Automated Segmentation of Question-Solution Documents from NPTEL-NOC

Master Thesis Project report submitted to

Indian Institute of Technology Kharagpur

in partial fulfilment for the award of the degree of

Master of Technology

in

Computer Science and Engineering

by

Heeramani Prasad (15CS30015)

Under the supervision of

Professor Partha Pratim Das

Department of Computer Science and Engineering

Indian Institute of Technology Kharagpur

Autumn Semester, 2019-20

November 05, 2019

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Automated Segmentation of Question-Solution Documents from NPTEL-NOC**" submitted by **Heeramani Prasad (15CS30015)** to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Master of Technology in Computer Science and Engineering is a record of bona fide work carried out by him under my supervision and guidance during Autumn Semester, 2019-20.

Date: November 05, 2019

Place: Kharagpur

Professor Partha Pratim Das
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur
Kharagpur - 721302, India

Acknowledgements

We would like to express our sincere gratitude to our Master Thesis Project guide Professor Partha Pratim Das for the continuous support for our study and related research, for his patience, motivation, and immense knowledge. His guidance helped us in all the time of research and writing of this report. We could not have imagined having a better advisor and mentor.

Besides our guide, we would like to thank Phd student Mr.Himadri B.G.S. Bhuyan for his insightful comments and encouragement to widen our research from various perspectives.

Thank you.

Contents

Abbreviations

PDF	**P**ortable **D**ocument **F**ormat

Chapter 1

Introduction

Portable document format (PDF) is a common output format for electronic documents.PDF documents preserve the look and feel of the original documents by describing the low-level structural objects such as a group of characters, lines, curves and images and associated style attributes such as font, color, stroke, fill, and shapes,etc.Chao and Fan [2004] However, most PDF documents are untagged and don't have the basic high level logical structure information such as words, text lines, paragraphs, logos, and figure illustrations, which makes reusing, editing or modifying the layout or the content of the document difficult. Although originally, the Portable Document Format (PDF) was designed for the final presentation of a document, there were trends to extend the capability of PDF to more than a viewable format Chao and Fan [2004] and to recover the PDF document layout and its content.

The absence of effective means to extract text from these PDF files in a layout-aware manner presents a significant challenge for developers of biomedical text mining or biocuration informatics systems that use published literature as an information source.Ramakrishnan et al. [2012]

A strategy for document analysis is presented which uses Portable Document Format (PDF the underlying file structure for Adobe Acrobat software) as its starting point. This strategy examines the appearance and geometric position of text and image blocks distributed over an entire document. PDF is shown to be a useful intermediate stage in the bottom-up analysis of document structure. Its information on line spacing and font usage gives important clues in bridging the semantic gap between the pdf page and its fully analysed, block-structured form. Analysis of PDF can yield not only accurate page decomposition but also sufficient document information for the later stages of structural analysis and document understanding.Lovegrove and Brailsford [1995]

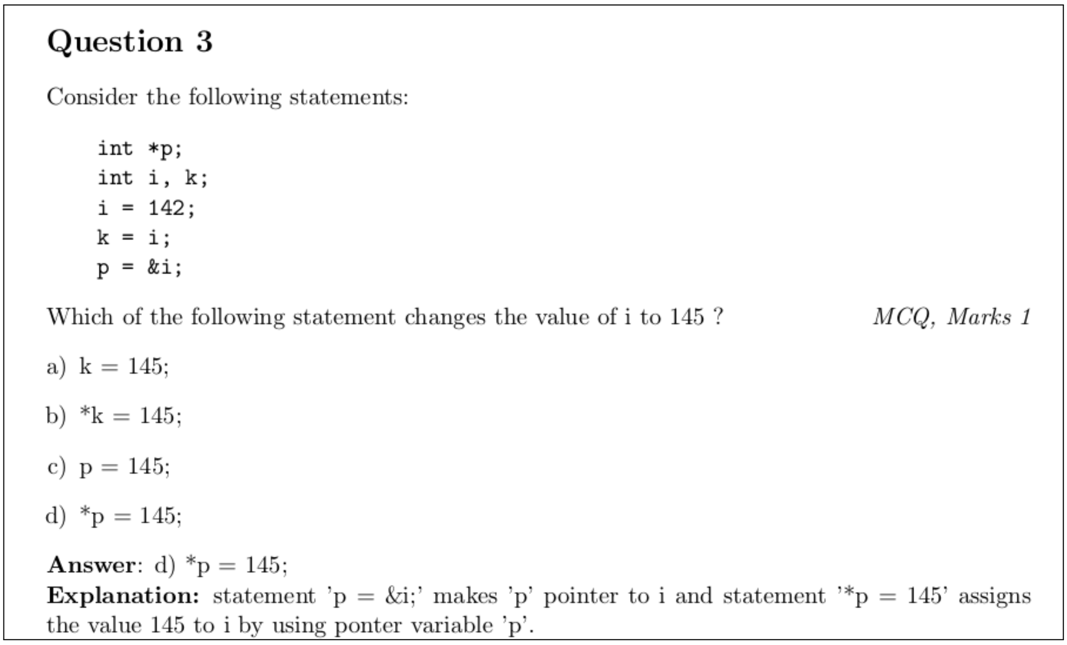
We generate vast number of NPTEL-NOC documents for students. These documents include *question with multiple options*(here we have only four option), *Fill in the blanks* and *comprehension question* having multiple question in it. It has a question part which start from "Question NUMBER" after that its content are there like code portion or query and question related to it. After question ended there are four uniformly latex environment generated option "a) b) c) d)". These option may include simple one line text, option with diagram and figure in it and only diagram in case of mostly DBMS.

Here following are some sample question for reference.

We can see in figure 1.1 that is it taken from NPTEL-NOC MOOC C++ 4th run. It has question with code embedded in it. After question ends, it has *four* option *a)*, *b)*, *c)* and *d)* respectively.

Next type of question is *Fill in the blanks* as figure 1.2.It is taken from NPTEL-NOC C++ 4th run. It has only question along with solution

Last type of question to segment is *comprehensive* type.It is taken from NPTEL-NOC DBMS 2nd run. It has some passage, code or diagram related for some set of question which may be ranges from to two to five -six. It may vary depending on



Question 3

Consider the following statements:

```
int *p;
int i, k;
i = 142;
k = i;
p = &i;
```

Which of the following statement changes the value of i to 145 ? *MCQ, Marks 1*

a) k = 145;

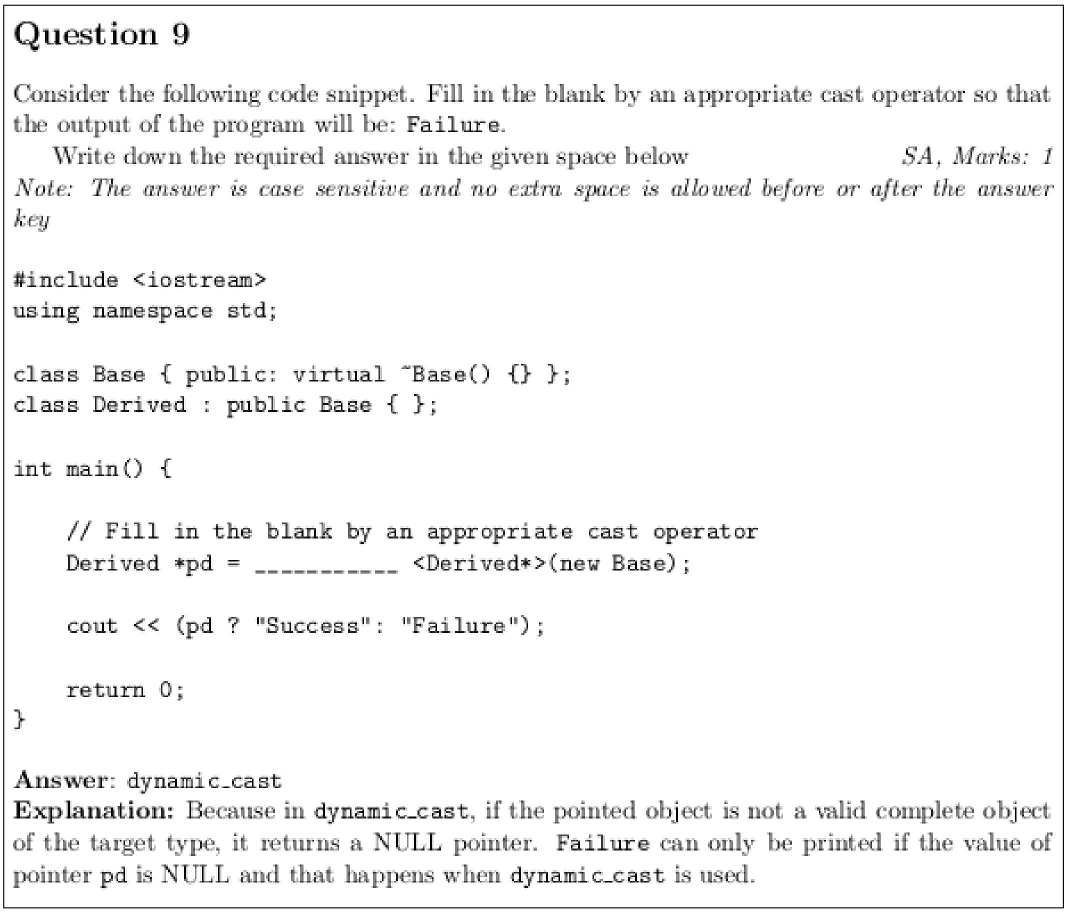
b) *k = 145;

c) p = 145;

d) *p = 145;

Answer: d) *p = 145;
Explanation: statement 'p = &i;' makes 'p' pointer to i and statement '*p = 145' assigns the value 145 to i by using ponter variable 'p'.

FIGURE 1.1: MOOC C++ 4th run, Assignment 1,Ques 3,Normal question.



Question 9

Consider the following code snippet. Fill in the blank by an appropriate cast operator so that the output of the program will be: `Failure`.

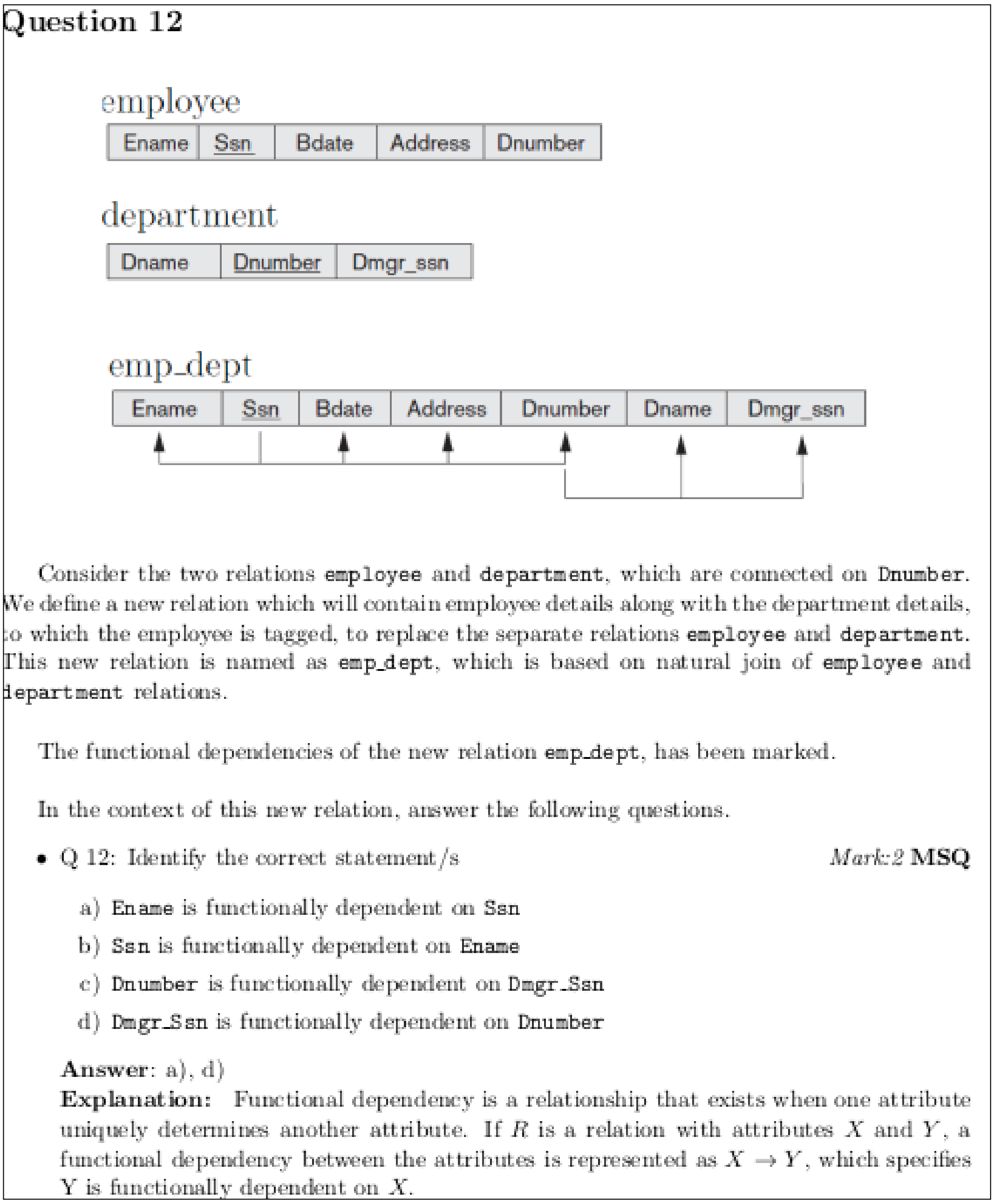
Write down the required answer in the given space below *SA, Marks: 1*

Note: The answer is case sensitive and no extra space is allowed before or after the answer key

```
#include <iostream>
using namespace std;

class Base { public: virtual ~Base() {} };
class Derived : public Base { };

int main() {

    // Fill in the blank by an appropriate cast operator
    Derived *pd = ___________ <Derived*>(new Base);

    cout << (pd ? "Success": "Failure");

    return 0;
}
```

Answer: `dynamic_cast`
Explanation: Because in `dynamic_cast`, if the pointed object is not a valid complete object of the target type, it returns a NULL pointer. `Failure` can only be printed if the value of pointer `pd` is NULL and that happens when `dynamic_cast` is used.

FIGURE 1.2: MOOC C++ 4th run, Assignment 7,Ques 9, Fill in the blanks.



Question 12

employee

Ename	Ssn	Bdate	Address	Dnumber

department

Dname	Dnumber	Dmgr_ssn

emp_dept

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn

Consider the two relations `employee` and `department`, which are connected on `Dnumber`. We define a new relation which will contain employee details along with the department details, to which the employee is tagged, to replace the separate relations `employee` and `department`. This new relation is named as `emp_dept`, which is based on natural join of `employee` and `department` relations.

The functional dependencies of the new relation `emp_dept`, has been marked.

In the context of this new relation, answer the following questions.

- Q 12: Identify the correct statement/s *Mark:2* **MSQ**

  a) `Ename` is functionally dependent on `Ssn`

  b) `Ssn` is functionally dependent on `Ename`

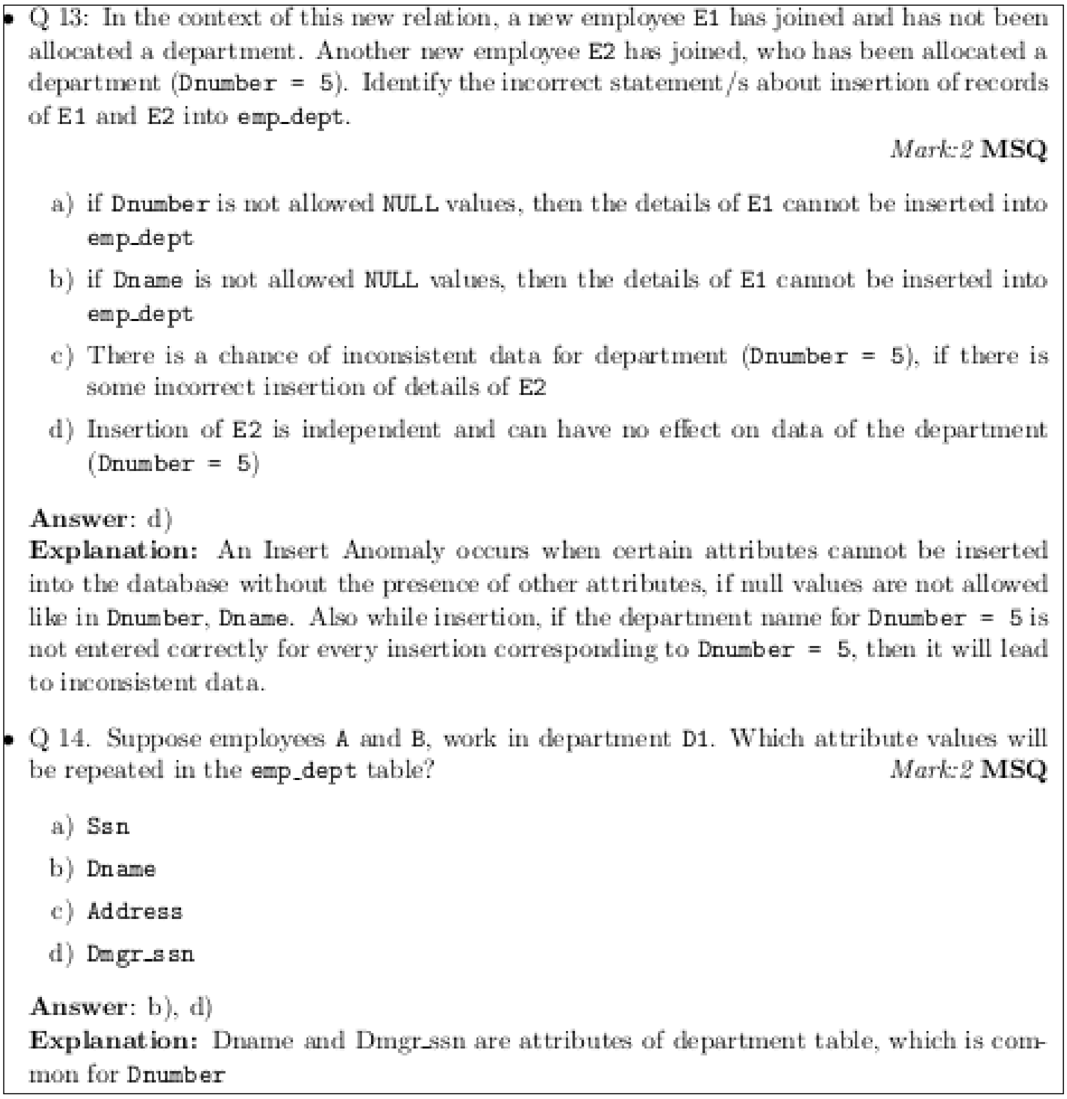
  c) `Dnumber` is functionally dependent on `Dmgr_Ssn`

  d) `Dmgr_Ssn` is functionally dependent on `Dnumber`

  Answer: a), d)
  Explanation: Functional dependency is a relationship that exists when one attribute uniquely determines another attribute. If R is a relation with attributes X and Y, a functional dependency between the attributes is represented as $X \to Y$, which specifies Y is functionally dependent on X.

