.

1.7 Which program in Anaconda that will give you a developer environment that provides the current variables in the system as well as an interactive Python console?
   a. Jupyter Notebook b. Spyder
   c. Qtconsole d. NetBean

1.8 What is Markdown?
   a. A library for statistical analysis
   b. A way to assign value to a variable
   c. A script language to write and format the output
   d. A way to compare values between variables

1.9 What is the data type of *x*, when x = ["1", 2, True]?
   a. integer
   b. character
   c. logical
   d. list

Name_____ ID_____ Seat _____

1.10 What will be the printed output of the following code?

```
import numpy as np
x = xp.arange(10)
y = x[3:]
y[2]
```

a. 2 b. 3

c. 5 d. 6

1.11 What is the Python data structure that stores the data in a tabular format that has columns represent variables and rows represent data record?

a. list b. tuple

c. dictionary d. data frame

1.12 Which of the following command that transforms an array of values into an array of tuples (index, value) for looping?

a. enumerate b. apply

c. where d. all are correct

1.13 What does the following function declarations mean for the input argument?

```
def calculate(*args)
```

a. It is a pointer. It can be accessed using &.
b. It is an array. The input must be in a list form.
c. It can receive arbitrary number of argument.
d. All are correct.

1.14 If we have a sales record in a tabular format. Columns are variables and rows are sales records. If we want to summarize the total sales per day, which of the following command will give us the result?

a. data.shape
b. pd.merge(data1, data2, on="key")
c. data.describe()
d. data.groupby('day').sum()

1.15 Which of the following command provide the results as data frame when df is data frame?

a. df.shape b. np.array()

c. df.columns d. df.head()

1.16 Which of the following command change the shape of data frame?

a. read_csv b. groupby

c. melt d. arange

Name_____ ID_____ Seat _____

1.17 Which of the following command convert categorical columns into multiple binary columns of levels in categorical columns?

a. transform
b. get_dummies
c. melt
d. spread

1.18 Which of the following steps in data science is NOT part of data preprocessing?

a. enrichment
b. tokenization
c. aggregation
d. modeling

1.19 What is the purpose of LabelEncoder in sklearn.preprocessing?

a. Transform a categorical variable into multiple binary variables. Each variable denotes a level in the categorical variable.
b. Convert a categorical variable into a numerical sequence of labels. The order is alphabetic.
c. Convert a numerical variable into a categorical variable by binning the range of variable. The categorical value is the range in which the actual data is located.
d. All are corrected.

1.20 What is the purpose of the function pd.cut?

a. Transform a categorical variable into multiple binary variables. Each variable denotes a level in the categorical variable.
b. Convert a categorical variable into a numerical sequence of labels. The order is alphabetic.
c. Convert a numerical variable into a categorical variable by binning the range of variable. The categorical value is the range in which the actual data is located.
d. All are corrected.

Name_____ ID_____ Seat _____

Section 1 – Answer Sheet: Please give your answer to all questions in Section 1 here. For each question, select only **one** choice and mark "X" on the choice you chose.

1.	a	b	c	d
2.	a	b	c	d
3.	a	b	c	d
4.	a	b	c	d
5.	a	b	c	d
6.	a	b	c	d
7.	a	b	c	d
8.	a	b	c	d
9.	a	b	c	d
10.	a	b	c	d
11.	a	b	c	d
12.	a	b	c	d
13.	a	b	c	d
14.	a	b	c	d
15.	a	b	c	d
16.	a	b	c	d
17.	a	b	c	d
18.	a	b	c	d
19.	a	b	c	d
20.	a	b	c	d

Name_____ ID_____ Seat _____

Section 2 (20 points)

2.1. *(10 points)* Explain the concepts of Hadoop Distributed File Systems. How does it store and distribute a big data? What are name nodes and data nodes? Why are they efficient in file reading? Why are the file more robust to failure than a single storage?

Name_____ ID_____ Seat _____

2. *(10 points)* We want to write MapReduce programs to detect the offensive words in the social network posts. You are tasked to write this program. You have the data in (doc_id, raw_text) format and a dictionary of a list of offensive word. Design a MapReduce program to do this. Explain in detail how it can be achieved.