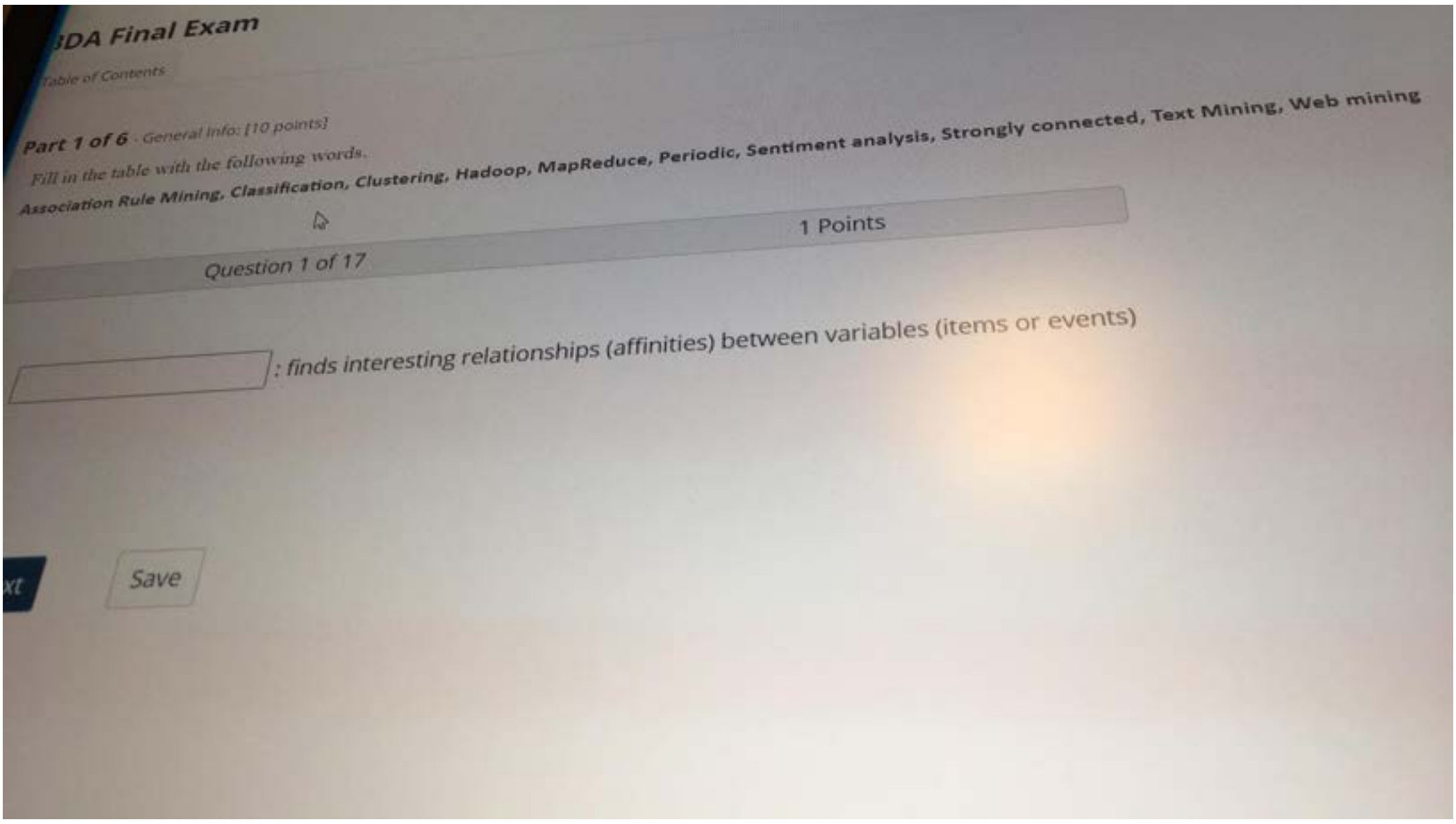
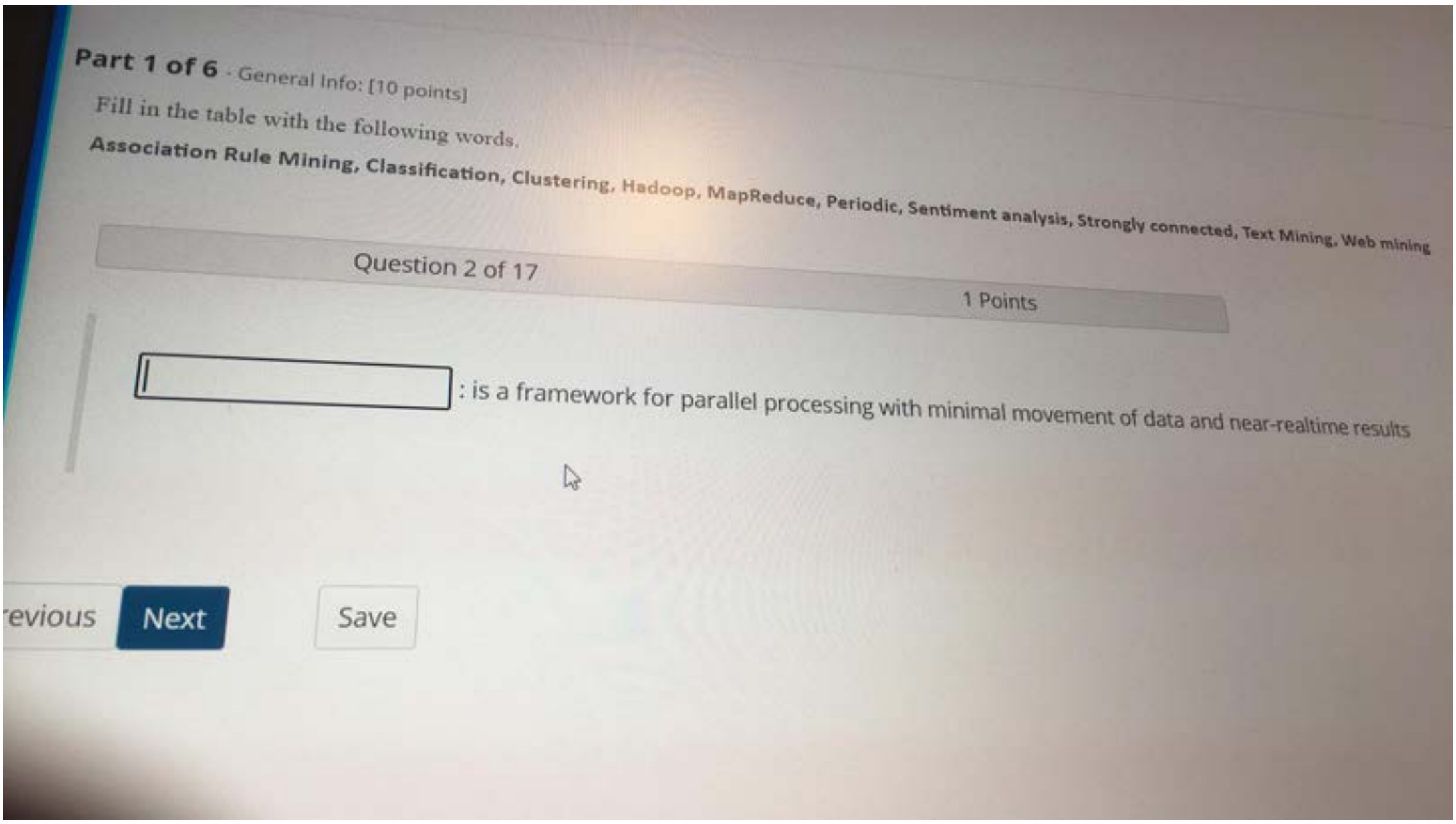
Question 1. General Info - Fill in blank