Figure 1.3: DBMS 2nd run, Assignment 4,Ques 12 - 14, Comprehension question. All these question have multiple option along with question.

For example, we can see in figure 1.3 and 1.4 that it has comprehension with DBMS data table diagram, then query related to given data. After that, each question has separate question and its four options.



- Q 13: In the context of this new relation, a new employee `E1` has joined and has not been allocated a department. Another new employee `E2` has joined, who has been allocated a department (`Dnumber = 5`). Identify the incorrect statement/s about insertion of records of `E1` and `E2` into `emp_dept`.

  Mark:2 **MSQ**

  a) if `Dnumber` is not allowed `NULL` values, then the details of `E1` cannot be inserted into `emp_dept`

  b) if `Dname` is not allowed `NULL` values, then the details of `E1` cannot be inserted into `emp_dept`

  c) There is a chance of inconsistent data for department (`Dnumber = 5`), if there is some incorrect insertion of details of `E2`

  d) Insertion of `E2` is independent and can have no effect on data of the department (`Dnumber = 5`)

  Answer: d)
  Explanation: An Insert Anomaly occurs when certain attributes cannot be inserted into the database without the presence of other attributes, if null values are not allowed like in `Dnumber`, `Dname`. Also while insertion, if the department name for `Dnumber = 5` is not entered correctly for every insertion corresponding to `Dnumber = 5`, then it will lead to inconsistent data.

- Q 14. Suppose employees `A` and `B`, work in department `D1`. Which attribute values will be repeated in the `emp_dept` table? *Mark:2* **MSQ**

  a) `Ssn`

  b) `Dname`

  c) `Address`

  d) `Dmgr_ssn`

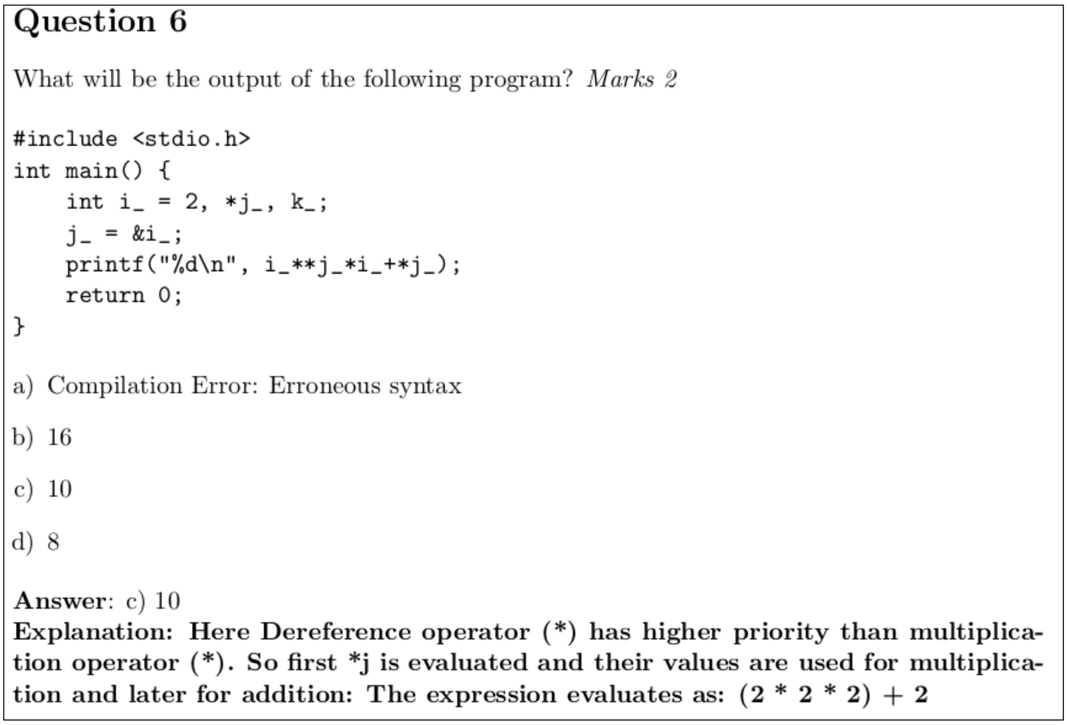
  Answer: b), d)
  Explanation: Dname and Dmgr_ssn are attributes of department table, which is common for `Dnumber`

FIGURE 1.4: DBMS 2nd run, Assignment 4,Ques 12 - 14, Comprehension

1.1 Problem Statement

1.1.1 Introduction

The goal of this work is to make a method for segment Question-Solution documents set in PDF. PDF is converted to image, so mainly it is automated segmentation of Question-Solution Documents(later converted to image file) from NPTEL-NOC. There is NPTEL- NOC assignment upload portal where question along with options



Question 6

What will be the output of the following program? *Marks 2*

```
#include <stdio.h>
int main() {
    int i_ = 2, *j_, k_;
    j_ = &i_;
    printf("%d\n", i_**j_*i_+*j_);
    return 0;
}
```

a) Compilation Error: Erroneous syntax

b) 16

c) 10

d) 8

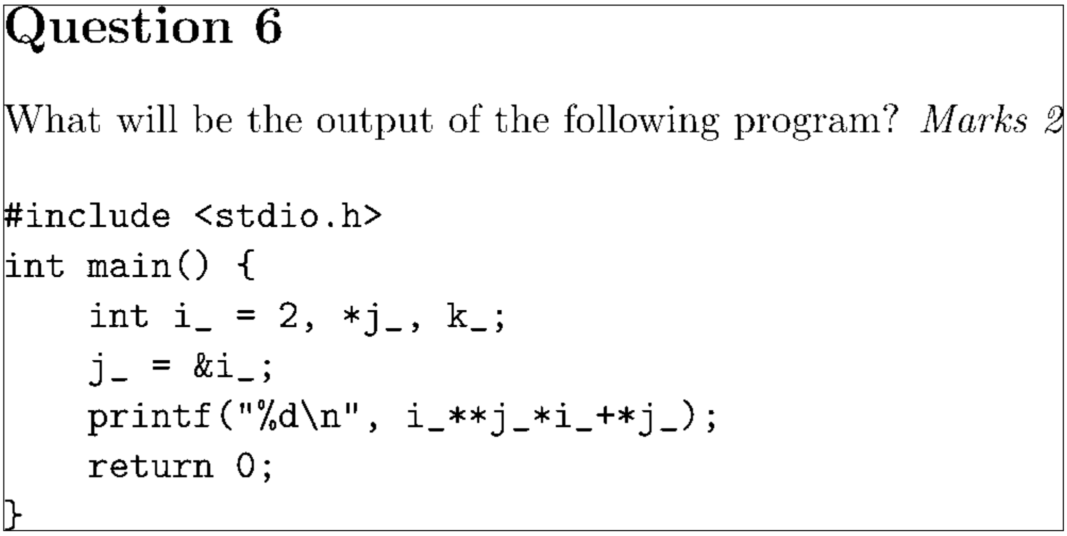
Answer: c) 10
Explanation: Here Dereference operator (*) has higher priority than multiplication operator (*). So first *j is evaluated and their values are used for multiplication and later for addition: The expression evaluates as: (2 * 2 * 2) + 2

FIGURE 1.5: C++ assignment, Week 1 quiz, Ques 6,Input Image

are uploaded for students. Currently, those who upload images have not any knowledge about content of question, options and its solution as they are not professors or Phd students who make question.So, they manually crop images and upload image to NPTEL- NOC portal as in respective question and option portion.

But, it takes lot of expensive time. So, we want to design **Automated Segmentation of Question-Solution Documents from NPTEL-NOC** so that it could be done automatically without using third party.

After segmentation, all question and its option are saved in separate for folder for each question or vary on demand. Following is a sample example where it is shown given input 1.5 and desired output as questions 1.6 , option *a)* 1.7, option *b)* 1.8, option *c)* 1.9 and option *d)* 1.10.

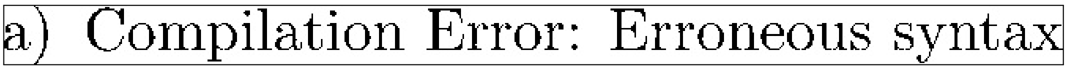


Question 6

What will be the output of the following program? *Marks 2*

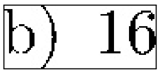
```
#include <stdio.h>
int main() {
    int i_ = 2, *j_, k_;
    j_ = &i_;
    printf("%d\n", i_**j_*i_+*j_);
    return 0;
}
```

FIGURE 1.6: C++ assignment, Week 1 quiz, Ques 6,required segmented question



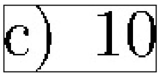
a) Compilation Error: Erroneous syntax

FIGURE 1.7: C++ assignment, Week 1 quiz, Ques 6, required option a)



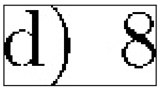
b) 16

FIGURE 1.8: C++ assignment, Week 1 quiz, Ques 6,required option b)



c) 10

FIGURE 1.9: C++ assignment, Week 1 quiz, Ques 6,required option c)



d) 8

FIGURE 1.10: C++ assignment, Week 1 quiz, Ques 6, required option d)

1.1.2 Related Work

1.1.2.1 Line spacing and font size

PDF information on line spacing and font usage gives important clues in bridging the semantic gap between the scanned bitmap page and its fully analysed, block-structured form. Analysis of PDF can yield not only accurate page decomposition but also sufficient document information for the later stages of structural analysis and document understanding.Lovegrove and Brailsford [1995]

1.1.2.2 Common Work

- Scanned document images were broken into different regions and different layers based on the texture of the objects on the pageHaffner et al. [1999], Fan [2003]
- Brailsford et al. have been able to segment a PDF document page into different image and text blocksSmith and Brailsford [1995], Lovegrove and Brailsford [1995]
- Anjewierden has reported his work on recovering the logical structure of the technical manuals in PDF Anjewierden [2001]
- Hadjar et al. have developed a tool for extracting the structures from PDF documents.Hadjar et al. [2004]
- Hui Chao and Jian Fan developed techniques that identified layout and logical components on a PDF document page Chao and Fan [2004]
- T. Hassan and R. baumgartner have proposed a flexible method for detecting and understanding tables in PDF files Hassan and Baumgartner [2007]
- Abhishek at al developed tool for layout-aware text extraction from full-text PDF of scientific articles.Ramakrishnan et al. [2012]
- Henry at al demonstrated pdf document layout study with page elements and bounding boxes, in document layout interpretation and its applications workshop.Chao et al. [2001]

Chapter 2

Dataset Preparation

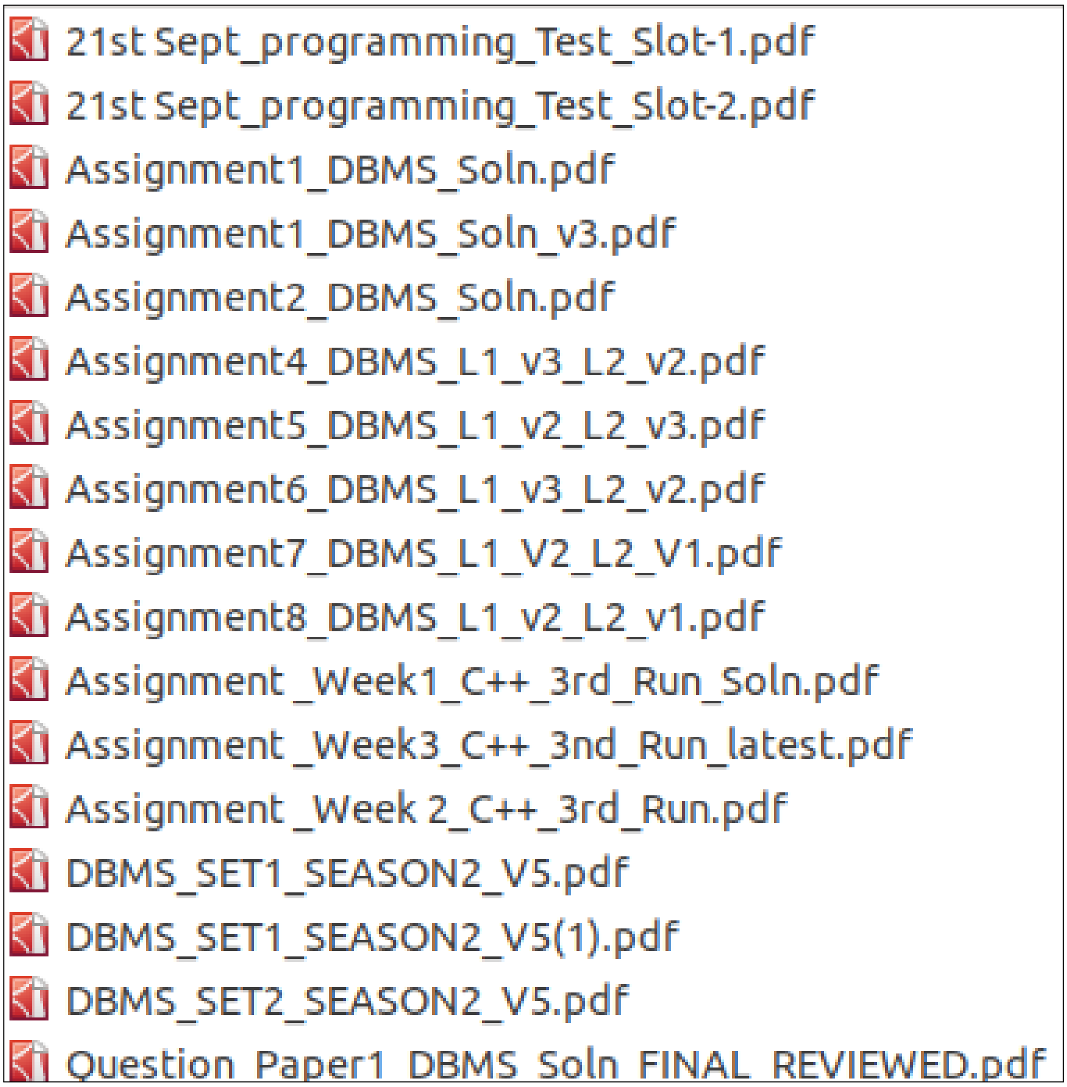
The output in most cases is a region of interest in image converted from pdf.

2.1 Data generated from Latex

Give a brief of the chapter and introduce what you will talk about.

2.1.1 Dataset description

- All the pdf files are generated from latex. So, we have uniform data font size, spacing between sentences
- Suppose there is question number 1 in page 1, font size 20 point and spacing 10 point.
- Then, same uniform font size and spacing would be every where.



21st Sept_programming_Test_Slot-1.pdf

21st Sept_programming_Test_Slot-2.pdf

Assignment1_DBMS_Soln.pdf

Assignment1_DBMS_Soln_v3.pdf

Assignment2_DBMS_Soln.pdf

Assignment4_DBMS_L1_v3_L2_v2.pdf

Assignment5_DBMS_L1_v2_L2_v3.pdf

Assignment6_DBMS_L1_v3_L2_v2.pdf

Assignment7_DBMS_L1_V2_L2_V1.pdf

Assignment8_DBMS_L1_v2_L2_v1.pdf

Assignment _Week1_C++_3rd_Run_Soln.pdf

Assignment _Week3_C++_3nd_Run_latest.pdf

Assignment _Week 2_C++_3rd_Run.pdf

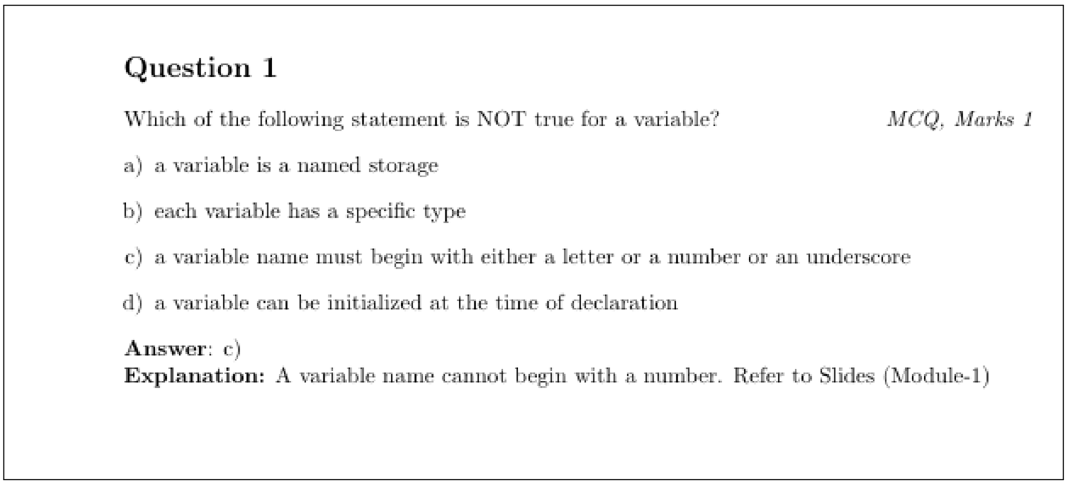
DBMS_SET1_SEASON2_V5.pdf

DBMS_SET1_SEASON2_V5(1).pdf

DBMS_SET2_SEASON2_V5.pdf

Question_Paper1_DBMS_Soln_FINAL_REVIEWED.pdf

Figure 2.1: Pdf Dataset generated from latex



Question 1

Which of the following statement is NOT true for a variable? *MCQ, Marks 1*

a) a variable is a named storage

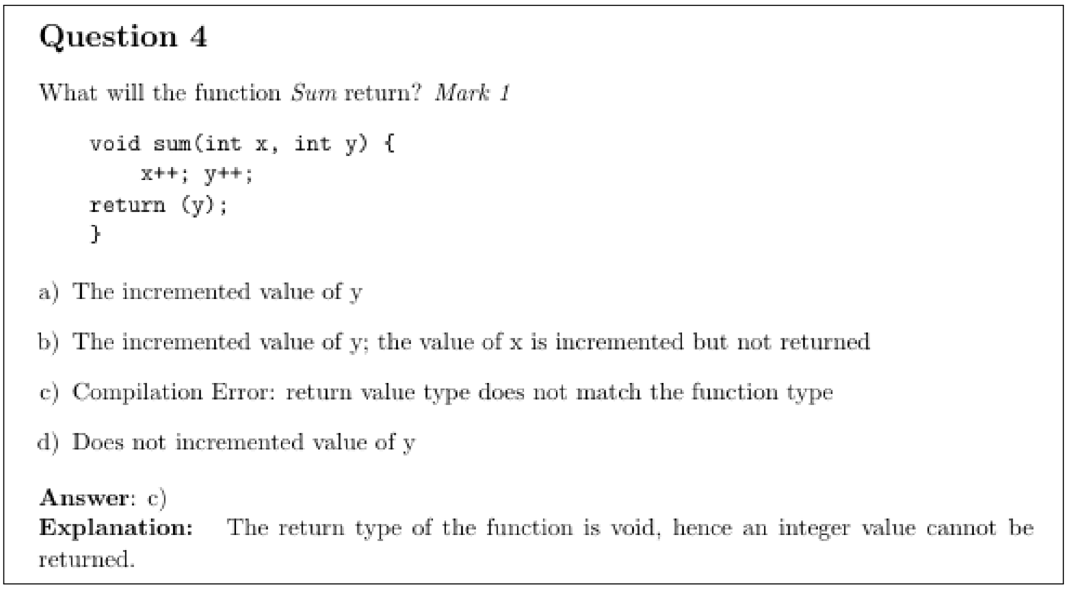
b) each variable has a specific type

c) a variable name must begin with either a letter or a number or an underscore

d) a variable can be initialized at the time of declaration

Answer: c)
Explanation: A variable name cannot begin with a number. Refer to Slides (Module-1)

Figure 2.2: MOOC Question 1



Question 4

What will the function *Sum* return? *Mark 1*

```
    void sum(int x, int y) {
        x++; y++;
    return (y);
    }
```

a) The incremented value of y

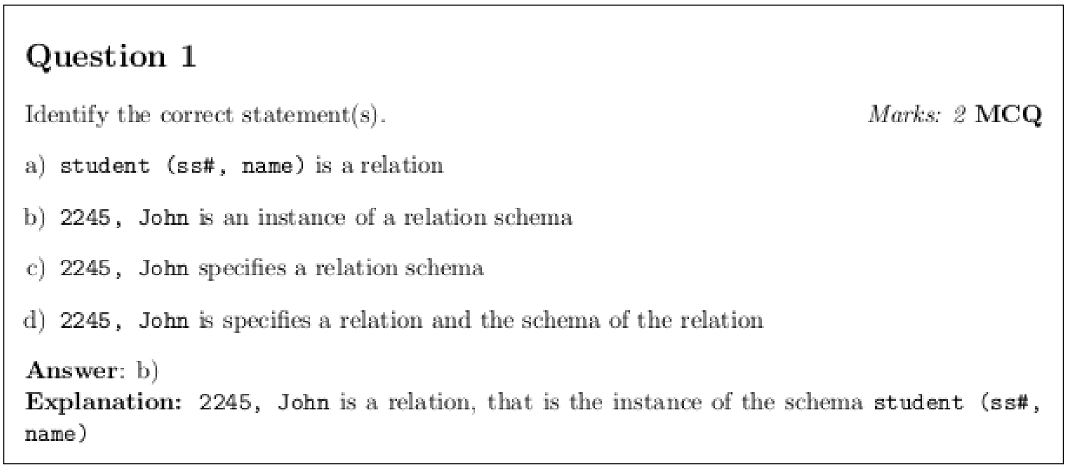
b) The incremented value of y; the value of x is incremented but not returned

c) Compilation Error: return value type does not match the function type

d) Does not incremented value of y

Answer: c)
Explanation: The return type of the function is void, hence an integer value cannot be returned.

