

Semester Two 2019

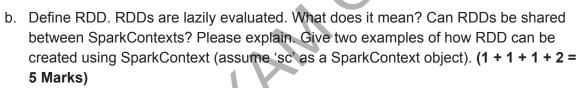
Examination Period

Faculty of Information Technology

5202
ΓA PROCESSING FOR BIG DATA
ours 10 minutes
STUDYING AT: (tick where applicable)
□ Parkville □ Peninsula
as Learning Malaysia Sth Africa
in your possession any item/material that has not been authorised poks, notes, paper, electronic device/s, mobile phone, smart e, or writing on any part of your body. Any authorised items are at desk, chair, in your clothing or otherwise on your person will be
be removed from the room. This includes retaining, copying, at of exam material for personal use or to share with any other exam. instructions, or attempting to cheat or cheating in an exam is a the Monash University (Council) Regulations, or a breach of ash University (Academic Board) Regulations.
NO
YES TEMS NO
his section if required to write answers within this paper DESK NUMBER:

a. What is Apache Spark? What are the two advantages of unified stack in Spark? (1 + 2 = 3 Marks)

Write your answer below



Continue your answer below

	be server logs called "logs.txt". The contents of "logs.txt" file is shown below. Please emplete the program in the space provided. (2 Marks)
	INFO This is a message with content INFO This is some other content WARN This is a warning ERROR Something bad happened WARN More details on the bad thing
	(}
	from pyspark import SparkContext
	# Start your code here
	sc = SparkContext(master="local[2]", appName="Errors and warnings Count") lines = sc.textFile("logs.txt")
C	
	# End your code here
	Office use only

c. Write a program that does word count of the words ERROR and WARN found in the

a. What are broadcast variables? Why do we need broadcast variables when working with Apache Spark? (1 + 1 = 2 Marks)
 Write your answer below

b. We want to perform a log analysis. The input data consists of log messages of varying degrees of severity, along with some blank lines. We want to compute how many log messages appear at each level of severity. The contents of "input.txt" file is shown below.

INFO This is a message with content INFO This is some other content (empty line)
INFO Here are more messages
WARN This is a warning (empty line)
ERROR Something bad happened
WARN More details on the bad thing
INFO back to normal messages

The expected output of the operations is as below.

[('INFO', 4), ('WARN', 2), ('ERROR', 1)]

Write the code that will produce the expected output Assume spark context object 'sc' has already been initialised. (5 Marks)

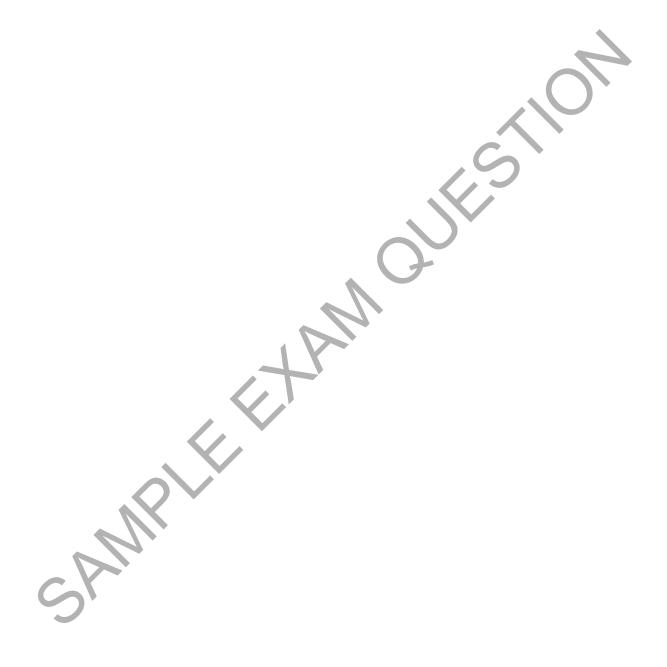
from pyspark import SparkContext

Start your code here

sc = SparkContext(master="local[2]", appName="Errors and warnings Count")
Input = ssc.textFile("input.txt")

End your code here

c. List four different sections of Spark Web UI and briefly explain all of them. (1 + 2 = 3 Marks)



<u>Offic</u>	<u>ce use only</u>

Question 3:

a. What is data visualisation? List two importance of data visualisation? What are the factors you need to be aware of before visualising the data? Please explain. (1 + 1 + 2 = 4 Marks)

- b. What are the benefits of using Apache Spark and MongoDB together? Assume you have a database named "FIT5202" and a collection named "zips" in MongoDB database. The information on the attributes are as follows
 - The _id field holds the zip code as a string.
 - The city field holds the city name. A city can have more than one zip code associated with it as different sections of the city can each have a different zip code
 - The **state** field holds the two-letter state abbreviation.
 - The **pop** field holds the population.
 - The **loc** field holds the location as a latitude longitude pair.