Association rule mining



Mapreduce

Part 1 of 6 - 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

Question 3 of 17 — 1 Points

[ ] : is a feature of a state i in a Markov that there exists a directed cycle that the chain has to traverse.

Previous | Next | Save

periodic

Part 1 of 6 - 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

Question 4 of 17 — 1 Points

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

Previous | Next | Save

Web mining

Part 1 of 6 - 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

Question 5 of 17 — 1 Points

[ ] : is a feature of a directed graph if and only if, for each pair of nodes u, v ∈ V, there is a path from u t

revious | Next | Save

Strongly connected

Part 1 of 6 - General Info: [10 points]

Fill in the table with the following words.

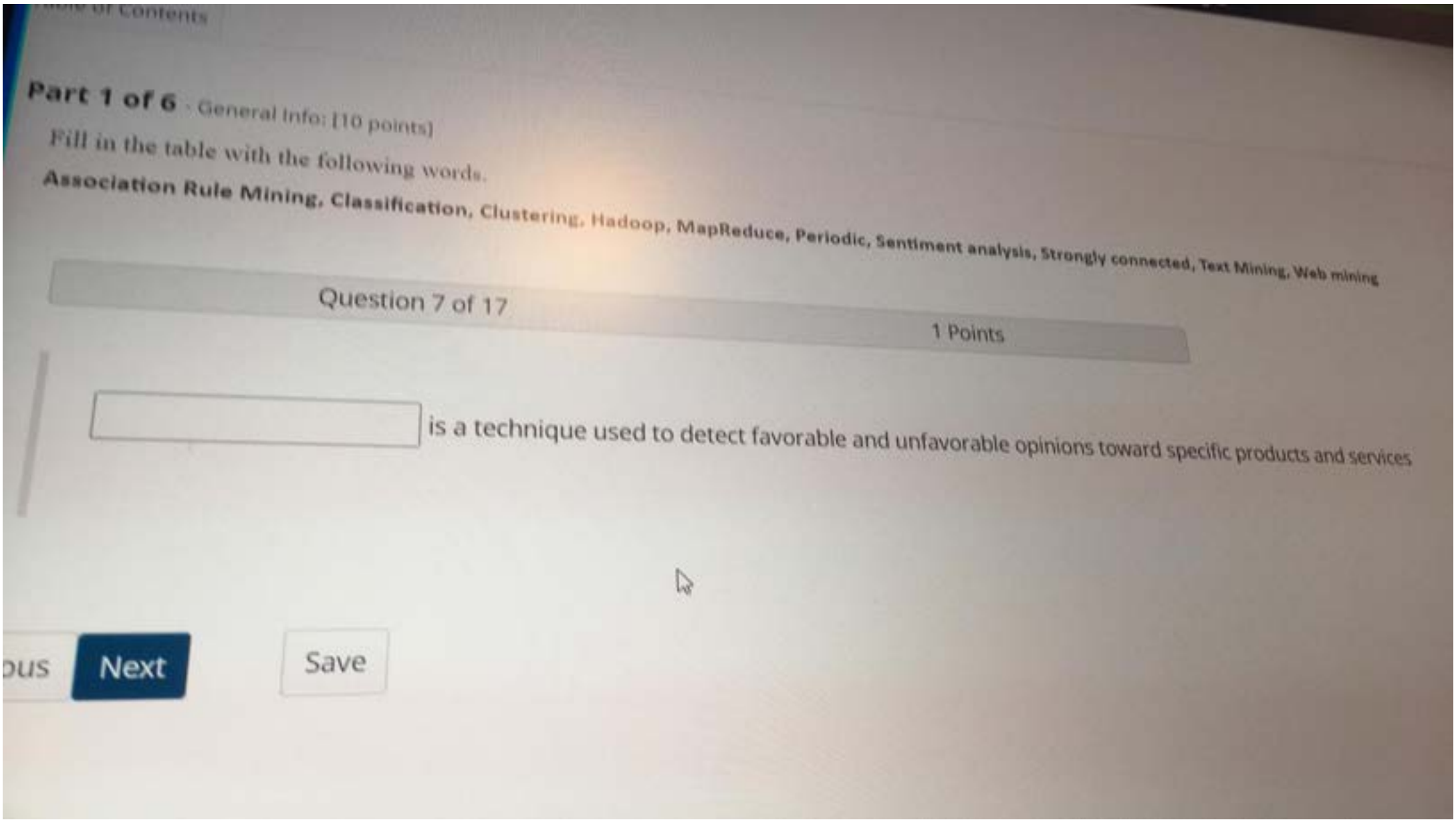
Association Rule Mining, Classification, Clustering, Hadoop, MapReduce, Periodic, Sentiment analysis, Strongly connected, Text Mining, Web mining