FIGURE 2.3: C++ Assignment Week 1, Question 4



Question 1

Identify the correct statement(s). *Marks: 2* **MCQ**

a) `student (ss#, name)` is a relation

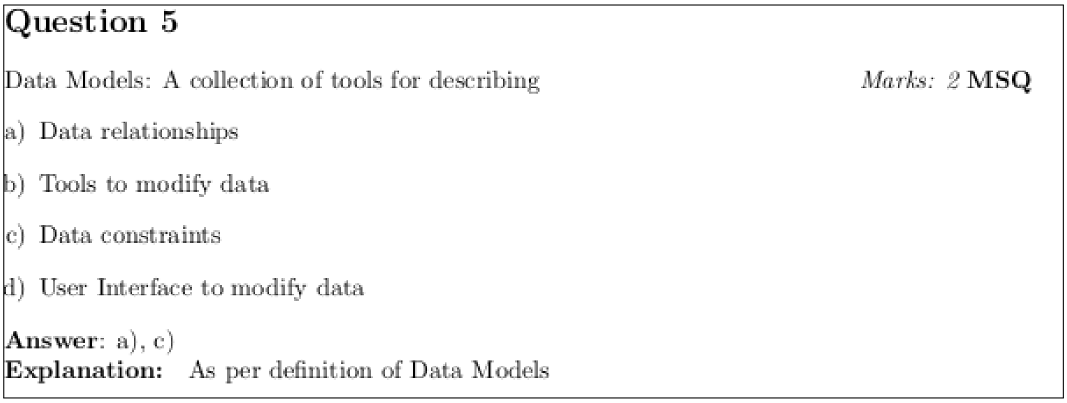
b) `2245, John` is an instance of a relation schema

c) `2245, John` specifies a relation schema

d) `2245, John` is specifies a relation and the schema of the relation

Answer: b)
Explanation: `2245, John` is a relation, that is the instance of the schema `student (ss#, name)`

FIGURE 2.4: Question 1



Question 5

Data Models: A collection of tools for describing *Marks: 2* **MSQ**

a) Data relationships

b) Tools to modify data

c) Data constraints

d) User Interface to modify data

Answer: a), c)
Explanation: As per definition of Data Models

FIGURE 2.5: Question 5

Chapter 3

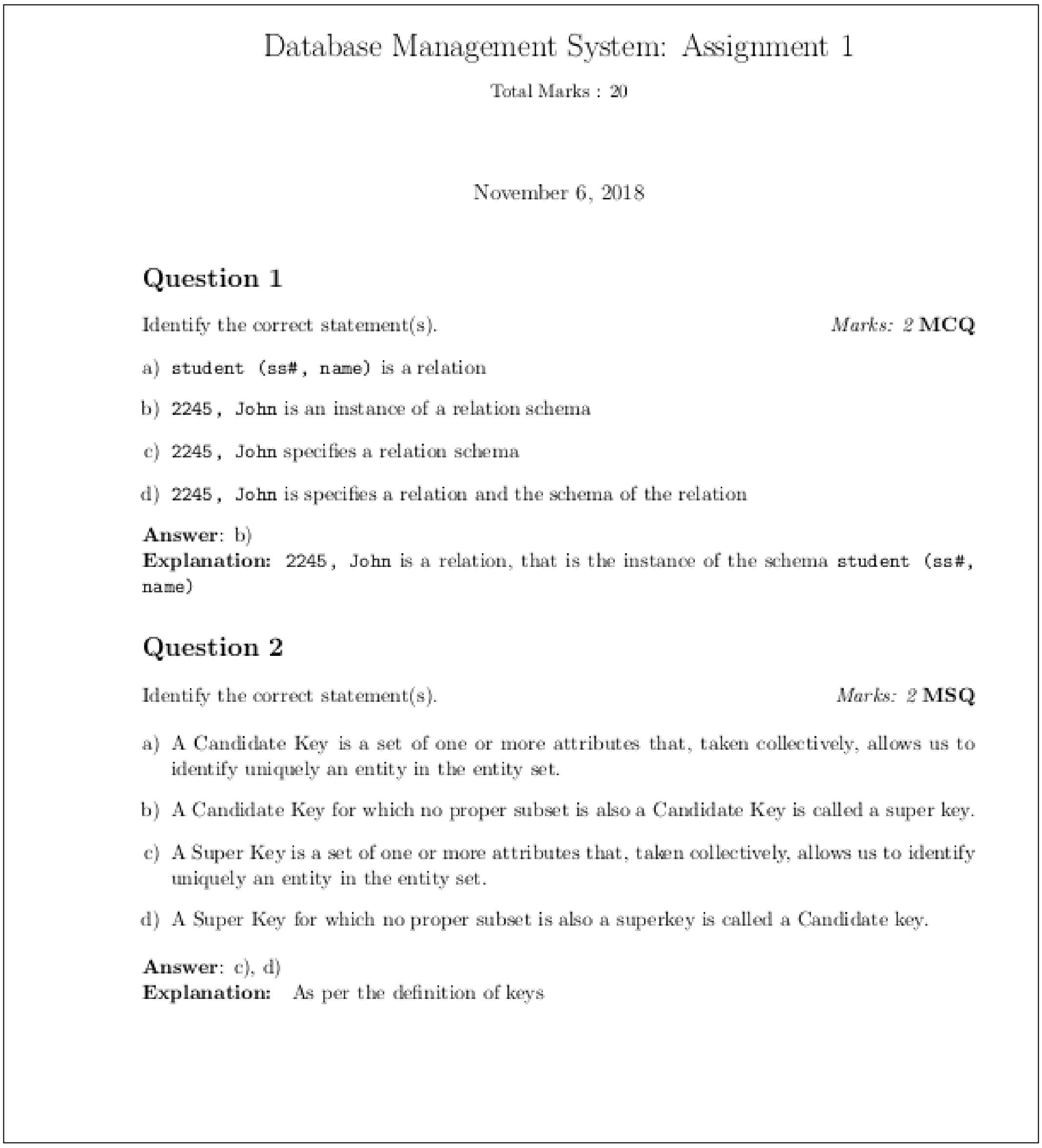
Approaches

3.1 Convert to Excel

3.1.1 Introduction

- It is very simple method. As all pdf are latex generated, so their would be uniformity.
- It gives basic overview of document. There may problem if there are images and tables in pdf file.
- Main motivation was some how extract option “a”, “b”, “c”, “d” as they would be left most and starting point for content of option.

3.1.2 Input



Database Management System: Assignment 1

Total Marks : 20

November 6, 2018

Question 1

Identify the correct statement(s). *Marks: 2* **MCQ**

a) `student (ss#, name)` is a relation

b) `2245, John` is an instance of a relation schema

c) `2245, John` specifies a relation schema

d) `2245, John` is specifies a relation and the schema of the relation

Answer: b)
Explanation: `2245, John` is a relation, that is the instance of the schema `student (ss#, name)`

Question 2

Identify the correct statement(s). *Marks: 2* **MSQ**

a) A Candidate Key is a set of one or more attributes that, taken collectively, allows us to identify uniquely an entity in the entity set.

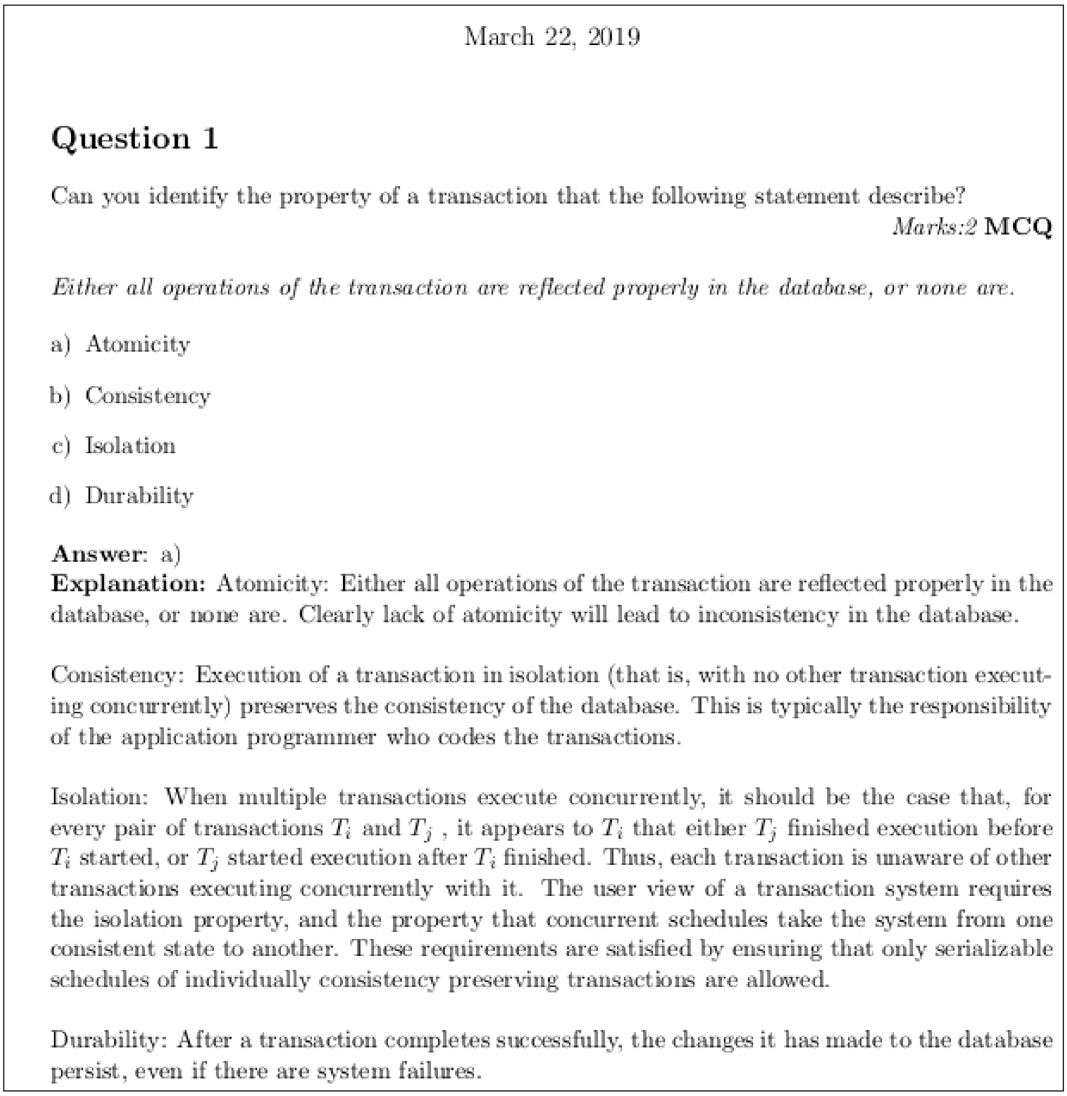
b) A Candidate Key for which no proper subset is also a Candidate Key is called a super key.

c) A Super Key is a set of one or more attributes that, taken collectively, allows us to identify uniquely an entity in the entity set.

d) A Super Key for which no proper subset is also a superkey is called a Candidate key.

Answer: c), d)
Explanation: As per the definition of keys

FIGURE 3.1: Possible Input



March 22, 2019

Question 1

Can you identify the property of a transaction that the following statement describe?

Marks:2 **MCQ**

Either all operations of the transaction are reflected properly in the database, or none are.

a) Atomicity

b) Consistency

c) Isolation

d) Durability

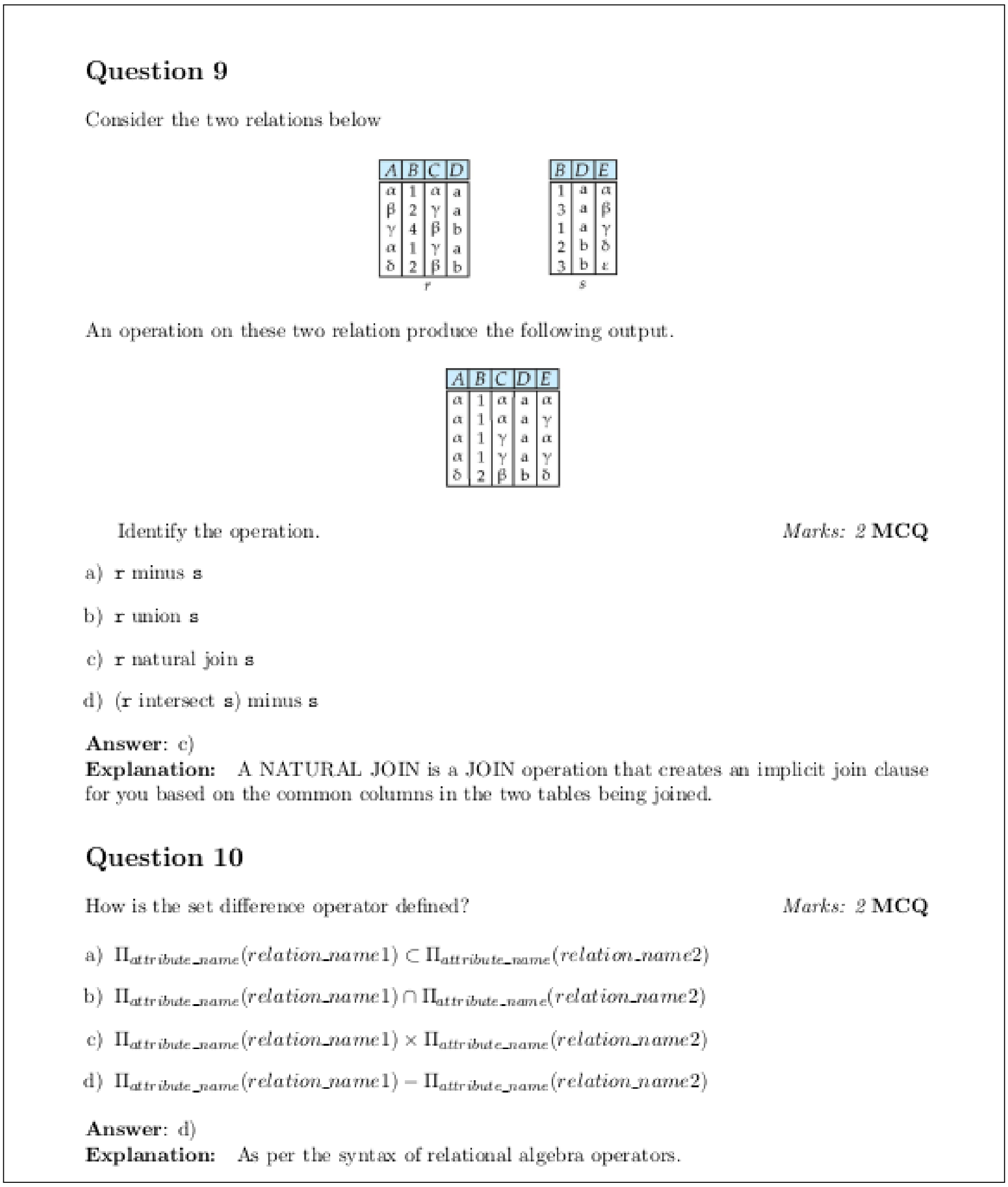
Answer: a)
Explanation: Atomicity: Either all operations of the transaction are reflected properly in the database, or none are. Clearly lack of atomicity will lead to inconsistency in the database.

Consistency: Execution of a transaction in isolation (that is, with no other transaction executing concurrently) preserves the consistency of the database. This is typically the responsibility of the application programmer who codes the transactions.

Isolation: When multiple transactions execute concurrently, it should be the case that, for every pair of transactions T_i and T_j , it appears to T_i that either T_j finished execution before T_i started, or T_j started execution after T_i finished. Thus, each transaction is unaware of other transactions executing concurrently with it. The user view of a transaction system requires the isolation property, and the property that concurrent schedules take the system from one consistent state to another. These requirements are satisfied by ensuring that only serializable schedules of individually consistency preserving transactions are allowed.

Durability: After a transaction completes successfully, the changes it has made to the database persist, even if there are system failures.

FIGURE 3.2: DBMS Run 2, Week 8 Input Snapshot



Question 9

Consider the two relations below

A	B	C	D
α	1	α	a
β	2	γ	a
γ	4	β	b
α	1	γ	a
δ	2	β	b

r

B	D	E
1	a	α
3	a	β
1	a	γ
2	b	δ
3	b	ϵ

s

An operation on these two relation produce the following output.

A	B	C	D	E
α	1	α	a	α
α	1	α	a	γ
α	1	γ	a	α
α	1	γ	a	γ
δ	2	β	b	δ

Identify the operation. *Marks: 2* **MCQ**

a) `r` minus `s`

b) `r` union `s`

c) `r` natural join `s`

d) (`r` intersect `s`) minus `s`

Answer: c)
Explanation: A NATURAL JOIN is a JOIN operation that creates an implicit join clause for you based on the common columns in the two tables being joined.