Assume that spark session object (i.e. spark) has been initialised. The analysis required is "Find the states with populations above 10 Million". Alice, our data analyst, is only familiar with SQL queries so she provided you with the following SQL query:

SELECT state, SUM(pop) AS totalPop FROM zips GROUP BY state HAVING totalPop >= (10*1000*1000)

You read the data from the MongoDB using the command below:

zips_df = spark.read.format("com.mongodb.spark.sql.DefaultSource").load()

Use the **functions provided by dataframe** to find the states with populations above 10 million. (1 + 5 = 6 Marks)

Write your answer below

SAMPLEEXAM

a. What is Machine Learning and why should you use machine learning with Spark? In Apache Spark, machine learning pipelines provide a uniform set of high-level APIs built on top of DataFrames. It makes easier to combine multiple algorithms into a single pipeline, or workflow. The key concepts introduced by the Pipelines API are DataFrame, Transformer, Estimator, Pipeline, and Parameter. What is a Transformer and an Estimator? (2 + 2 = 4 Marks)

Write your answer below

b. Suppose we have a set of data comprising; height, weight and shoe size of some customers. The aim is to predict the shoe size of a new customer given only height and weight information.

Height (in cm)	Weight (in kg)	Shoe Size
158	58	36
158	59	36
158	63	36
160	59	38
160	60	38
163	60	38
163	61	38
163	64	40
165	64	40
165	61	40
165	62	40
168	65	40
168	62	40

Write the formula to calculate the Euclidean distance? A new customer named "Matthew" has height 161 cm and weight 61 kg. Using kNN, for k = 5, what is the (most) unlikely shoes size of Matthew? (2 + 4 = 10 Marks)

SAMPLEETAMOULESTION

a. What is the difference between Supervised Learning and Unsupervised Learning? Mention any two differences. (2 Marks)

b. Consider the following data set consisting of the scores of two variables on each of seven individuals:

A	В
1.0	1.0
1.5	2.0
3.0	4.0
5.0	7.0
3.5	5.0
4.5	5.0
3.5	4.5
	3.0 5.0 3.5 4.5

Use the k-means algorithm to cluster the data in two clusters. The distance of each data point and the centroid (or mean) is calculated using Euclidean distance. The formula to calculate the euclidean distance is given below. (8 Marks)

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}.$$

SAMPLE EXAMINATES FILOS

SAMPLE FILOS

SAMPLE FILOS

SAMPLE FILOS

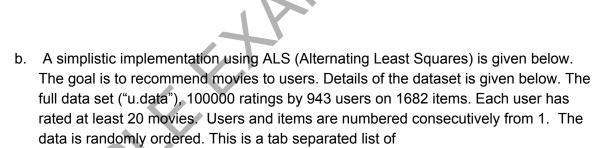
SAMPLE FILOS

SAMPLE FILOS

S

Office use only		

 a. 'People who bought this also bought...' recommendations seen on Amazon is based on which algorithm? What is the difference between Association Rules and Collaborative Filtering? (1 + 1 = 2 Marks)



user id | item id | rating | timestamp

The time stamps are unix seconds since 1/1/1970 UTC. The sample data contents of "u.data" file is shown below.

N				
1	196	242	3	881250949
	186	302	3	891717742
	22	377	1	878887116
	244	51	2	880606923
	166	346	1	886397596

Write the necessary code below to develop an ALS based recommendation model for movie recommendations. First examine the dataset and perform necessary steps to convert the dataset into DataFrame to make it ready for the algorithm. (8 Marks) Write your answer below

from pyspark import SparkContext from pyspark.sql import SparkSession, Row from pyspark.ml.recommendation import from pyspark.ml.evaluation import_ appName="Collaborative Filtering with PySpark" # initialize the spark session spark = SparkSession.builder.appName(appName).getOrCreate() # get sparkcontext from the sparksession sc = spark.sparkContext # Step1: the data is loaded to an RDD movielens_rdd = # Step 2: process the data into appropriate structure # Step 3: convert the rdd to dataframes # Step 4: split the dataset into training and test data (70% training and 30% test) (trainingData, testData) = # Step 5: build the recommendation model using ALS on the training data # Use maxIter = 5, regParam = 0.01, coldStartStrategy = "drop",

implicitPrefs = False

Step 6: predict the top movies for some selected users predictions = model.transform(testData)
Step 7: find and print the accuracy of the model
/, V
Office use on

This page can be utilised to perform rough calculations and will not be marked unless clearly indicated.



This page can be utilised to perform rough calculations and will not be marked unless clearly indicated.

SAMPLEENAMOULESTION