Question 6 of 17 — 1 Points

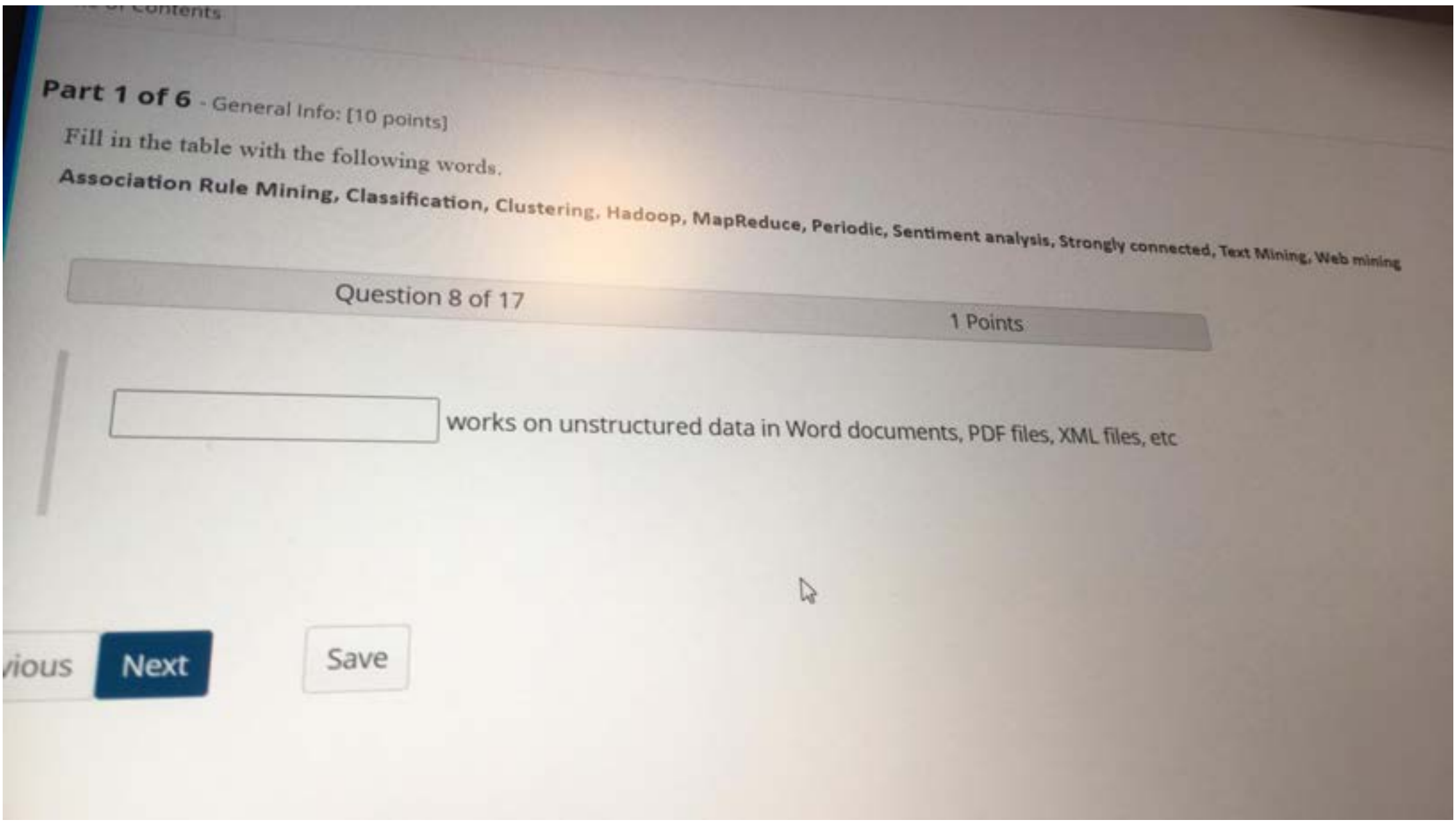
[ ] : is non-relational system of distributed and cost-effective data storage on commodity hardware