Question 10

How is the set difference operator defined? *Marks: 2* **MCQ**

a) $\Pi_{attribute\_name}(relation\_name1) \subset \Pi_{attribute\_name}(relation\_name2)$

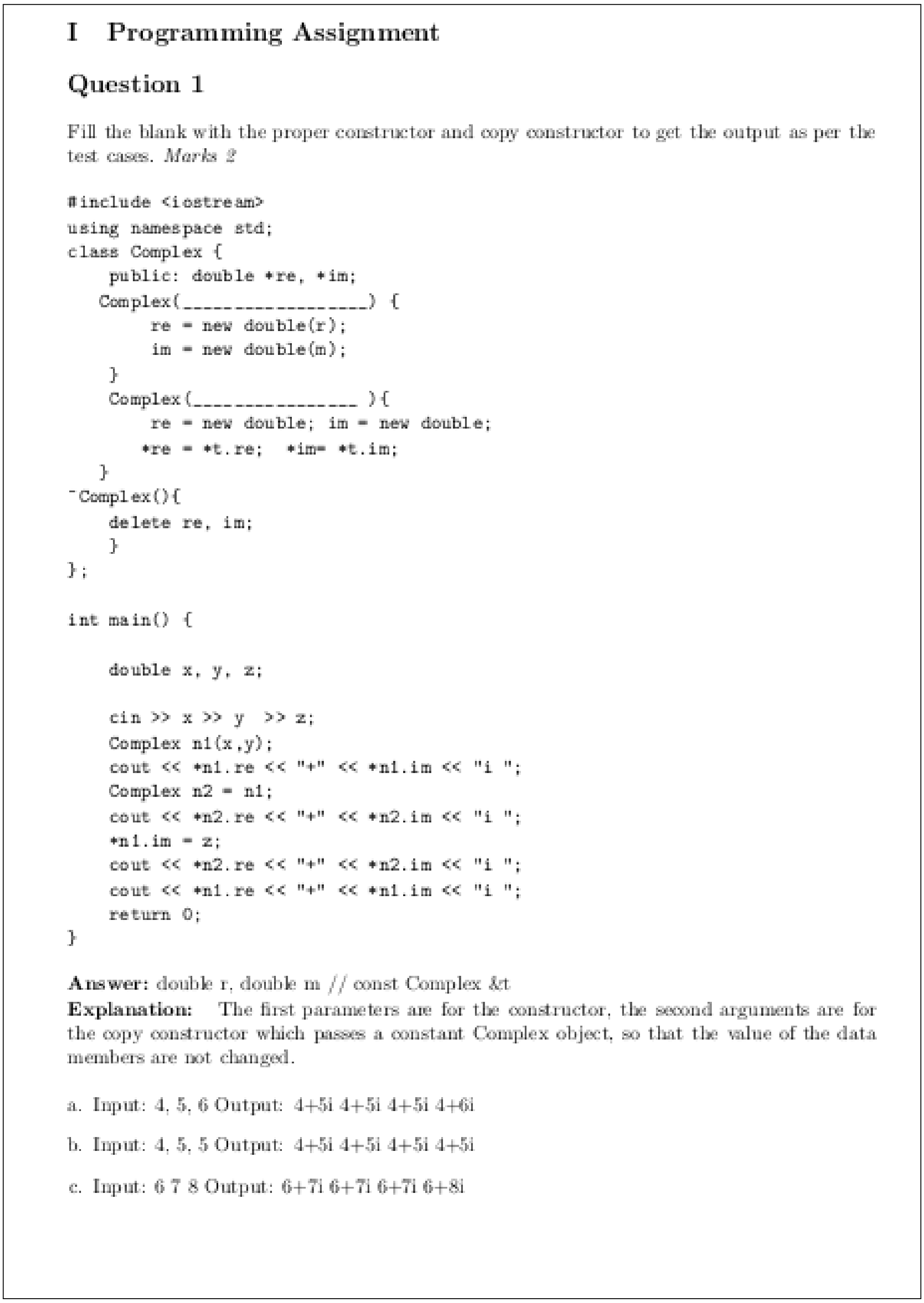
b) $\Pi_{attribute\_name}(relation\_name1) \cap \Pi_{attribute\_name}(relation\_name2)$

c) $\Pi_{attribute\_name}(relation\_name1) \times \Pi_{attribute\_name}(relation\_name2)$

d) $\Pi_{attribute\_name}(relation\_name1) - \Pi_{attribute\_name}(relation\_name2)$

Answer: d)
Explanation: As per the syntax of relational algebra operators.

FIGURE 3.3: Possible Input



I Programming Assignment

Question 1

Fill the blank with the proper constructor and copy constructor to get the output as per the test cases. *Marks 2*

```
#include <iostream>
using namespace std;
class Complex {
    public: double *re, *im;
   Complex(_________________) {
        re = new double(r);
        im = new double(m);
    }
    Complex(________________ ){
        re = new double; im = new double;
       *re = *t.re;  *im= *t.im;
   }
~Complex(){
    delete re, im;
    }
};

int main() {

    double x, y, z;

    cin >> x >> y  >> z;
    Complex n1(x,y);
    cout << *n1.re << "+" << *n1.im << "i ";
    Complex n2 = n1;
    cout << *n2.re << "+" << *n2.im << "i ";
    *n1.im = z;
    cout << *n2.re << "+" << *n2.im << "i ";
    cout << *n1.re << "+" << *n1.im << "i ";
    return 0;
}
```

Answer: double r, double m // const Complex &t
Explanation: The first parameters are for the constructor, the second arguments are for the copy constructor which passes a constant Complex object, so that the value of the data members are not changed.

a. Input: 4, 5, 6 Output: 4+5i 4+5i 4+5i 4+6i

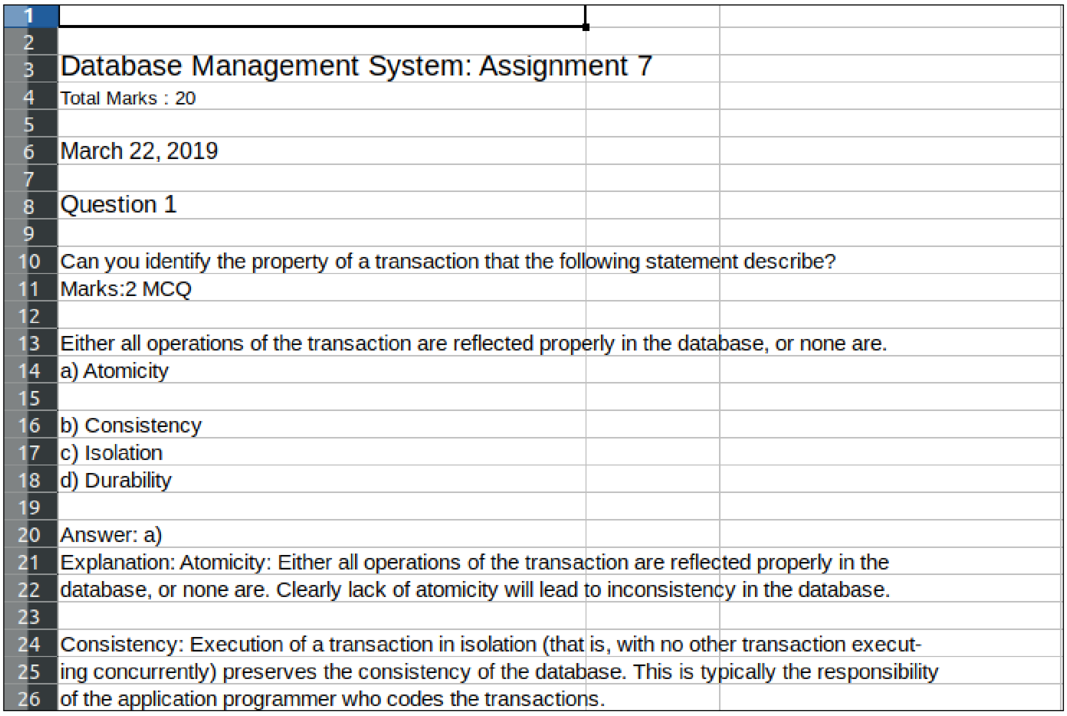
b. Input: 4, 5, 5 Output: 4+5i 4+5i 4+5i 4+5i

c. Input: 6 7 8 Output: 6+7i 6+7i 6+7i 6+8i

Figure 3.4: Possible Input

3.1.3 Result

We get Some idea of formatting of latex generated file.



Database Management System: Assignment 7

Total Marks : 20

March 22, 2019

Question 1

Can you identify the property of a transaction that the following statement describe?

Marks:2 MCQ

Either all operations of the transaction are reflected properly in the database, or none are.

a) Atomicity

b) Consistency

c) Isolation

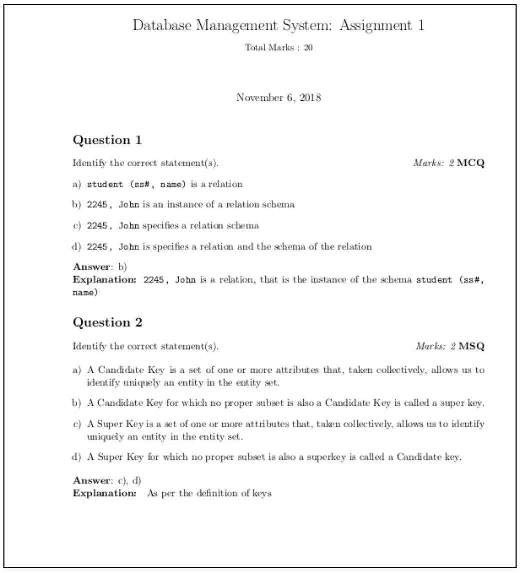
d) Durability

Answer: a)

Explanation: Atomicity: Either all operations of the transaction are reflected properly in the
database, or none are. Clearly lack of atomicity will lead to inconsistency in the database.

Consistency: Execution of a transaction in isolation (that is, with no other transaction execut-
ing concurrently) preserves the consistency of the database. This is typically the responsibility
of the application programmer who codes the transactions.

Figure 3.5: DBMS Run 2, Week 8 Output Excel Snapshot



Database Management System: Assignment 1

Total Marks : 20

November 6, 2018

Question 1

Identify the correct statement(s). *Marks: 2* **MCQ**

a) `student (ss#, name)` is a relation

b) `2245, John` is an instance of a relation schema

c) `2245, John` specifies a relation schema

d) `2245, John` is specifies a relation and the schema of the relation

Answer: b)
Explanation: `2245, John` is a relation, that is the instance of the schema `student (ss#, name)`

Question 2

Identify the correct statement(s). *Marks: 2* **MSQ**

a) A Candidate Key is a set of one or more attributes that, taken collectively, allows us to identify uniquely an entity in the entity set.

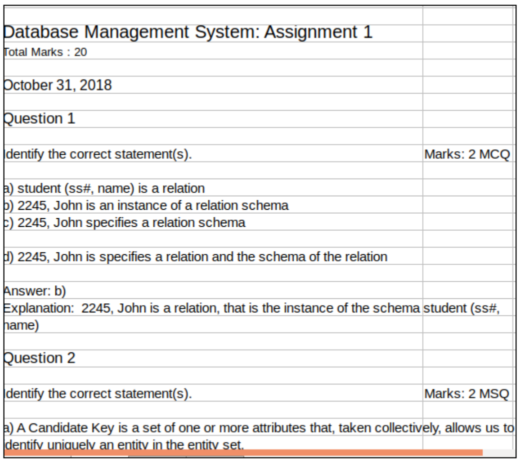
b) A Candidate Key for which no proper subset is also a Candidate Key is called a super key.

c) A Super Key is a set of one or more attributes that, taken collectively, allows us to identify uniquely an entity in the entity set.

d) A Super Key for which no proper subset is also a superkey is called a Candidate key.

Answer: c), d)
Explanation: As per the definition of keys

Figure 3.6: Input Image



Database Management System: Assignment 1

Total Marks : 20

October 31, 2018

Question 1

Identify the correct statement(s). Marks: 2 MCQ

a) student (ss#, name) is a relation
b) 2245, John is an instance of a relation schema
c) 2245, John specifies a relation schema

d) 2245, John is specifies a relation and the schema of the relation

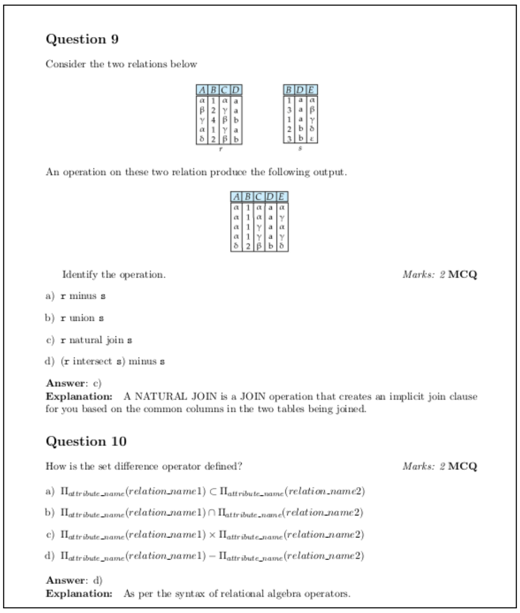
Answer: b)
Explanation: 2245, John is a relation, that is the instance of the schema student (ss#,
name)

Question 2

Identify the correct statement(s). Marks: 2 MSQ

a) A Candidate Key is a set of one or more attributes that, taken collectively, allows us to
identify uniquely an entity in the entity set.

Figure 3.7: Output,Expected



Question 9

Consider the two relations below

A	B	C	D
α	1	α	a
β	2	γ	a
γ	4	β	b
α	1	γ	a
δ	2	β	b

r

B	D	E
1	a	α
3	a	β
1	a	γ
2	b	δ
3	b	ϵ

s

An operation on these two relation produce the following output.

A	B	C	D	E
α	1	α	a	α
α	1	α	a	γ
α	1	γ	a	α
α	1	γ	a	γ
δ	2	β	b	δ

Identify the operation. *Marks: 2* **MCQ**

a) r minus s

b) r union s

c) r natural join s

d) (r intersect s) minus s

Answer: c)
Explanation: A NATURAL JOIN is a JOIN operation that creates an implicit join clause for you based on the common columns in the two tables being joined.

Question 10

How is the set difference operator defined? *Marks: 2* **MCQ**

a) $\Pi_{attribute\_name}(relation\_name1) \subset \Pi_{attribute\_name}(relation\_name2)$

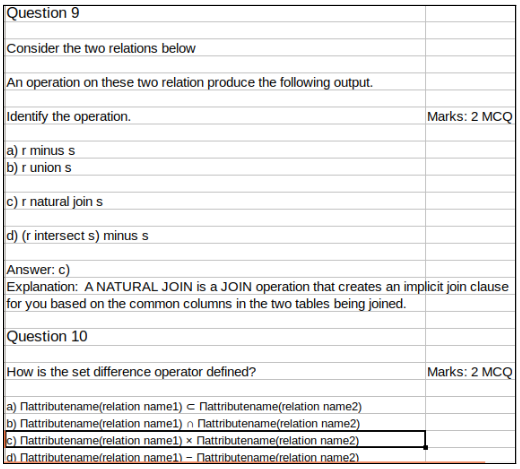
b) $\Pi_{attribute\_name}(relation\_name1) \cap \Pi_{attribute\_name}(relation\_name2)$

c) $\Pi_{attribute\_name}(relation\_name1) \times \Pi_{attribute\_name}(relation\_name2)$

d) $\Pi_{attribute\_name}(relation\_name1) - \Pi_{attribute\_name}(relation\_name2)$

Answer: d)
Explanation: As per the syntax of relational algebra operators.

FIGURE 3.8: Input Image



Question 9	
Consider the two relations below	
An operation on these two relation produce the following output.	
Identify the operation.	Marks: 2 MCQ
a) r minus s	
b) r union s	
c) r natural join s	
d) (r intersect s) minus s	
Answer: c)	
Explanation: A NATURAL JOIN is a JOIN operation that creates an implicit join clause	
for you based on the common columns in the two tables being joined.	
Question 10	
How is the set difference operator defined?	Marks: 2 MCQ
a) Πattributename(relation name1) ⊂ Πattributename(relation name2)	
b) Πattributename(relation name1) ∩ Πattributename(relation name2)	
c) Πattributename(relation name1) × Πattributename(relation name2)	
d) Πattributename(relation name1) − Πattributename(relation name2)	

FIGURE 3.9: Image is missing in output image

3.2 Tesseract hOCR

3.2.1 Introduction

An optical character recognition (OCR) engine

3.2.1.1 OCR

The method of extracting text from images is also called Optical Character Recognition (OCR) or sometimes simply text recognition. Tesseract is an OCR engine with support for unicode and the ability to recognize more than 100 languages out of the box. It can be trained to recognize other languages.

3.2.1.2 Tesseract

Tesseract was developed as a proprietary software by Hewlett Packard Labs. In 2005, it was open sourced by HP in collaboration with the University of Nevada, Las Vegas. Since 2006 it has been actively developed by Google and many open source contributors.

3.2.1.3 Tesseract Deep Leaning

In the past few years, Deep Learning based methods have surpassed traditional machine learning techniques by a huge margin in terms of accuracy in many areas of Computer Vision. Handwriting recognition is one of the prominent examples. So, it was just a matter of time before Tesseract too had a Deep Learning based recognition engine.

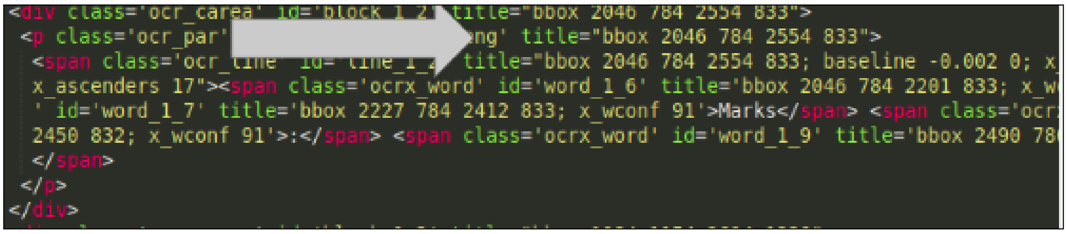
LSTM In currently used version, Tesseract has implemented a Long Short Term Memory (LSTM) based recognition engine. LSTM is a kind of Recurrent Neural Network (RNN).

How Google uses Tesseract OCR Tesseract is used for text detection on mobile devices, in video, and in Gmail image spam detection.

