Name: ________________________

Score: ______ / ______

BDA Final Exam - Makeup

Part 1: General Info: [10 points]

Fill in the table with the following words.

Association Rule Mining, Classification, Clustering, Hadoop, MapReduce, Periodic, Sentiment analysis, Strongly connected, Text Mining, Web mining, Irreducible, Dangling nodes

____: finds interesting relationships (affinities) between variables (items or events)

____: is a framework for parallel processing with minimal movement of data and near-realtime results

____: pages with no outgoing edges.

____: is the process of discovering intrinsic relationships from Web data (textual, linkage, or usage)

____: strongly connected Web graph

____: is non-relational system of distributed and cost-effective data storage on commodity hardware

____ is a technique used to detect favorable and unfavorable opinions toward specific products and services

____ works on unstructured data in Word documents, PDF files, XML files, etc

____: is an example for supervised learning

____: is a technique used for automatic identification of natural groupings of things

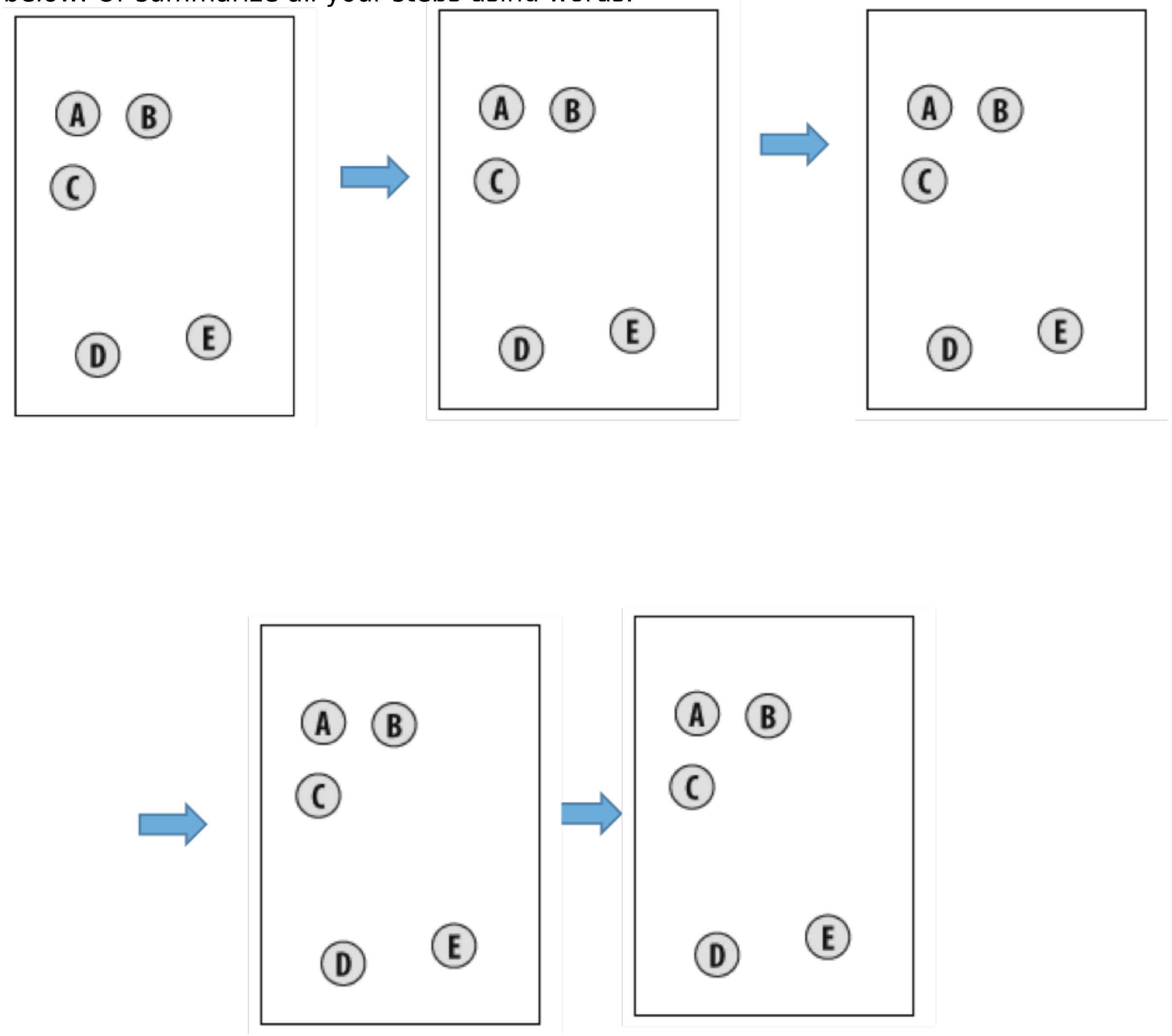
Part 2: Hadoop: [20 points]