vious | Next | Save

Hadoop



Sentiment analysis



Text mining

Part 1 of 6 - 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

Question 9 of 17 — 1 Points

[ ] : is an example for supervised learning

Previous | Next | Save

Classification

Table of Contents

Part 1 of 6 - 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

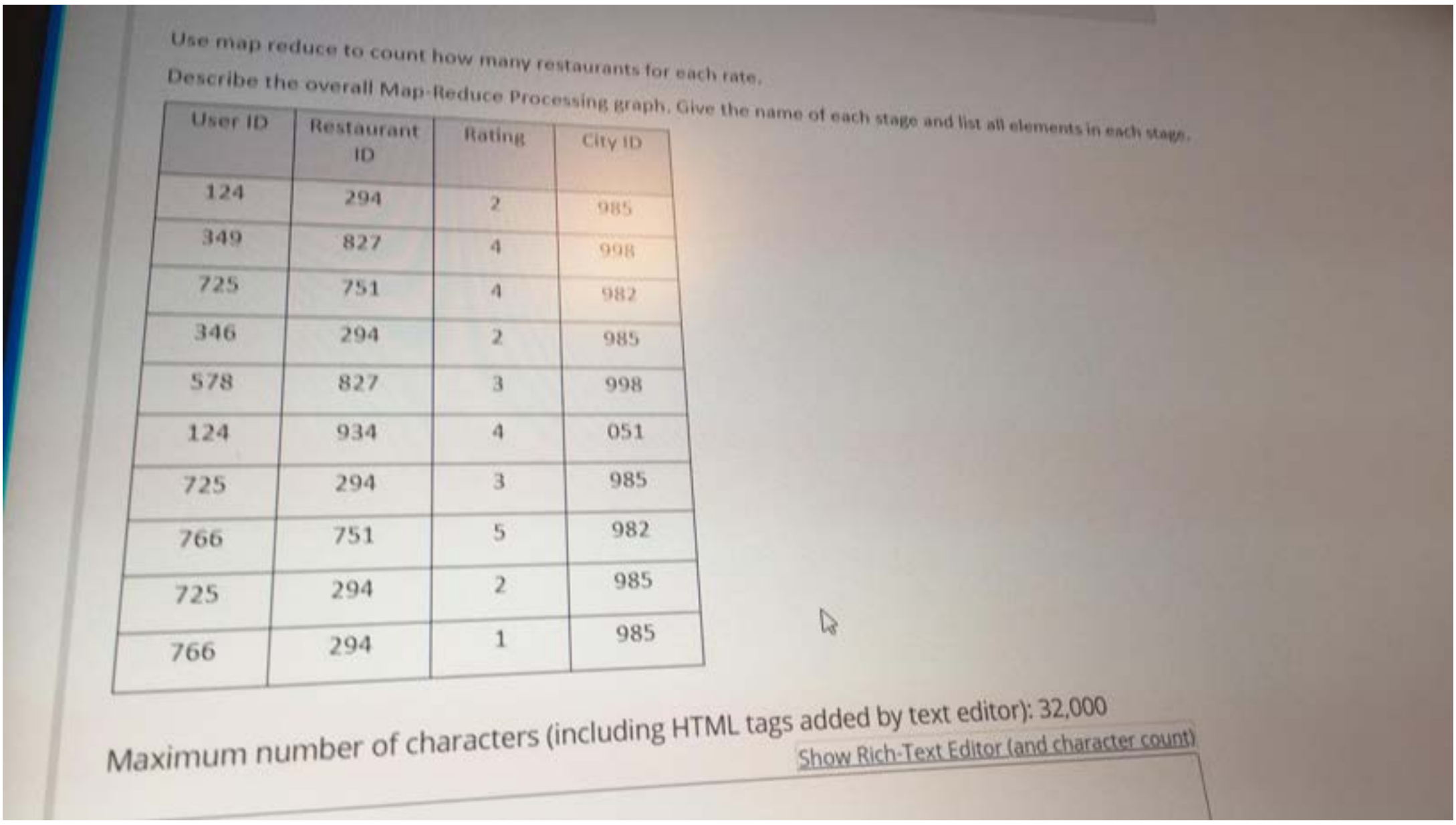
Question 10 of 17 — 1 Points

[ ] : are examples for unsupervised learning

Previous | Next | Save

clustering

Question 2. Hadoop - Map Reduce



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

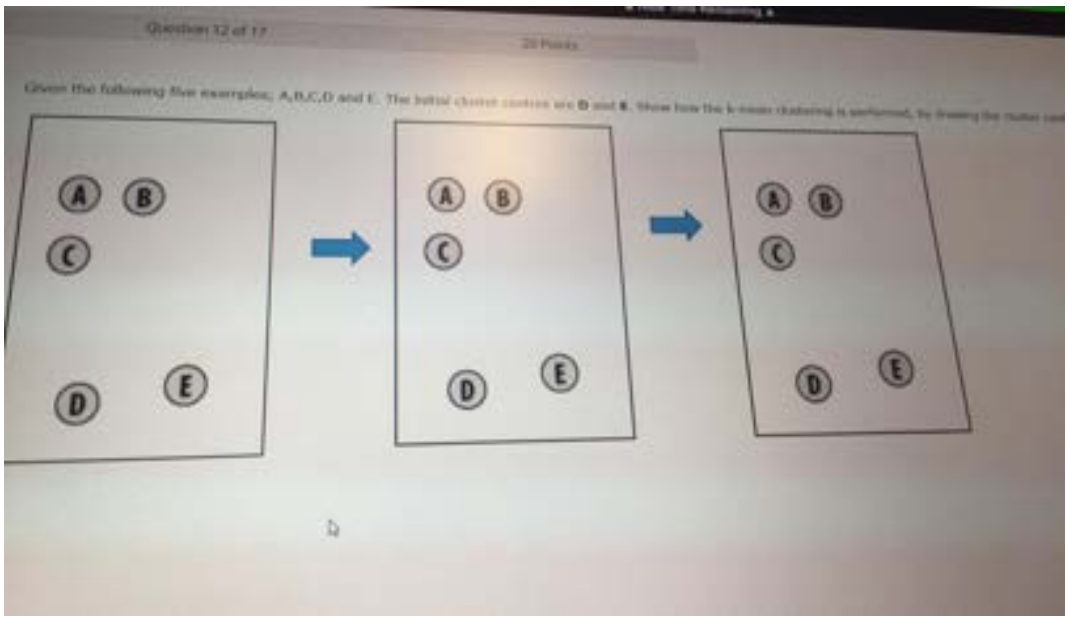
Maximum number of characters (including HTML tags added by text editor): 32,000

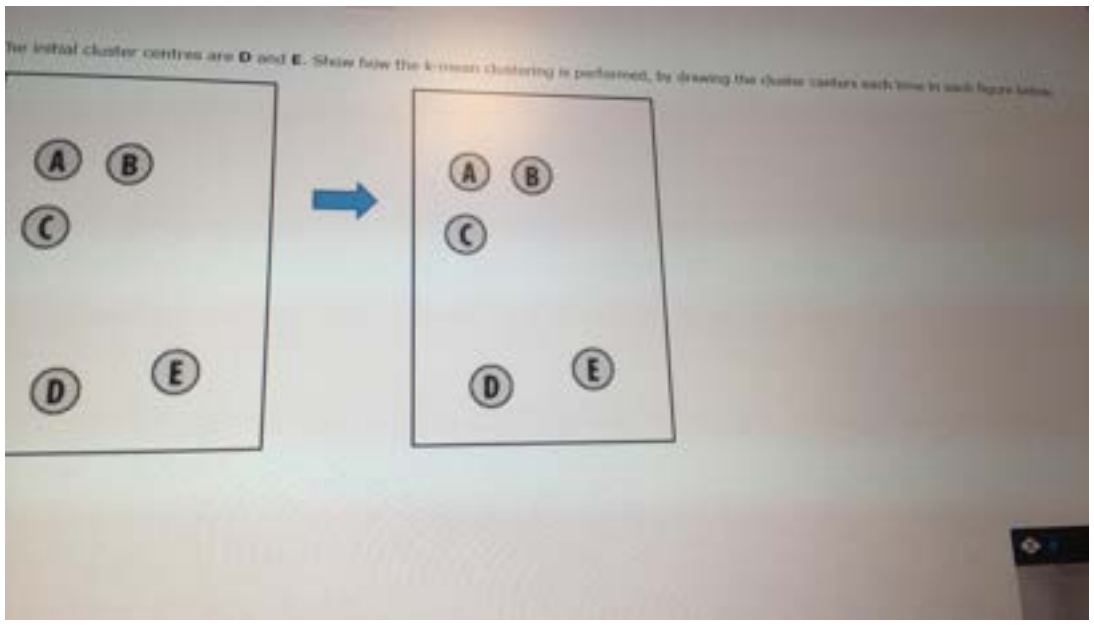
Show Rich-Text Editor (and character count)