3.2.2 Algorithm

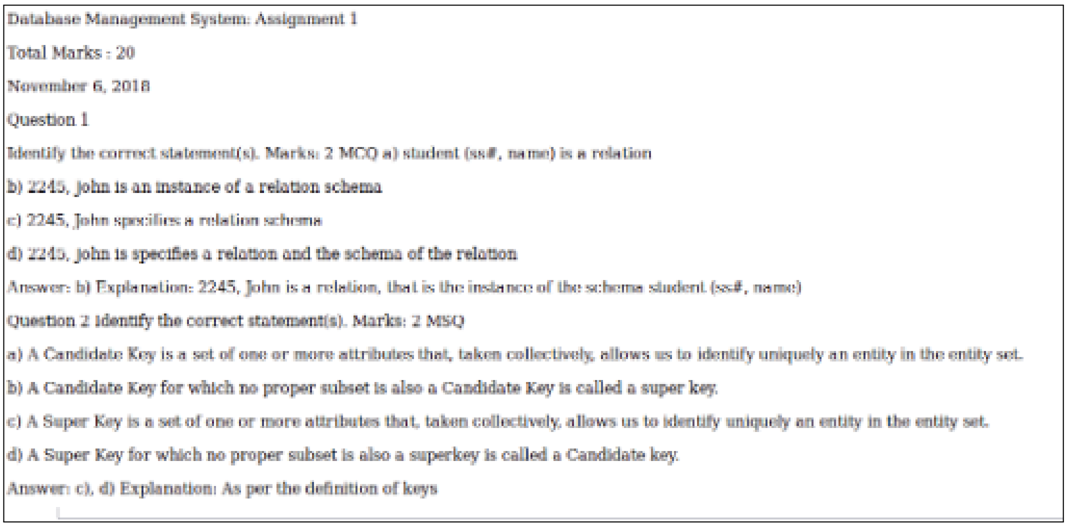
1. Pdf converted to image.
2. Each image is taken as input.
3. It would produce html file where for each line there is co-ordinates which would be bounding.

3.2.3 Result



```
<div class='ocr_carea' id='block_1_2  title="bbox 2046 784 2554 833">
 <p class='ocr_par'                    ng' title="bbox 2046 784 2554 833">
  <span class='ocr_line' id='line_1_2  title="bbox 2046 784 2554 833; baseline -0.002 0; x_
  x_ascenders 17"><span class='ocrx_word' id='word_1_6' title='bbox 2046 784 2201 833; x_w
  ' id='word_1_7' title='bbox 2227 784 2412 833; x_wconf 91'>Marks</span> <span class='ocr
  2450 832; x_wconf 91'>:</span> <span class='ocrx_word' id='word_1_9' title='bbox 2490 78
  </span>
 </p>
</div>
```

Figure 3.10: Bounding box co-ordinates



Database Management System: Assignment 1

Total Marks : 20

November 6, 2018

Question 1

Identify the correct statement(s). Marks: 2 MCQ a) student (ss#, name) is a relation

b) 2245, John is an instance of a relation schema

c) 2245, John specifies a relation schema

d) 2245, John is specifies a relation and the schema of the relation

Answer: b) Explanation: 2245, John is a relation, that is the instance of the schema student (ss#, name)

Question 2 Identify the correct statement(s). Marks: 2 MSQ

a) A Candidate Key is a set of one or more attributes that, taken collectively, allows us to identify uniquely an entity in the entity set.

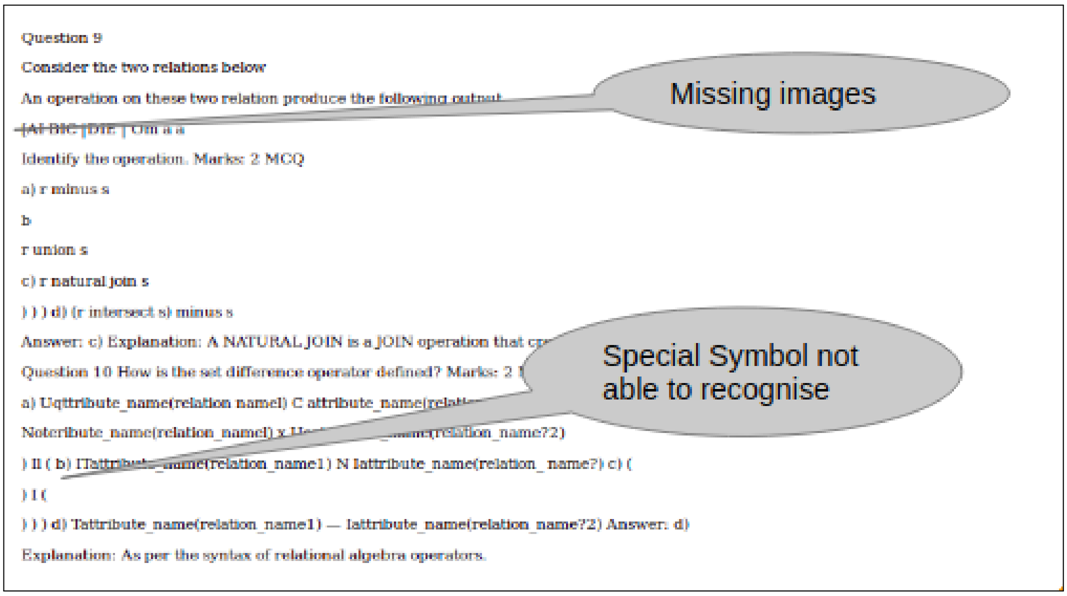
b) A Candidate Key for which no proper subset is also a Candidate Key is called a super key.

c) A Super Key is a set of one or more attributes that, taken collectively, allows us to identify uniquely an entity in the entity set.

d) A Super Key for which no proper subset is also a superkey is called a Candidate key.

Answer: c), d) Explanation: As per the definition of keys

Figure 3.11: Parse in browser



Question 9

Consider the two relations below

An operation on these two relation produce the following output

|A| BIC |DIE | Om a a

Identify the operation. Marks: 2 MCQ

a) r minus s

b

r union s

c) r natural join s

))) d) (r intersect s) minus s

Answer: c) Explanation: A NATURAL JOIN is a JOIN operation that cr

Question 10 How is the set difference operator defined? Marks: 2

a) Uqttribute_name(relation namel) C attribute_name(relati

Noteribute_name(relation_namel) x H name(relation_name?2)

) Il (b) ITattribute_name(relation_name1) N lattribute_name(relation_ name?) c) (

) I (

))) d) Tattribute_name(relation_name1) — lattribute_name(relation_name?2) Answer: d)

Explanation: As per the syntax of relational algebra operators.

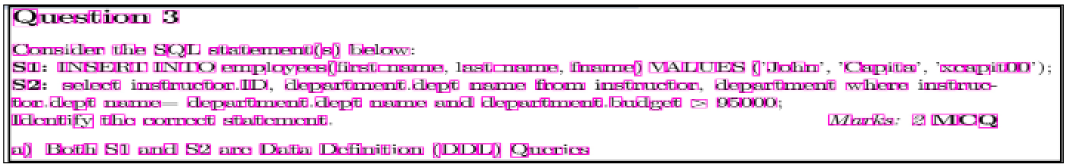
Figure 3.12: When there are images and special symbol in images. It could not identify it.

3.3 Bounding Box

3.3.1 Algorithm

1. Convert pdf to image
2. Applied threshold (simple binary threshold,with handpicked value of 150)
3. Applied dilation to thicken lines in image, leading to more compact object and less white space fragments.
4. Identified contours of objects in resulted image using opencv findContours function.‘
5. Draw a bounding box (rectangle) circumscribing each contoured object.

3.3.2 Result

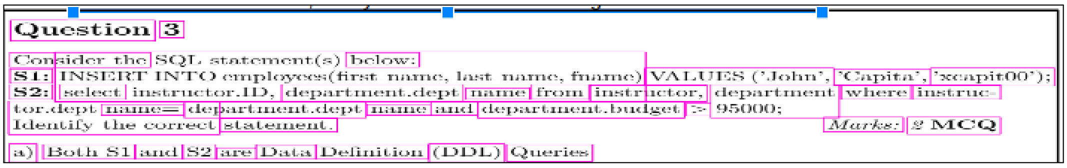


Question 3

Consider the SQL statement(s) below:
S1: INSERT INTO employees(first name, last name, fname) VALUES ('John', 'Capita', 'xcapit00');
S2: select instructor.ID, department.dept name from instructor, department where instructor.dept name= department.dept name and department.budget > 95000;
Identify the correct statement. *Marks:* 2 MCQ

a) Both S1 and S2 are Data Definition (DDL) Queries

FIGURE 3.13: For dilution iteration = 1, Every character has bounding box



Question 3

Consider the SQL statement(s) below:
S1: INSERT INTO employees(first name, last name, fname) VALUES ('John', 'Capita', 'xcapit00');
S2: select instructor.ID, department.dept name from instructor, department where instructor.dept name= department.dept name and department.budget > 95000;
Identify the correct statement. *Marks:* 2 MCQ

a) Both S1 and S2 are Data Definition (DDL) Queries

FIGURE 3.14: For dilution iteration = 30, Some lines are missing in bounding box

3.4 Space Segmentation

3.4.1 Introduction

As pdf is latex generated, So document is uniform. Font size and spacing between Question, its option and explanation would be uniform.

3.4.2 Algorithm

1. Convert pdf to image
2. As image is 2D matrix of 0 and 225.
3. Scan document from top to bottom and left to right to find all lines of text in the document.
4. While scanning note down the starting and ending pixels in vertical and horizontal direction.

5. Spacing = nextlineStart - CurrentlineEnd

6. Fontsize = CurrentlineEnd - CurrentlineStart + 1

7. If (spacing LESSTHAN fontsize): Consider in paragraph

8. Crop required region as we get all coordinates.

3.4.3 Result

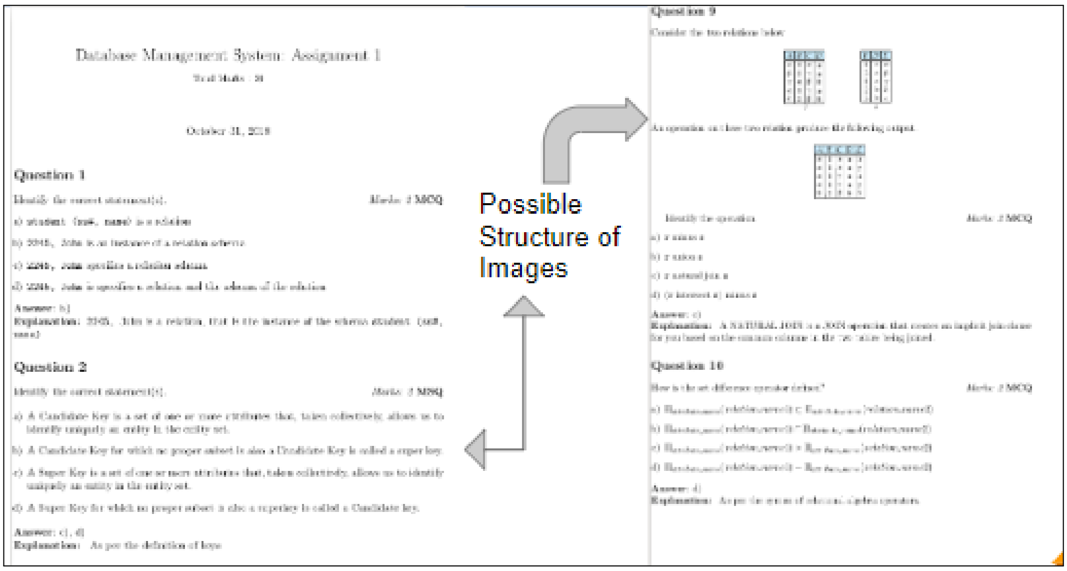
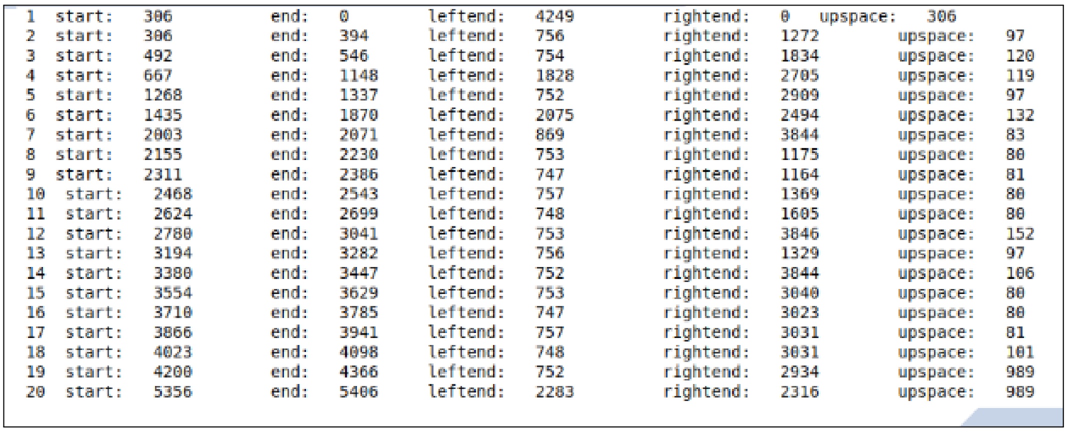
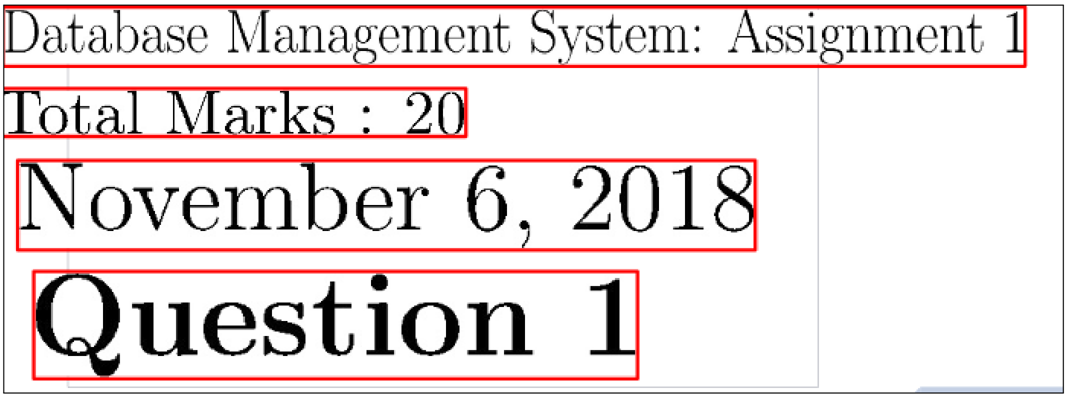

Database Management System: Assignment 1
October 31, 2018
Question 1
Question 2
Possible Structure of Images
Question 9
Question 10

Figure 3.15: Possible Structure of images



```
1  start:   306          end:   0      leftend:   4249        rightend:   0   upspace:   306
2  start:   306          end:   394    leftend:   756         rightend:   1272       upspace:   97
3  start:   492          end:   546    leftend:   754         rightend:   1834       upspace:   120
4  start:   667          end:   1148   leftend:   1828        rightend:   2705       upspace:   119
5  start:   1268         end:   1337   leftend:   752         rightend:   2909       upspace:   97
6  start:   1435         end:   1870   leftend:   2075        rightend:   2494       upspace:   132
7  start:   2003         end:   2071   leftend:   869         rightend:   3844       upspace:   83
8  start:   2155         end:   2230   leftend:   753         rightend:   1175       upspace:   80
9  start:   2311         end:   2386   leftend:   747         rightend:   1164       upspace:   81
10  start:   2468        end:   2543   leftend:   757         rightend:   1369       upspace:   80
11  start:   2624        end:   2699   leftend:   748         rightend:   1605       upspace:   80
12  start:   2780        end:   3041   leftend:   753         rightend:   3846       upspace:   152
13  start:   3194        end:   3282   leftend:   756         rightend:   1329       upspace:   97
14  start:   3380        end:   3447   leftend:   752         rightend:   3844       upspace:   106
15  start:   3554        end:   3629   leftend:   753         rightend:   3040       upspace:   80
16  start:   3710        end:   3785   leftend:   747         rightend:   3023       upspace:   80
17  start:   3866        end:   3941   leftend:   757         rightend:   3031       upspace:   81
18  start:   4023        end:   4098   leftend:   748         rightend:   3031       upspace:   101
19  start:   4200        end:   4366   leftend:   752         rightend:   2934       upspace:   989
20  start:   5356        end:   5406   leftend:   2283        rightend:   2316       upspace:   989
```

Figure 3.16: Coordinates for croping image,start, end, leftend and rightend



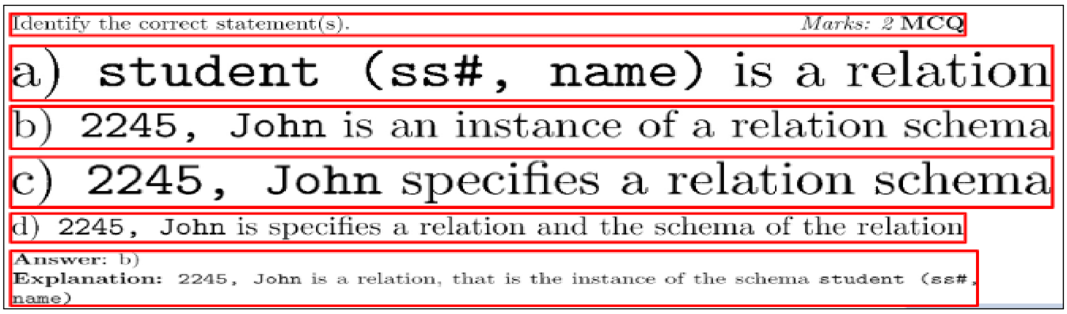
Database Management System: Assignment 1

Total Marks : 20

November 6, 2018

Question 1

FIGURE 3.17: Normal cases, as expected output.



Identify the correct statement(s). *Marks: 2* **MCQ**

a) student (ss#, name) is a relation

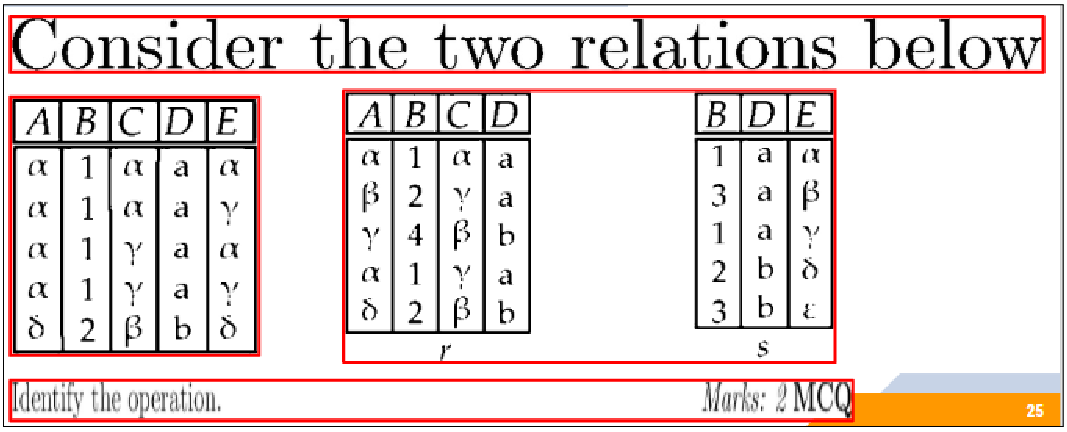
b) 2245, John is an instance of a relation schema

c) 2245, John specifies a relation schema

d) 2245, John is specifies a relation and the schema of the relation

Answer: b)
Explanation: 2245, John is a relation, that is the instance of the schema student (ss#, name)

FIGURE 3.18: Normal cases, as expected output



Consider the two relations below

A	*B*	*C*	*D*	*E*
α	1	α	a	α
α	1	α	a	γ
α	1	γ	a	α
α	1	γ	a	γ
δ	2	β	b	δ

A	*B*	*C*	*D*
α	1	α	a
β	2	γ	a
γ	4	β	b
α	1	γ	a
δ	2	β	b

r

B	*D*	*E*
1	a	α
3	a	β
1	a	γ
2	b	δ
3	b	ε

s

Identify the operation. *Marks: 2* MCQ

FIGURE 3.19: Four figure are separated as separate block, But they are required within same block

3.5 Space Segmentation and Using OCR to recognize option - a) b) c) d)

3.5.1 Introduction

Pdf file is latex generated. It is converted to image file for reading in open cv and passing to OCR. OCR give approximate text for image but our main goal is only segmentation. So, Option " a) b) c) d)" and "Answer" are only useful for us in OCR output.