Use map reduce to count how many restaurants for each rate.
Describe the overall Map-Reduce Processing graph. Give the name of each stage and list all elements in each stage.

User ID	Restaurant ID	Rating	City ID
124	294	2	985
349	827	4	998
725	751	4	982
346	294	2	985
578	827	3	998
124	934	4	051
725	294	3	985
766	751	5	982
725	294	2	985
766	294	1	985

Part 3: Clustering: [20 points]

Given the following five examples; A,B,C,D and E. The initial cluster centres are **A** and **C**. Show how the k-mean clustering is performed, by drawing the cluster canters each time in each figure below. Or summarize all your steps using words.



Part 4: Association Rule Mining: [20 points]

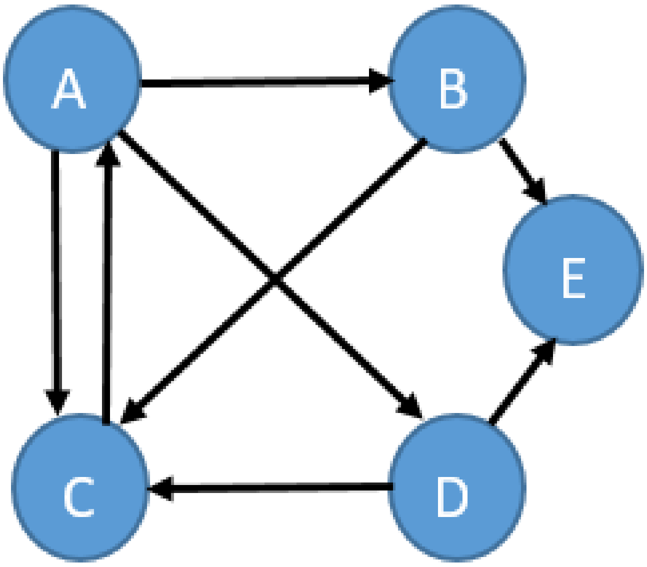
Use Apriori Algorithm to find all frequent itemsets with minimum support percentage of 25%. List the itemsets together with their supports

Transaction ID	Items
1	Monitor, Tablet
2	Printer, Tablet, headset
3	monitor, Printer, Tablet, headset
4	Monitor, Printer, Tablet
5	monitor, Printer, Tablet, headset
6	Tablet, Printer
7	Monitor, Tablet, headset, Printer
8	Tablet, monitor, headset
9	monitor, Tablet, headset, Monitor
10	Monitor, Printer, monitor

Part 5: PageRank: [20 points]

Write two advantages of PageRank.

For the following web graph:



1- [8 points] Write down the transition matrix ***A*** of the above web graph

2- [6 points] Use a damping factor of 0.85 and write down the equations of PageRank for the four pages in the graphs.

3- [2 points] Is matrix ***A*** Irreducible? explain why

4- [2 points] Is matrix ***A*** stochastic? explain why

Hint:

$$PR(i) = (1 - d) + d\sum_{j=1}^{n} A_{ji}P(j)$$

Part 6: [10 points] SCI Question

Write a short essay that connects one of the studied topics to the Science of Creative Intelligence (SCI). You can pick any topic.