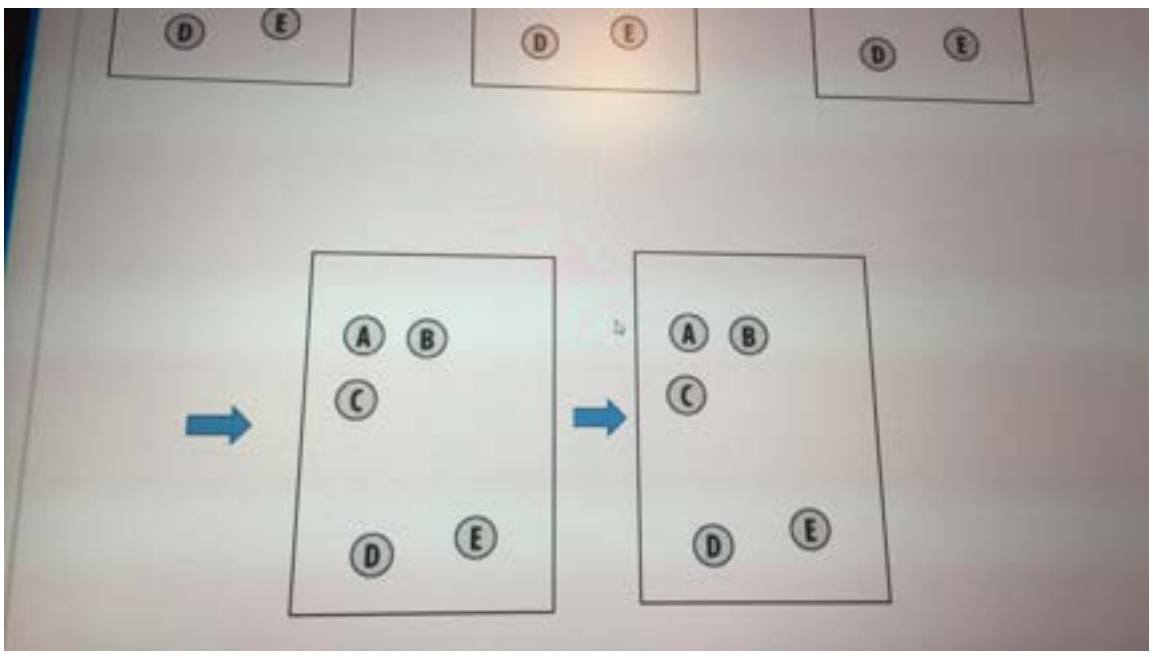
```
Input --> Splitting --> Mapping --> Shuffling --> Reducing --> Final result
=== Input Rate occurrence
Rate2 Rate4 Rate4 Rate2 Rate3 Rate4 Rate3 Rate5 Rate2 Rate1

=== Splitting phase

Splitter1: Rate2 Rate4 Rate4 Rate2 Rate3 Rate4 Rate3 Rate5 Rate2 Rate1

=== Mapping phase
Mapper1: ('Rate1', 1), ('Rate2', 3), ('Rate3', 2), ('Rate4', 3), ('Rate5', 1)
=== Shuffling phase
Shuffler1: ('Rate1', 1)
Shuffler2: ('Rate2', 3)
Shuffler3: ('Rate3', 2)
Shuffler4: ('Rate4', 3)
Shuffler5: ('Rate5', 1)
=== Reducing phase
Reducer1: Rate1,1
Reducer2: Rate2,3
Reducer3: Rate3,2
Reducer4: Rate4,3
Reducer5: Rate5,1
=== Final result
('Rate1', 1), ('Rate2', 3), ('Rate3', 2), ('Rate4', 3), ('Rate5', 1)
```

Input split>an input to a MapReduce is split into fixed-size parts called input splits. The term "input split" refers to a portion of the input that is processed by a single map.
Mapping>In this step, each split's data is handed to a mapping function, which generates output values.
Shuffling>This step consumes the Mapping phase's output. Its job is to bring together all of the relevant records from the Mapping phase's output.
Reducing>The output values from the Shuffling step are consolidated in this phase. This phase takes the values from the Shuffling phase and merges them into a single output value. In a nutshell, this stage summarizes the whole dataset.

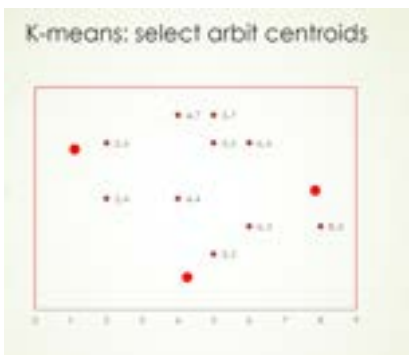
Question 3. Cluster - Kmean



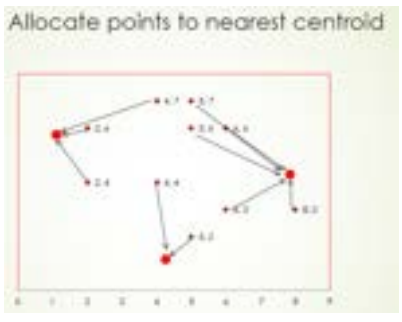




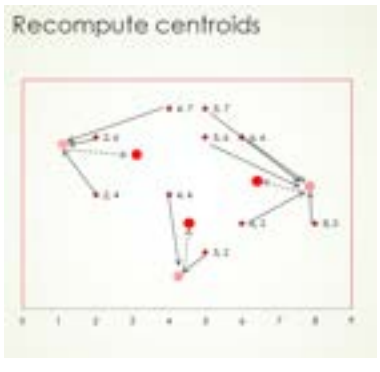
Step 1: centrioids D, E, calculate distance from A,B, C to divide the clusters