3.5.2 Algorithm

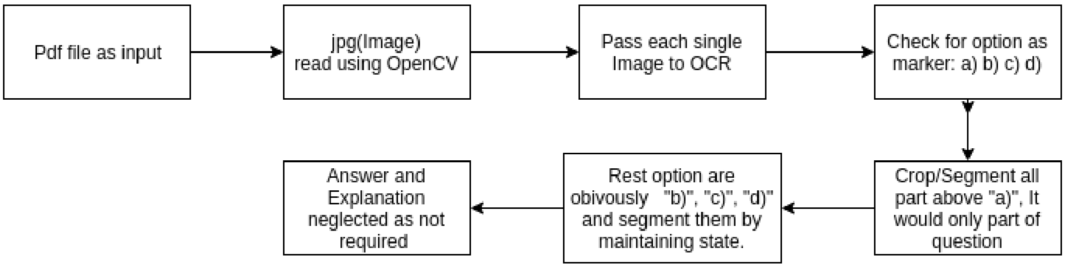
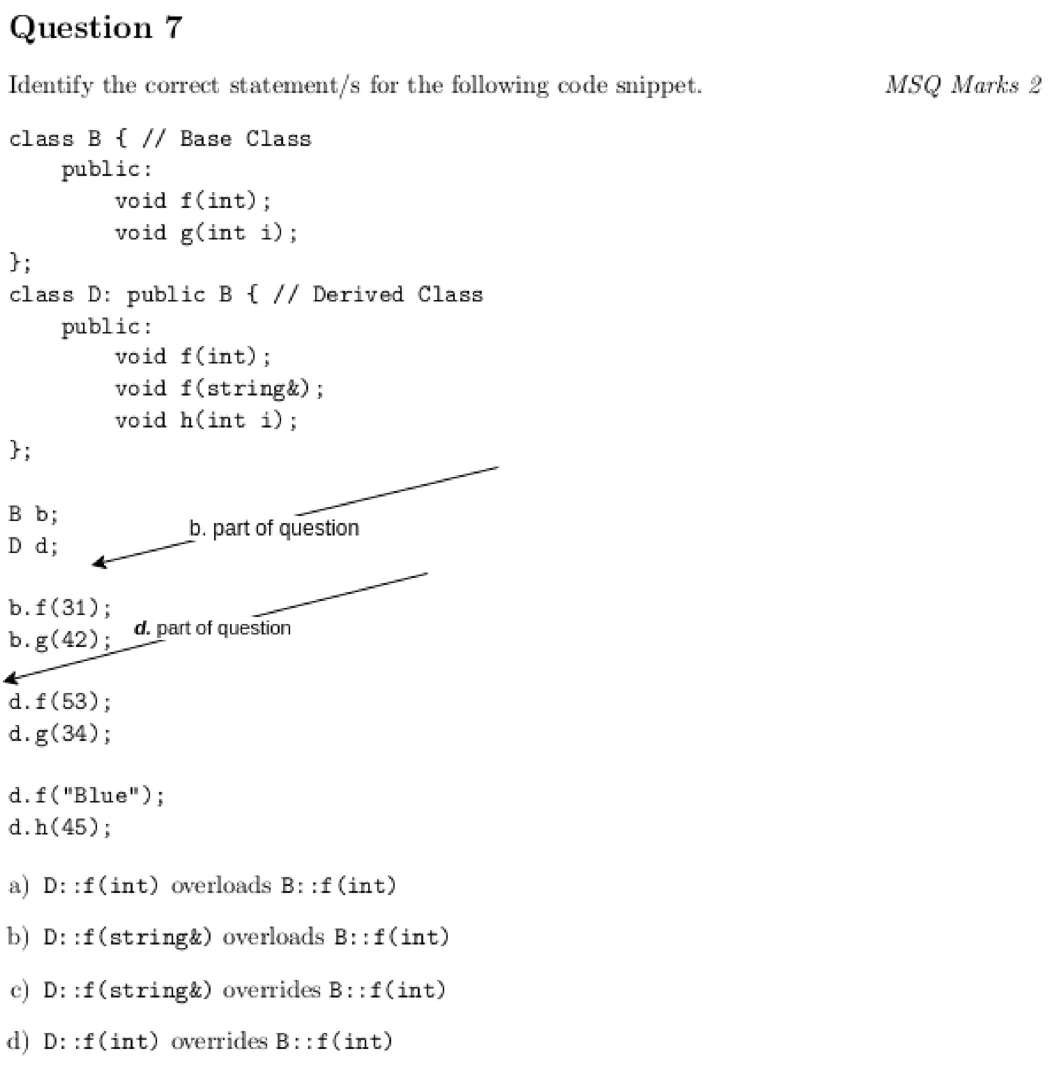

Pdf file as input
jpg(Image) read using OpenCV
Pass each single Image to OCR
Check for option as marker: a) b) c) d)
Crop/Segment all part above "a)", It would only part of question
Rest option are obivously "b)", "c)", "d)" and segment them by maintaining state.
Answer and Explanation neglected as not required

Figure 3.20: Overview flow diagram

While using this there are following restriction.

1. All question should start from new page.
2. All question have only four options.
3. Some question are multiple pages which are not handled, as it is assumed that each page start with new question.
4. We cannot use "a." "b." "c." "d." as option because some question are have "b.text" as part of question and options as shown in figure 3.21



Question 7

Identify the correct statement/s for the following code snippet. *MSQ Marks 2*

```
class B { // Base Class
    public:
        void f(int);
        void g(int i);
};
class D: public B { // Derived Class
    public:
        void f(int);
        void f(string&);
        void h(int i);
};

B b;
D d;

b.f(31);
b.g(42);

d.f(53);
d.g(34);

d.f("Blue");
d.h(45);
```

a) `D::f(int)` overloads `B::f(int)`

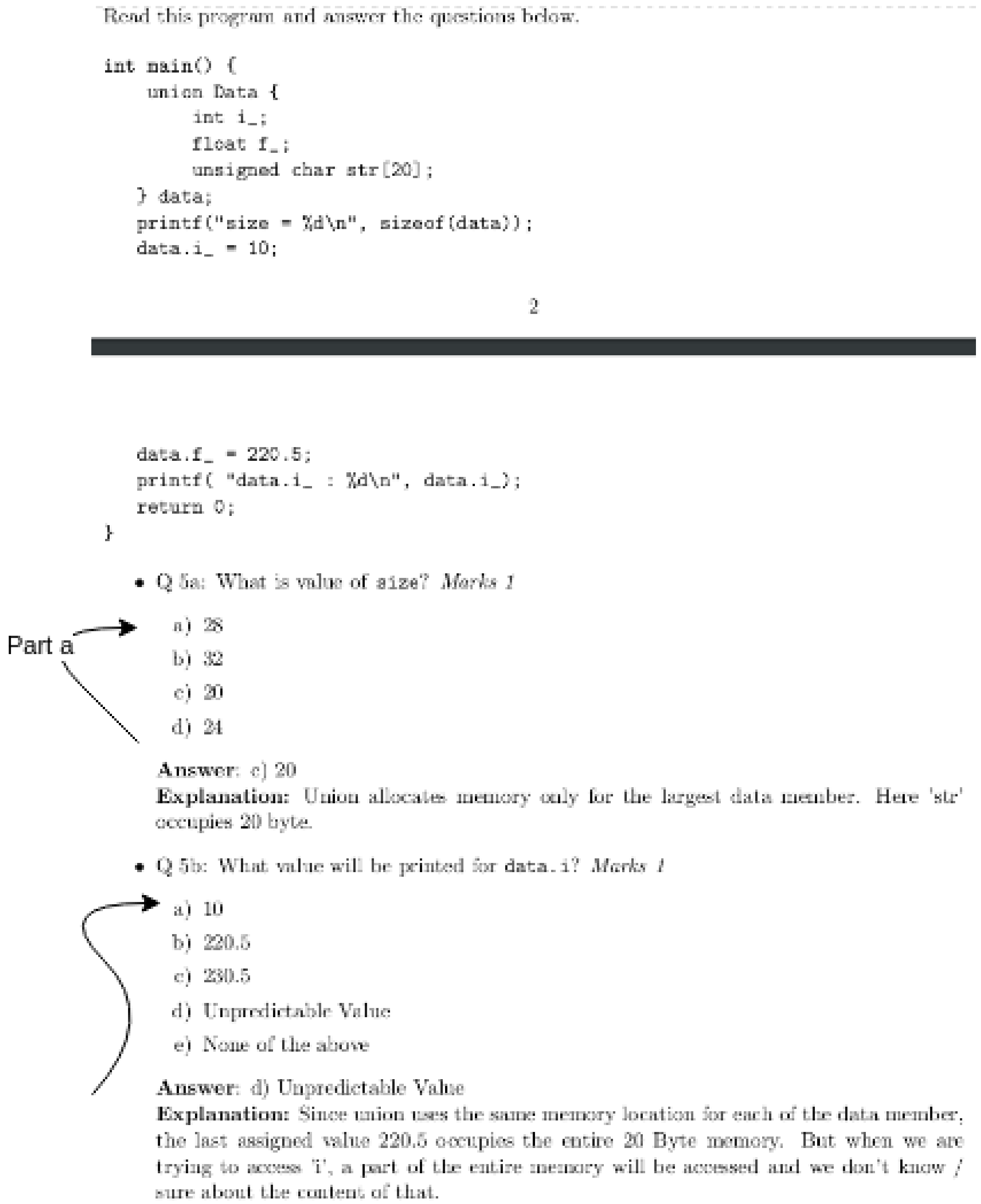
b) `D::f(string&)` overloads `B::f(int)`

c) `D::f(string&)` overrides `B::f(int)`

d) `D::f(int)` overrides `B::f(int)`

FIGURE 3.21: MOOC Assignment 5, Question 7 has ”b.” as part of question.

5. There should not be separate bullet and numbering or layout for Question as part in comprehension. Example is question **5a** and **5b** in figure 3.34.



Read this program and answer the questions below.

```
int main() {
    union Data {
        int i_;
        float f_;
        unsigned char str[20];
   } data;
   printf("size = %d\n", sizeof(data));
   data.i_ = 10;
```

2

```
   data.f_ = 220.5;
   printf( "data.i_ : %d\n", data.i_);
   return 0;
}
```

- Q 5a: What is value of size? *Marks 1*

  a) 28

  b) 32

  c) 20

  d) 24

  Answer: c) 20
  Explanation: Union allocates memory only for the largest data member. Here 'str' occupies 20 byte.

- Q 5b: What value will be printed for data.i? *Marks 1*

  a) 10

  b) 220.5

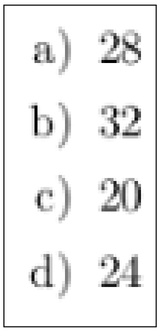
  c) 230.5

  d) Unpredictable Value

  e) None of the above

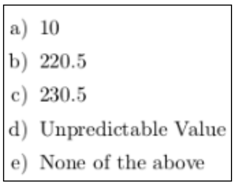
  Answer: d) Unpredictable Value
  Explanation: Since union uses the same memory location for each of the data member, the last assigned value 220.5 occupies the entire 20 Byte memory. But when we are trying to access 'i', a part of the entire memory will be accessed and we don't know / sure about the content of that.

FIGURE 3.22: Comprehension Question, Layout for question changes.



a) 28
b) 32
c) 20
d) 24

Figure 3.23: Ques 5a, Option as single entity due to layout problem

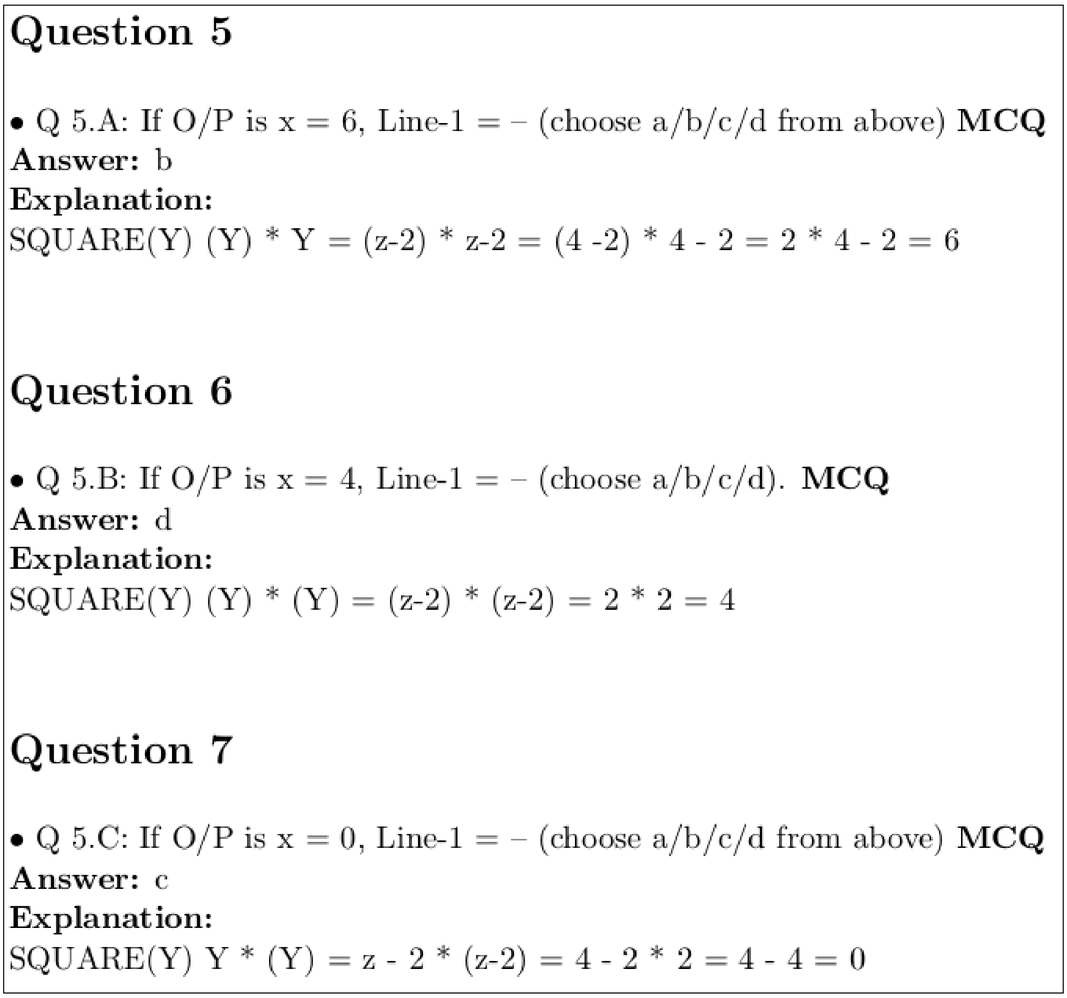


a) 10
b) 220.5
c) 230.5
d) Unpredictable Value
e) None of the above

Figure 3.24: Ques 5b, Option as single entity due to layout problem

Due to this layout change, options are not segmented as separate entity as in figure 3.23 and 3.24.

6. Some question are have default option, But we have to add in option in these so that OCR could recognize them and thus we could segment them as in figure 3.25



Question 5

- Q 5.A: If O/P is x = 6, Line-1 = – (choose a/b/c/d from above) **MCQ**

Answer: b
Explanation:
SQUARE(Y) (Y) * Y = (z-2) * z-2 = (4 -2) * 4 - 2 = 2 * 4 - 2 = 6

Question 6

- Q 5.B: If O/P is x = 4, Line-1 = – (choose a/b/c/d). **MCQ**

Answer: d
Explanation:
SQUARE(Y) (Y) * (Y) = (z-2) * (z-2) = 2 * 2 = 4

Question 7

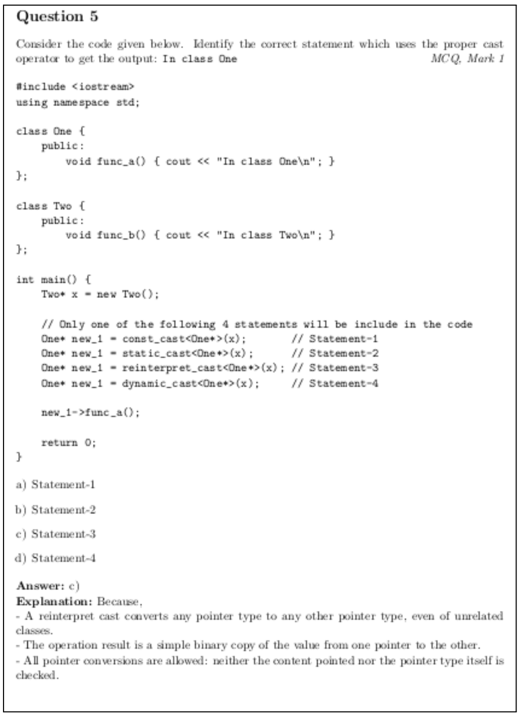
- Q 5.C: If O/P is x = 0, Line-1 = – (choose a/b/c/d from above) **MCQ**

Answer: c
Explanation:
SQUARE(Y) Y * (Y) = z - 2 * (z-2) = 4 - 2 * 2 = 4 - 4 = 0

FIGURE 3.25: Default Non Optional Question

3.5.3 Input and Result

In figure 3.26 input image is given and after segmentation its question in figure 3.27 and its correct option are in figure 3.28, 3.29, 3.30, 3.31.



Question 5

Consider the code given below. Identify the correct statement which uses the proper cast operator to get the output: `In class One` *MCQ, Mark 1*

```
#include <iostream>
using namespace std;

class One {
    public:
        void func_a() { cout << "In class One\n"; }
};

class Two {
    public:
        void func_b() { cout << "In class Two\n"; }
};

int main() {
    Two* x = new Two();

    // Only one of the following 4 statements will be include in the code
    One* new_1 = const_cast<One*>(x);       // Statement-1
    One* new_1 = static_cast<One*>(x);      // Statement-2
    One* new_1 = reinterpret_cast<One*>(x); // Statement-3
    One* new_1 = dynamic_cast<One*>(x);     // Statement-4

    new_1->func_a();

    return 0;
}
```

a) Statement-1

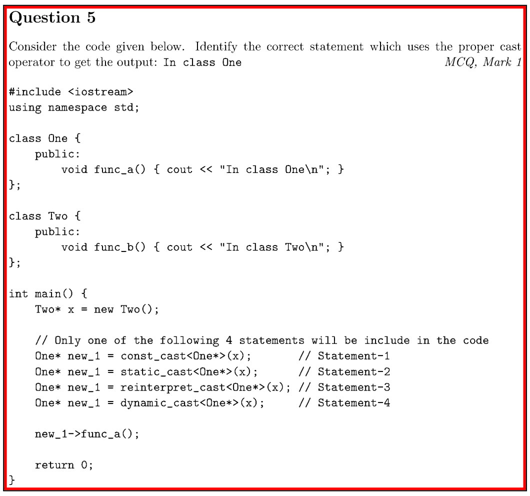
b) Statement-2

c) Statement-3

d) Statement-4

Answer: c)
Explanation: Because,
- A reinterpret cast converts any pointer type to any other pointer type, even of unrelated classes.
- The operation result is a simple binary copy of the value from one pointer to the other.
- All pointer conversions are allowed: neither the content pointed nor the pointer type itself is checked.