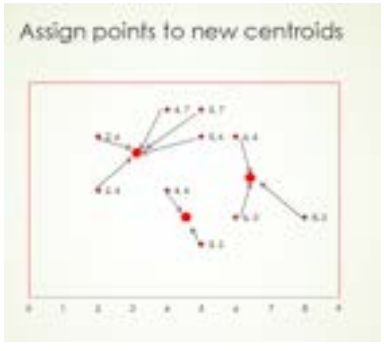
Step 2 AC to D, B to E because the AC closer to D than E, B closer to E than D



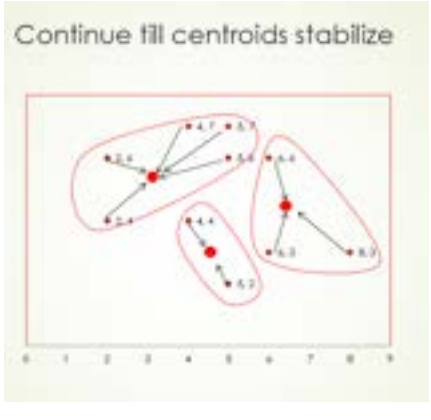
Step 3 Calculate new D’ E’



Step 4 - Reassigning points to new centroids D’ E’



Step 5 - Repeating until controids stabilize



Question 4. Association Rule - Apriori

```
=== Input
hairbrush shampoo
hairdryer shampoo haircond
mirror hairdryer shampoo
hairbrush hairdryer shampoo
mirror hairdryer shampoo haircond
shampoo hairdryer
hairbrush shampoo haircond hairdryer
shampoo mirror haircond
mirror shampoo haircond hairbrush
hairbrush hairdryer mirror

=== raw items: ['hairbrush', 'haircond', 'hairdryer', 'mirror', 'shampoo']
total trans=10, support*trans =2.5
=== 1-item itemsets
hairbrush   5
haircond    5
hairdryer   7
mirror      5
shampoo     9
=== 2-item itemsets
hairbrush, hairdryer   3
hairbrush, shampoo     4
haircond, hairdryer    3
haircond, mirror 3
haircond, shampoo5
hairdryer, mirror3
hairdryer, shampoo     6
mirror, shampoo  4
---> drop itemsets
      hairbrush,haircond     2
      hairbrush,mirror 2
=== 3-item itemsets
haircond, hairdryer, shampoo      3
haircond, mirror, shampoo    3
---> drop itemsets
      hairbrush,haircond,hairdryer      1
      hairbrush,haircond,mirror    1
      hairbrush,haircond,shampoo   2
      hairbrush,hairdryer,mirror   1
      hairbrush,hairdryer,shampoo  2
      hairbrush,mirror,shampoo     1
      haircond,hairdryer,mirror    1
      hairdryer,mirror,shampoo     2
=== 4-item itemsets
NO 4-item itemsets
---> drop itemsets
      hairbrush,haircond,hairdryer,mirror     0
      hairbrush,haircond,hairdryer,shampoo    1
      hairbrush,haircond,mirror,shampoo 1
      hairbrush,hairdryer,mirror,shampoo      0
      haircond,hairdryer,mirror,shampoo 1
```

Question 13 of 17 — 20 Points

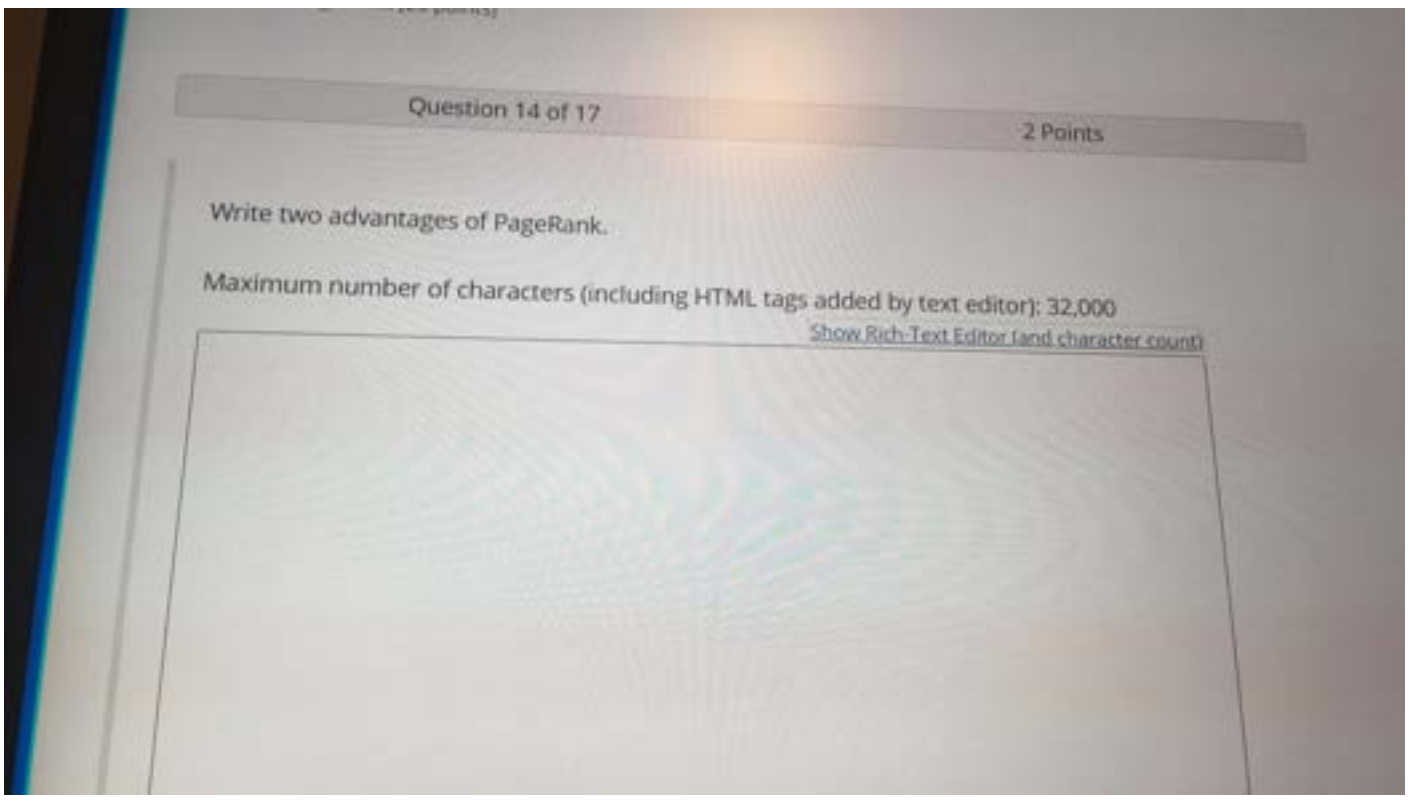
Use Apriori Algorithm to find all frequent itemsets with minimum support percentage of 25%. List the itemsets together with their supports.

Transaction ID	Items
1	Hairbrush, Shampoo
2	Hair dryer, Shampoo, Hair conditioner
3	Mirror, Hair dryer, Shampoo, Hair conditioner
4	Hairbrush, Hair dryer, Shampoo
5	Mirror, Hair dryer, Shampoo, Hair conditioner
6	Shampoo, Hair dryer
7	Hairbrush, Shampoo, Hair conditioner, Hair dryer
8	Shampoo, Mirror, Hair conditioner
9	Mirror, Shampoo, Hair conditioner, Hairbrush
10	Hairbrush, Hair dryer, Mirror

Maximum number of characters (including HTML tags added by text editor): 32,000

Question 5. PageRank

https://colab.research.google.com/drive/1ZE0SMPDozimHKoEDlmH60DJ8yYA5peDS?usp=sharing

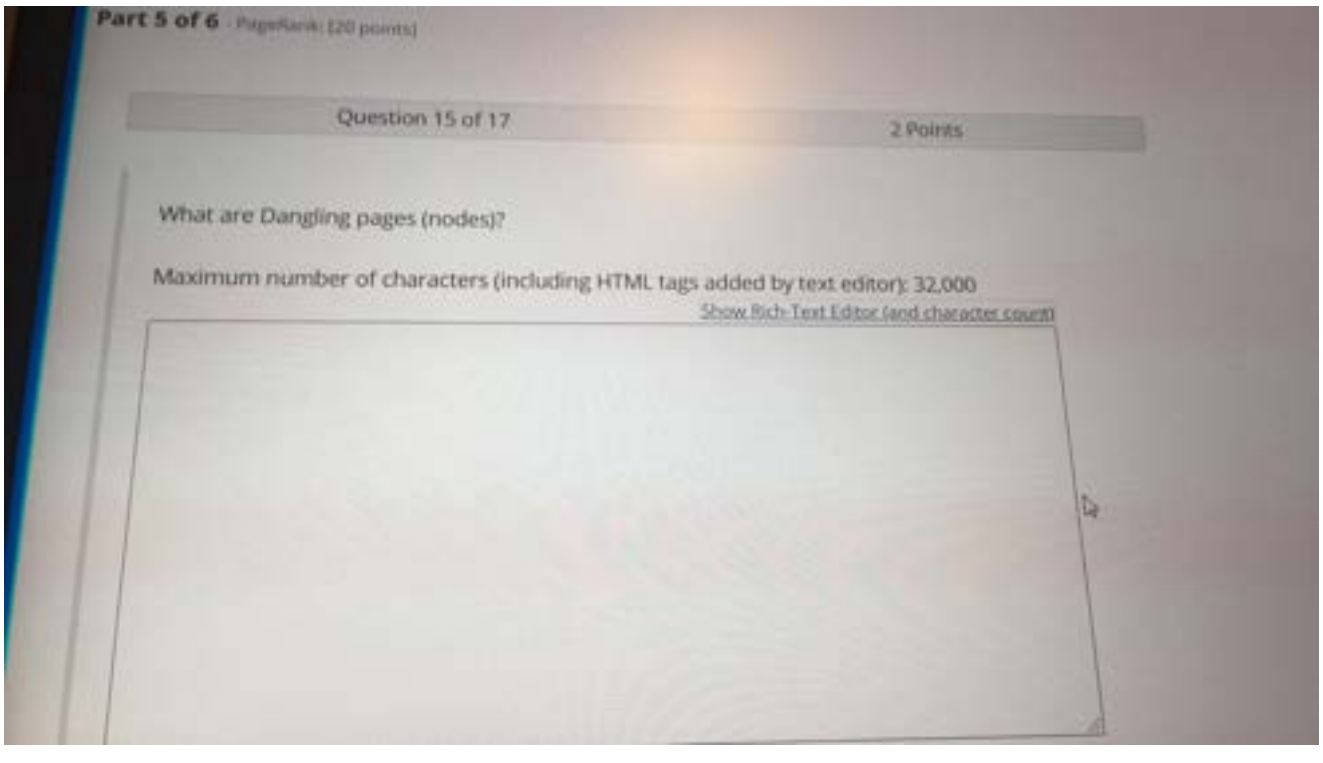


Choose 2 of 5 - **dont write out all**

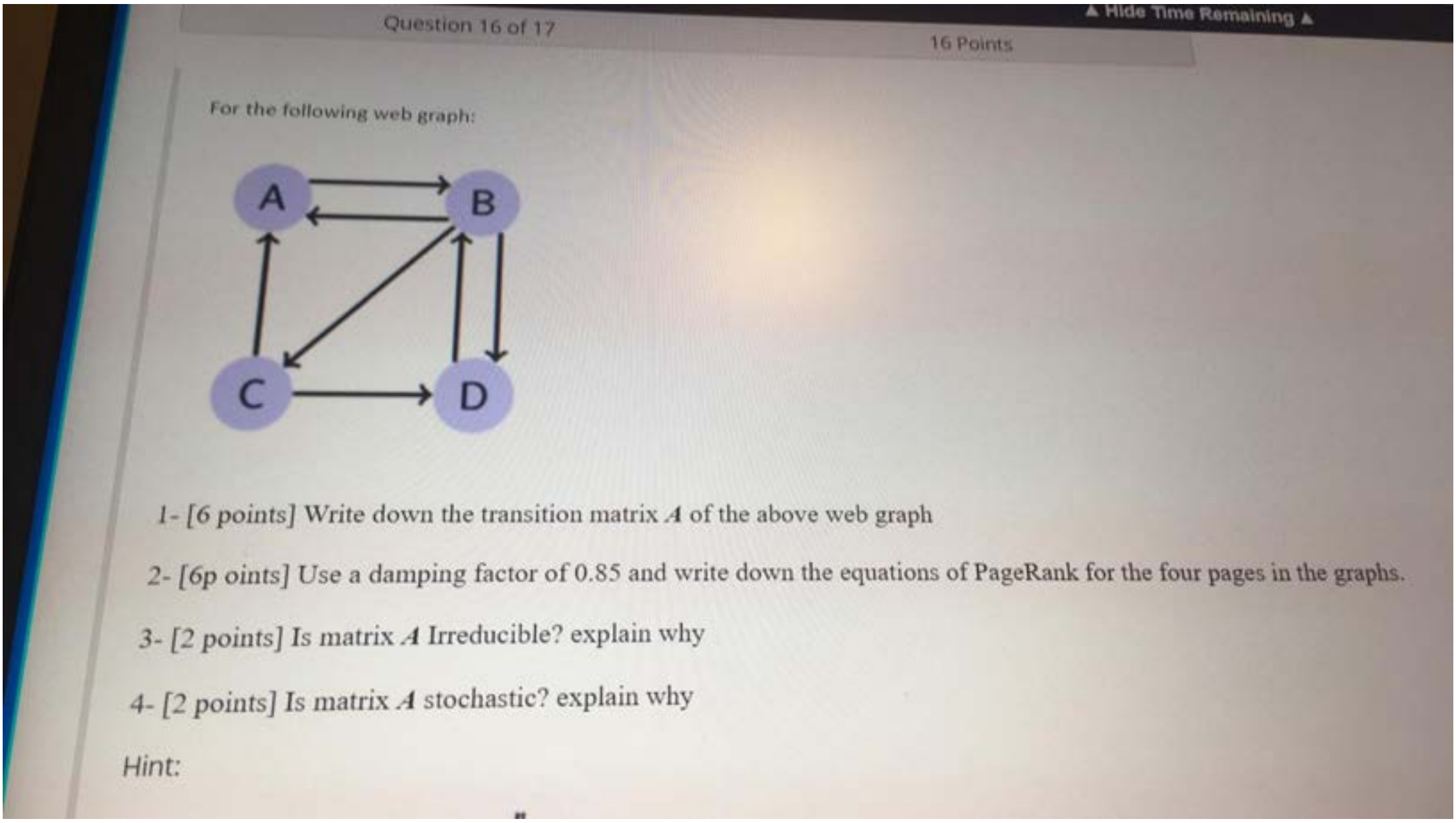
1. Since it pre-computes the rank score it takes less time and hence it is fast.
2. It is more feasible as it computes rank score at indexing time not at query time
3. It returns important pages as Rank is calculated on the basis of the popularity of a page

Hey guys this is for weighted page rank algorithm,not the page rank algorithm, so we probably should not include it

4. Quality of the pages returned by this algorithm is high as compared to PageRank algorithm.
5. It is more efficient than PageRank because rank value of a page is divided among it’s outlink pages according to importance of that page.



Dangling links are simply links that point to any page with no outgoing links. They affect the model because it is not clear where their weight should be distributed, and there are a large number of them.




```
=== Input
A -> B
B -> A
B -> C
B -> D
C -> A
C -> D
D -> B

Influence relationship