FIGURE 3.26: MOOC Assignment 7, Input question 2

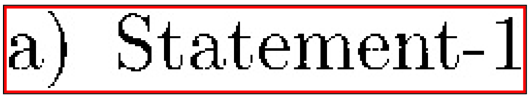


Question 5

Consider the code given below. Identify the correct statement which uses the proper cast operator to get the output: `In class One` *MCQ, Mark 1*

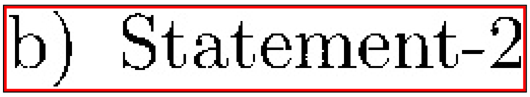
```
#include <iostream>
using namespace std;

class One {
    public:
        void func_a() { cout << "In class One\n"; }
};

class Two {
    public:
        void func_b() { cout << "In class Two\n"; }
};

int main() {
    Two* x = new Two();

    // Only one of the following 4 statements will be include in the code
    One* new_1 = const_cast<One*>(x);       // Statement-1
    One* new_1 = static_cast<One*>(x);      // Statement-2
    One* new_1 = reinterpret_cast<One*>(x); // Statement-3
    One* new_1 = dynamic_cast<One*>(x);     // Statement-4

    new_1->func_a();

    return 0;
}
```

FIGURE 3.27: Segmented Question 5



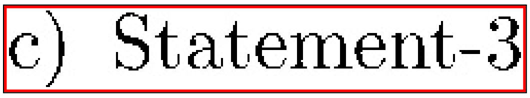
a) Statement-1

FIGURE 3.28: Question 5, Segmented Option a



b) Statement-2

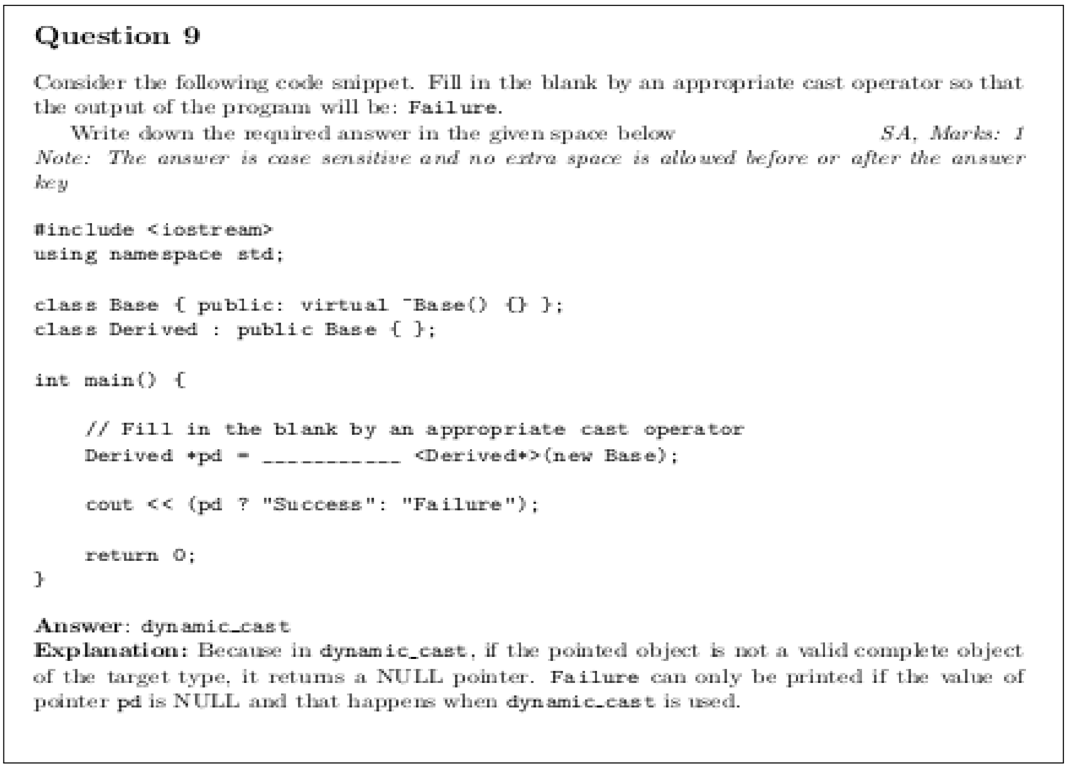
FIGURE 3.29: Question 5,Segmented Option b



c) Statement-3

FIGURE 3.30: Question 5,Segmented Option c

There is *Fill in the blanks* question where segmentation is only for question. There should not be any option related to it as shown in figure 3.32



Question 9

Consider the following code snippet. Fill in the blank by an appropriate cast operator so that the output of the program will be: `Failure`.

Write down the required answer in the given space below *SA, Marks: 1*

Note: The answer is case sensitive and no extra space is allowed before or after the answer key

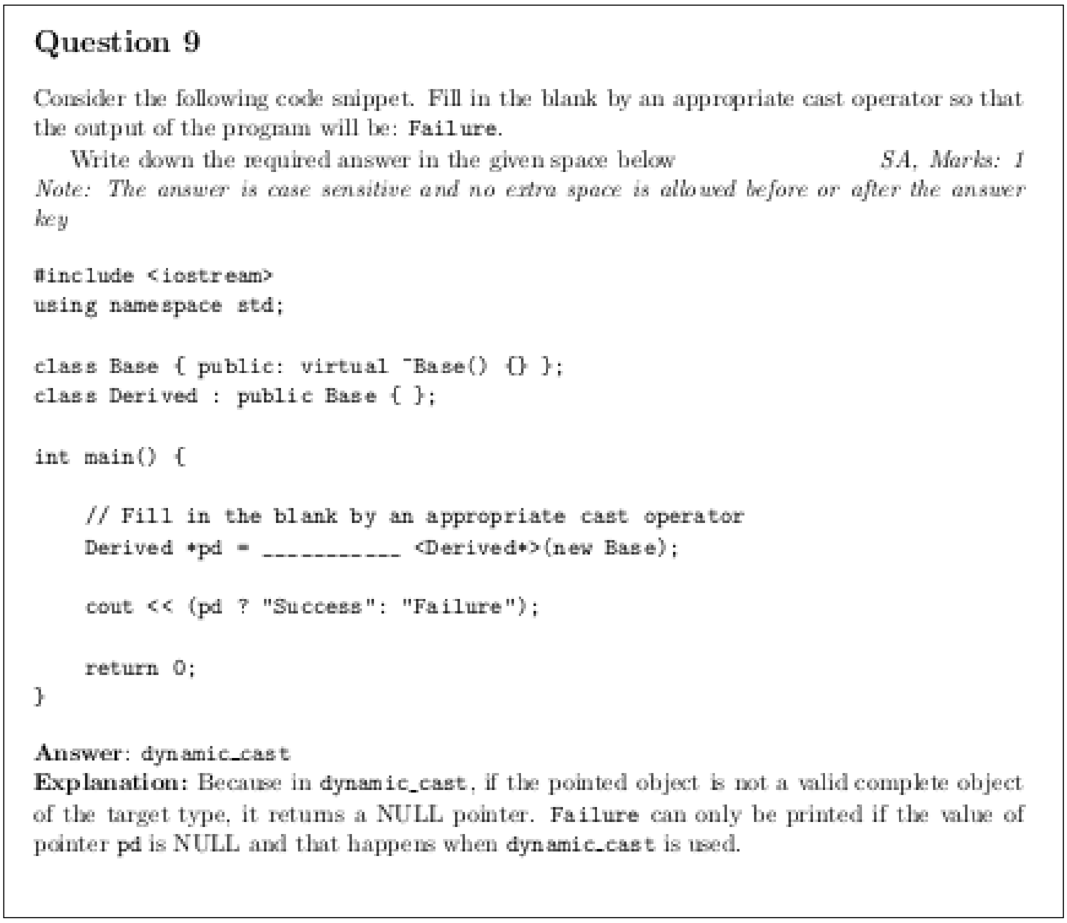
```
#include <iostream>
using namespace std;

class Base { public: virtual ~Base() {} };
class Derived : public Base { };

int main() {

    // Fill in the blank by an appropriate cast operator
    Derived *pd = ___________ <Derived*>(new Base);

    cout << (pd ? "Success": "Failure");

    return 0;
}
```

Answer: `dynamic_cast`

Explanation: Because in `dynamic_cast`, if the pointed object is not a valid complete object of the target type, it returns a NULL pointer. `Failure` can only be printed if the value of pointer `pd` is NULL and that happens when `dynamic_cast` is used.

Figure 3.32: MOOC Assignment 7, Question 9, Fill in the blanks

Also, in case of fill in the blanks, we have not any options so we could not recognize "*Question*" as shown in figure 3.33 and 3.32.



Question 9

Consider the following code snippet. Fill in the blank by an appropriate cast operator so that the output of the program will be: `Failure`.

Write down the required answer in the given space below *SA, Marks: 1*

Note: The answer is case sensitive and no extra space is allowed before or after the answer key

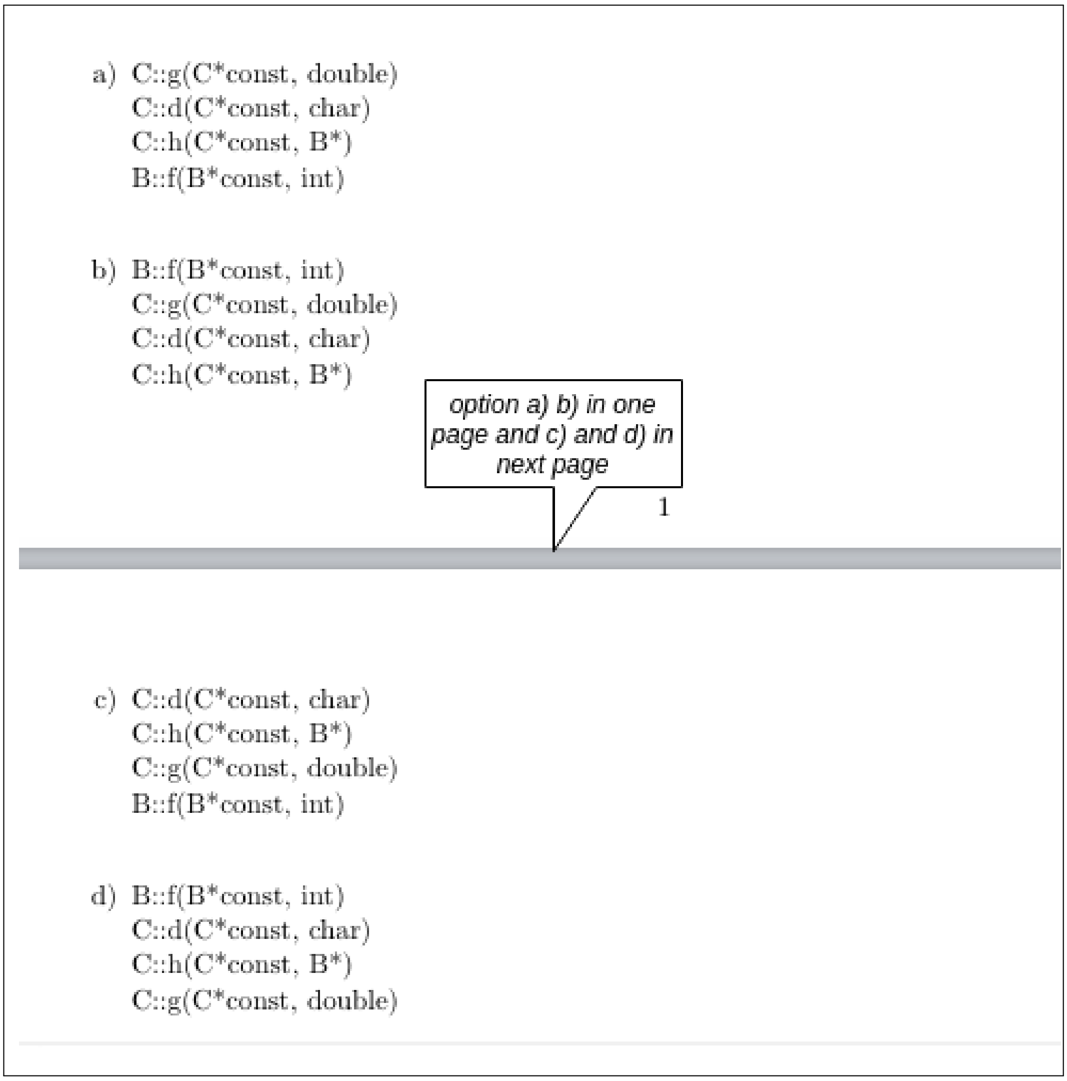
```
#include <iostream>
using namespace std;

class Base { public: virtual ~Base() {} };
class Derived : public Base { };

int main() {

    // Fill in the blank by an appropriate cast operator
    Derived *pd = ___________ <Derived*>(new Base);

    cout << (pd ? "Success": "Failure");

    return 0;
}
```

Answer: `dynamic_cast`

Explanation: Because in `dynamic_cast`, if the pointed object is not a valid complete object of the target type, it returns a NULL pointer. `Failure` can only be printed if the value of pointer `pd` is NULL and that happens when `dynamic_cast` is used.

Figure 3.33: Fill in the blanks

Suppose, some option are in first page and remaining few in next page for some question. Then it would be problem as in following shown figure 3.34.



a) C::g(C*const, double)
C::d(C*const, char)
C::h(C*const, B*)
B::f(B*const, int)

b) B::f(B*const, int)
C::g(C*const, double)
C::d(C*const, char)
C::h(C*const, B*)

option a) b) in one page and c) and d) in next page

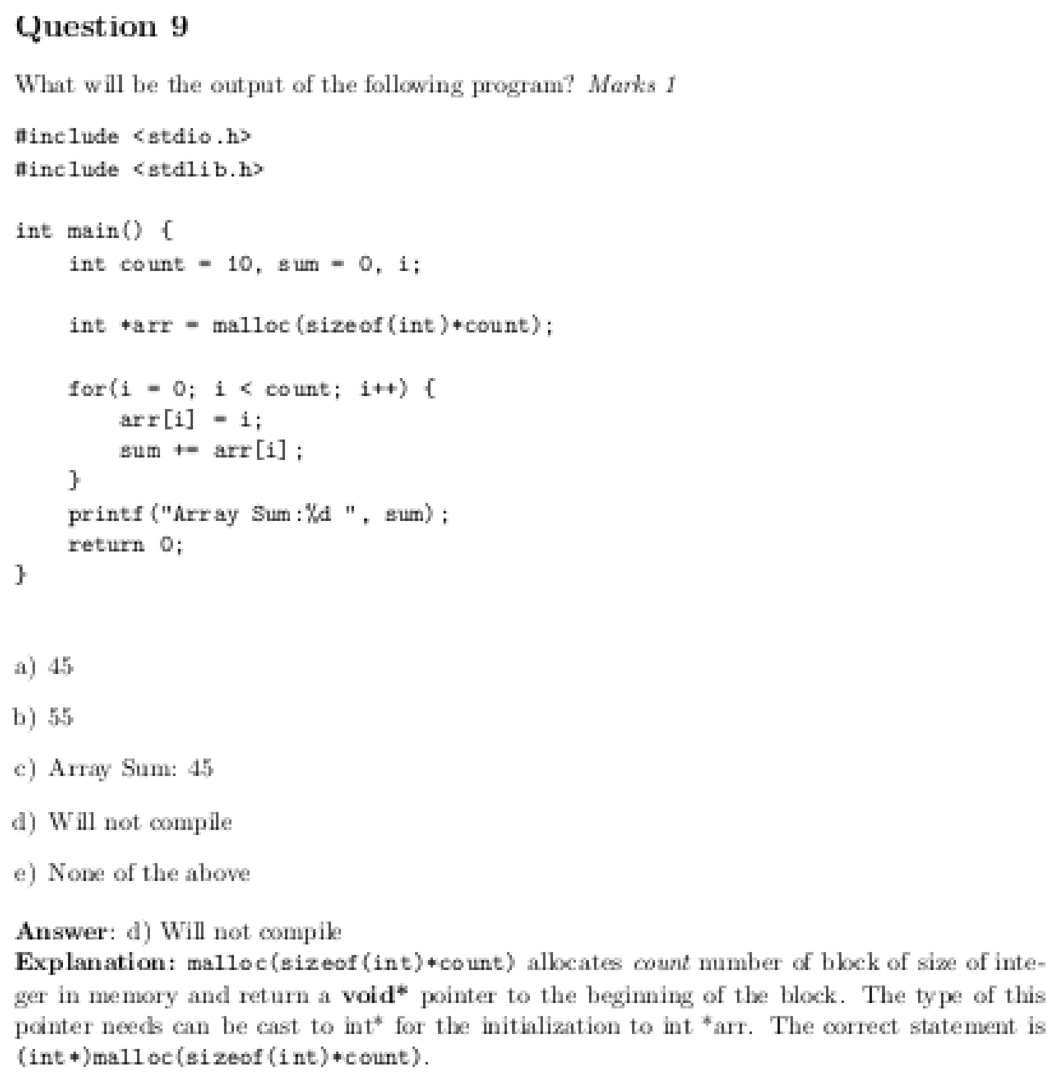
1

c) C::d(C*const, char)
C::h(C*const, B*)
C::g(C*const, double)
B::f(B*const, int)

d) B::f(B*const, int)
C::d(C*const, char)
C::h(C*const, B*)
C::g(C*const, double)

Figure 3.34: MOOC Assignment 7, Question 1, Few option are in next page

These are some question which have more than four options or less than four options.So, we restrict original question to only four question. In figure 3.35 has five options and figure 3.36 has six options.