Ra = 0.33Rb + 0.5Rc
Rb = Ra +  Rd
Rc = 0.33Rb
Rd = 0.33Rb + 0.5Rc

===>> Transition Matrix A
[[0.         0.33333333 0.5        0.        ]
 [1.         0.         0.         1.        ]
 [0.         0.33333333 0.         0.        ]
 [0.         0.33333333 0.5        0.        ]]
=== Page Rank with damping factor = 0.85, init value = 1
PRa = 0.15 + 0.85*(0.33Rb + 0.5Rc)
PRb = 0.15 + 0.85*(Ra +  Rd)
PRc = 0.15 + 0.85*(0.33Rb)
PRd = 0.15 + 0.85*(0.33Rb + 0.5Rc)
```

```
Khong can ghi tinh toan iteration
iter1 = [0.858, 1.85, 0.433, 0.858]
iter2 = [0.858, 1.609, 0.674, 0.858]
iter3 = [0.892, 1.609, 0.606, 0.892]
iter4 = [0.863, 1.667, 0.606, 0.863]
iter5 = [0.88, 1.618, 0.622, 0.88]
iter6 = [0.873, 1.646, 0.608, 0.873]
iter7 = [0.875, 1.634, 0.616, 0.875]
iter8 = [0.875, 1.637, 0.613, 0.875]
```

A **matrix** is **reducible** if and only if it can be placed into block upper-triangular form by simultaneous row/column permutations. In addition, a **matrix** is **reducible** if and only if its associated digraph is not strongly connected. A square **matrix** that is not **reducible** is said to be **irreducible**.

In mathematics, a **stochastic matrix** is a square matrix used to describe the transitions of a Markov chain. Each of its entries is a nonnegative real number representing a probability. It is also called a probability matrix, transition matrix, substitution matrix, or Markov matrix.

Question 17 of 17

10 Points

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

Maximum number of characters (including HTML tags added by text editor): 32,000

Show Rich-Text Editor (and character count)