Question 9

What will be the output of the following program? *Marks 1*

```
#include <stdio.h>
#include <stdlib.h>

int main() {
    int count = 10, sum = 0, i;

    int *arr = malloc(sizeof(int)*count);

    for(i = 0; i < count; i++) {
        arr[i] = i;
        sum += arr[i];
    }
    printf("Array Sum:%d ", sum);
    return 0;
}
```

a) 45

b) 55

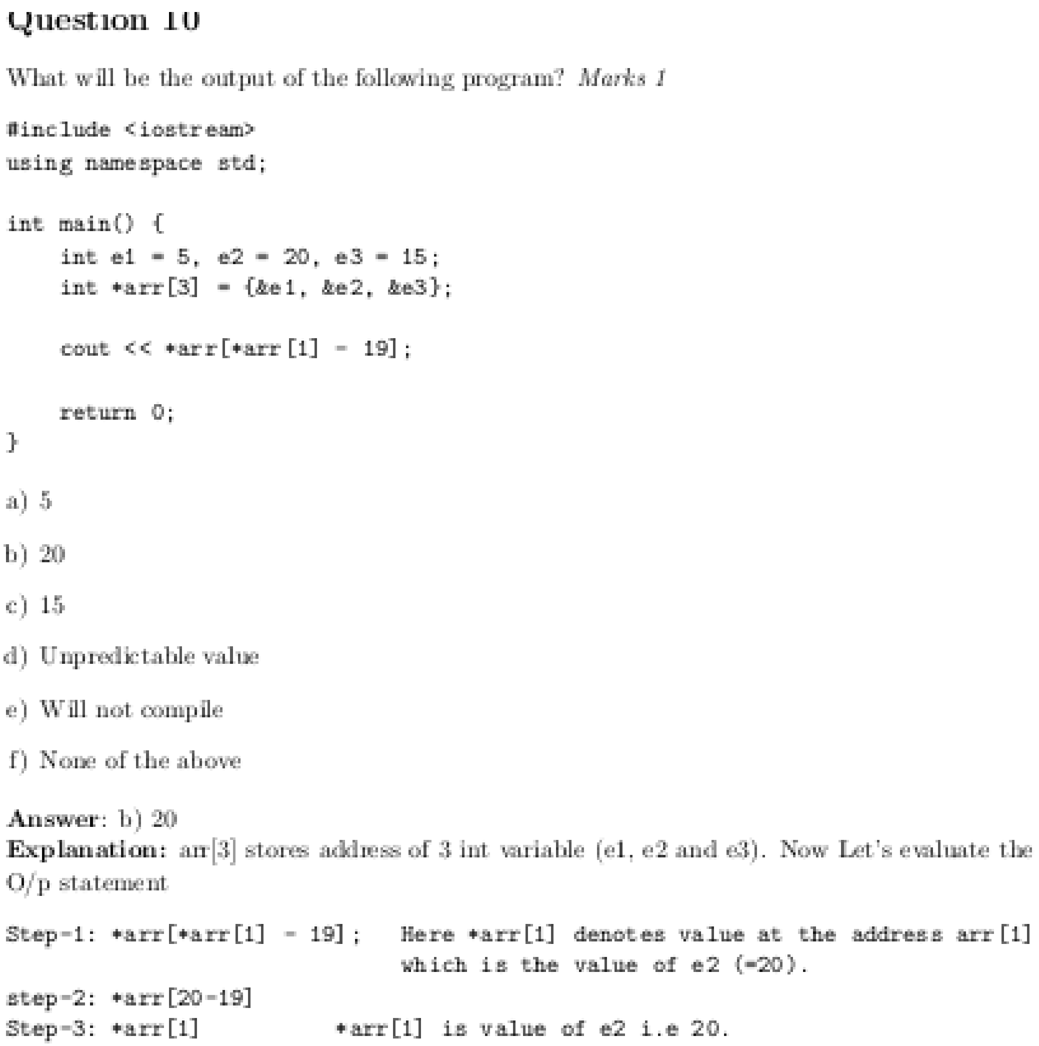
c) Array Sum: 45

d) Will not compile

e) None of the above

Answer: d) Will not compile
Explanation: `malloc(sizeof(int)*count)` allocates *count* number of block of size of integer in memory and return a **void*** pointer to the beginning of the block. The type of this pointer needs can be cast to int* for the initialization to int *arr. The correct statement is `(int*)malloc(sizeof(int)*count)`.

Figure 3.35: Five Options



Question 10

What will be the output of the following program? *Marks 1*

```
#include <iostream>
using namespace std;

int main() {
    int e1 = 5, e2 = 20, e3 = 15;
    int *arr[3] = {&e1, &e2, &e3};

    cout << *arr[*arr[1] - 19];

    return 0;
}
```

a) 5

b) 20

c) 15

d) Unpredictable value

e) Will not compile

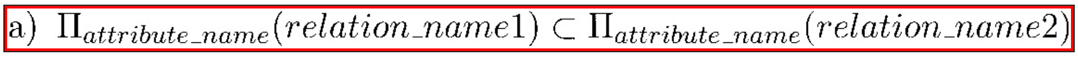
f) None of the above

Answer: b) 20
Explanation: arr[3] stores address of 3 int variable (e1, e2 and e3). Now Let's evaluate the O/p statement

```
Step-1: *arr[*arr[1] - 19];   Here *arr[1] denotes value at the address arr[1]
                              which is the value of e2 (=20).
step-2: *arr[20-19]
Step-3: *arr[1]          *arr[1] is value of e2 i.e 20.
```

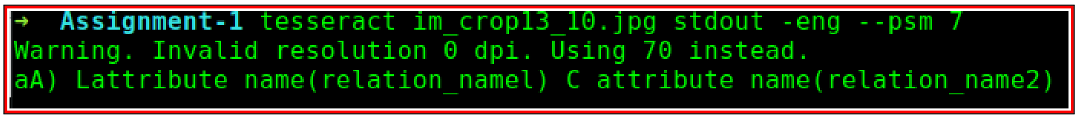
FIGURE 3.36: Six options

OCR give wrong output for some images so we could not fully depend on OCR. Because, our work is layout detection for Question and options, segmenting them accordingly. In figure 3.37, we had taken simple image but OCR gives wrong output as shown in figure 3.38. We are fully dependent for first two character on which this segmentation and work for recognizing *a)*. For, solving this issue we later try to include special symbol before question as shown in figure 3.62 and option as shown in figure 3.63.



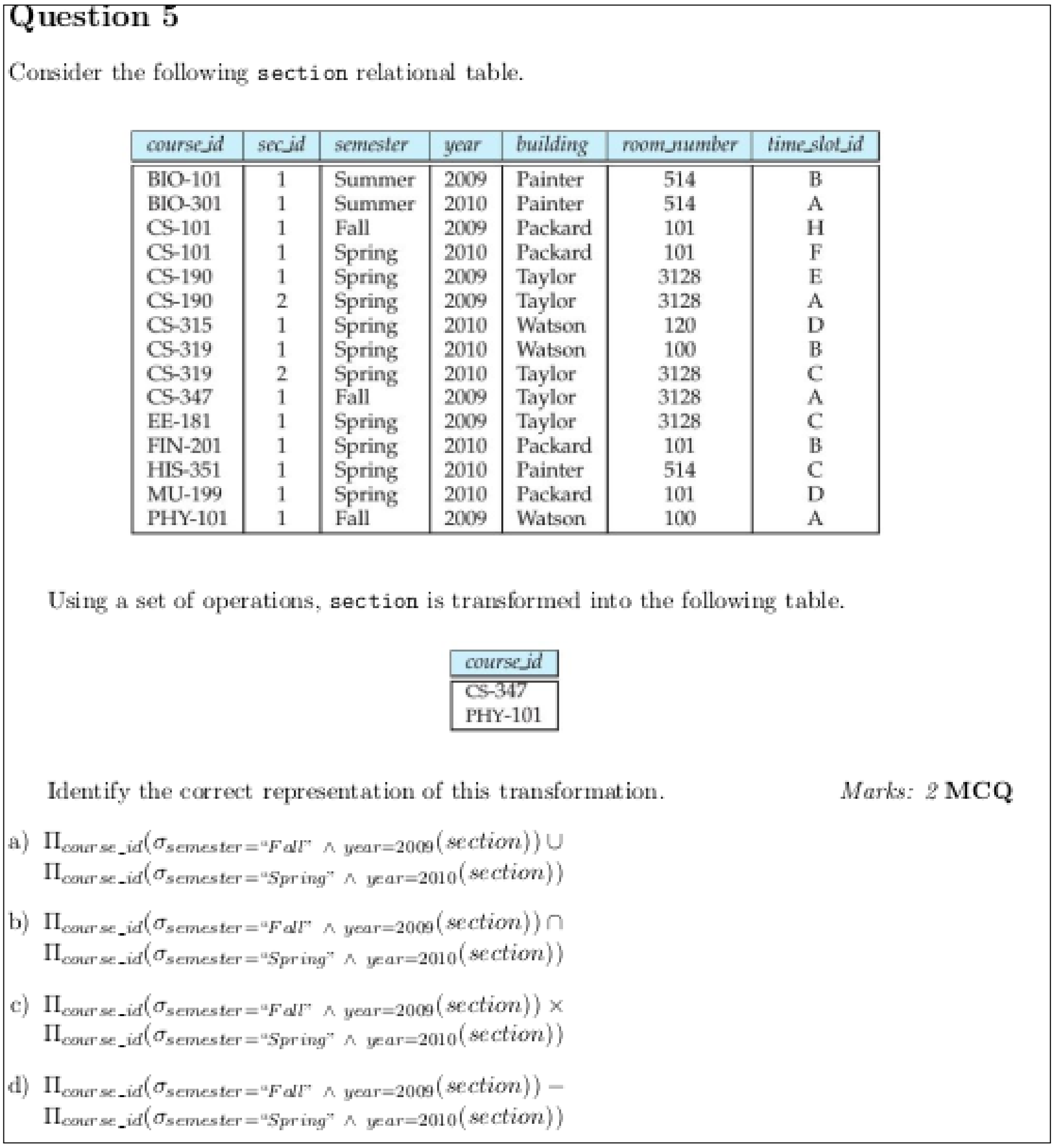
a) $\Pi_{attribute\_name}(relation\_name1) \subset \Pi_{attribute\_name}(relation\_name2)$

FIGURE 3.37: Input image



```
→  Assignment-1 tesseract im_crop13_10.jpg stdout -eng --psm 7
Warning. Invalid resolution 0 dpi. Using 70 instead.
aA) Lattribute name(relation_namel) C attribute name(relation_name2)
```

FIGURE 3.38: Output image, Expected output was ”a)” but we get ”A)”



Question 5

Consider the following `section` relational table.

course_id	*sec_id*	*semester*	*year*	*building*	*room_number*	*time_slot_id*
BIO-101	1	Summer	2009	Painter	514	B
BIO-301	1	Summer	2010	Painter	514	A
CS-101	1	Fall	2009	Packard	101	H
CS-101	1	Spring	2010	Packard	101	F
CS-190	1	Spring	2009	Taylor	3128	E
CS-190	2	Spring	2009	Taylor	3128	A
CS-315	1	Spring	2010	Watson	120	D
CS-319	1	Spring	2010	Watson	100	B
CS-319	2	Spring	2010	Taylor	3128	C
CS-347	1	Fall	2009	Taylor	3128	A
EE-181	1	Spring	2009	Taylor	3128	C
FIN-201	1	Spring	2010	Packard	101	B
HIS-351	1	Spring	2010	Painter	514	C
MU-199	1	Spring	2010	Packard	101	D
PHY-101	1	Fall	2009	Watson	100	A

Using a set of operations, `section` is transformed into the following table.

course_id
CS-347
PHY-101

Identify the correct representation of this transformation. *Marks: 2* **MCQ**

a) $\Pi_{course\_id}(\sigma_{semester="Fall" \wedge year=2009}(section)) \cup$
$\Pi_{course\_id}(\sigma_{semester="Spring" \wedge year=2010}(section))$

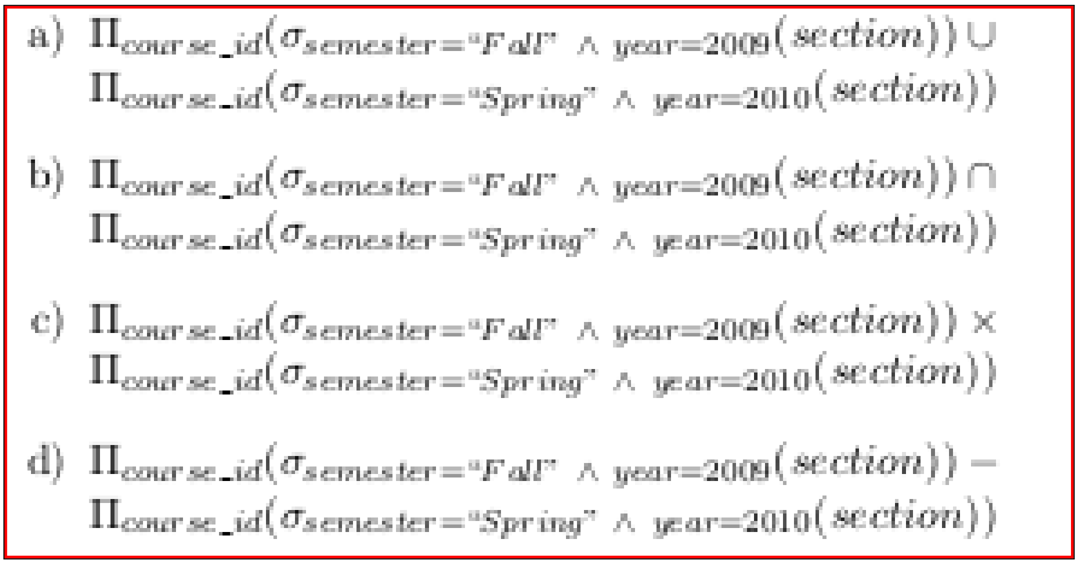
b) $\Pi_{course\_id}(\sigma_{semester="Fall" \wedge year=2009}(section)) \cap$
$\Pi_{course\_id}(\sigma_{semester="Spring" \wedge year=2010}(section))$

c) $\Pi_{course\_id}(\sigma_{semester="Fall" \wedge year=2009}(section)) \times$
$\Pi_{course\_id}(\sigma_{semester="Spring" \wedge year=2010}(section))$

d) $\Pi_{course\_id}(\sigma_{semester="Fall" \wedge year=2009}(section)) -$
$\Pi_{course\_id}(\sigma_{semester="Spring" \wedge year=2010}(section))$

FIGURE 3.39: Input Image, Options should be segmented as separate entity

As we can see taking figure 3.39 as reference, It has many special character in options. These special character are not segmented by our algorithm because all of them could not be recognized by OCR. We have shown as sample example in figure 3.40 that all option are separated as single entity. But, our required layout or segmentation was as in figure 3.41, 3.42, 3.43 and 3.44.



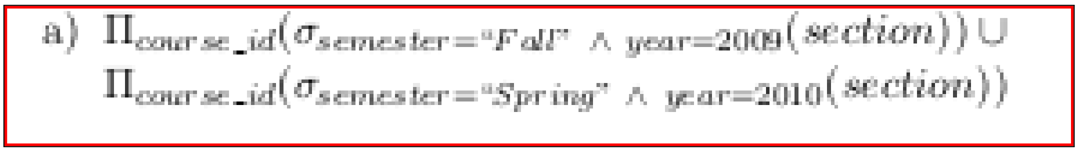
a) $\Pi_{course\_id}(\sigma_{semester=\text{“Fall”} \wedge year=2009}(section)) \cup$
$\Pi_{course\_id}(\sigma_{semester=\text{“Spring”} \wedge year=2010}(section))$

b) $\Pi_{course\_id}(\sigma_{semester=\text{“Fall”} \wedge year=2009}(section)) \cap$
$\Pi_{course\_id}(\sigma_{semester=\text{“Spring”} \wedge year=2010}(section))$

c) $\Pi_{course\_id}(\sigma_{semester=\text{“Fall”} \wedge year=2009}(section)) \times$
$\Pi_{course\_id}(\sigma_{semester=\text{“Spring”} \wedge year=2010}(section))$

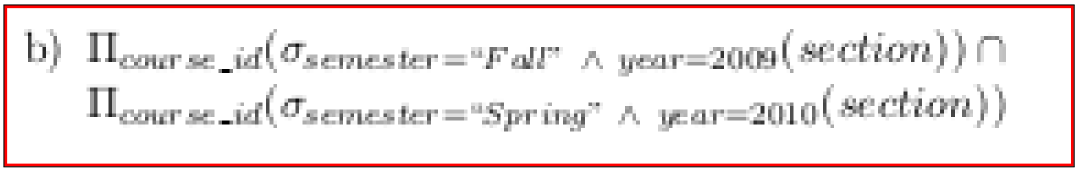
d) $\Pi_{course\_id}(\sigma_{semester=\text{“Fall”} \wedge year=2009}(section)) -$
$\Pi_{course\_id}(\sigma_{semester=\text{“Spring”} \wedge year=2010}(section))$

FIGURE 3.40: DBMS 1st run, Assignment 3, Ques 5,All options are separated as single entity



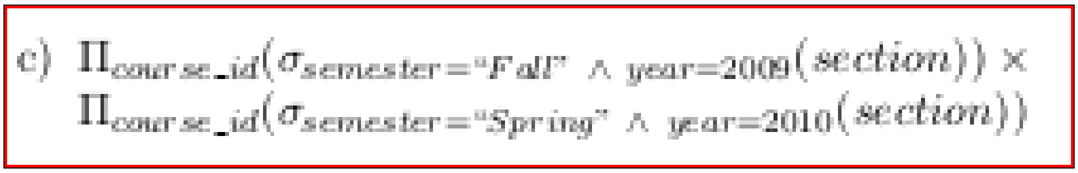
a) $\Pi_{course\_id}(\sigma_{semester=\text{“Fall”} \wedge year=2009}(section)) \cup$
$\Pi_{course\_id}(\sigma_{semester=\text{“Spring”} \wedge year=2010}(section))$

FIGURE 3.41: Expected ”a)” image



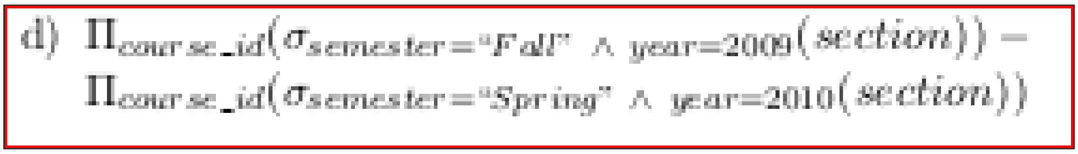
b) $\Pi_{course\_id}(\sigma_{semester=\text{“Fall”} \wedge year=2009}(section)) \cap$
$\Pi_{course\_id}(\sigma_{semester=\text{“Spring”} \wedge year=2010}(section))$

FIGURE 3.42: Expected ”b)” image



c) $\Pi_{course\_id}(\sigma_{semester=\text{“Fall”} \wedge year=2009}(section)) \times$
$\Pi_{course\_id}(\sigma_{semester=\text{“Spring”} \wedge year=2010}(section))$

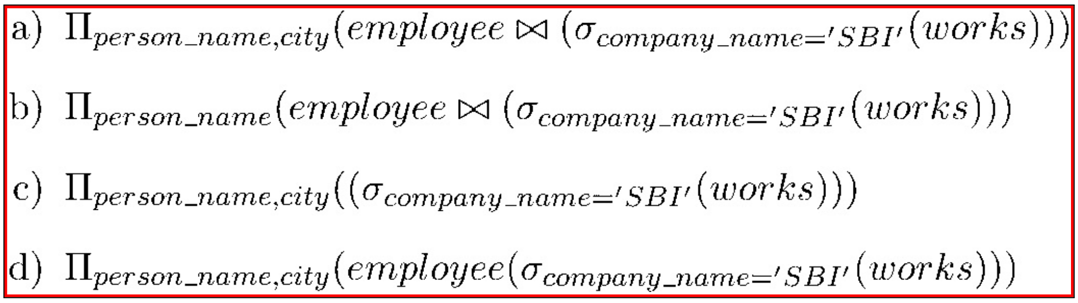
FIGURE 3.43: Expected ”c)” image



d) $\Pi_{course\_id}(\sigma_{semester=\text{“Fall”} \wedge year=2009}(section)) -$
$\Pi_{course\_id}(\sigma_{semester=\text{“Spring”} \wedge year=2010}(section))$

FIGURE 3.44: Expected ”d)” image

Same problem is faced in some other quesiton also as shown in figure 3.45 as single entity but required was as shown in figure. 3.46, 3.47, 3.48 and 3.49 as separate entity.



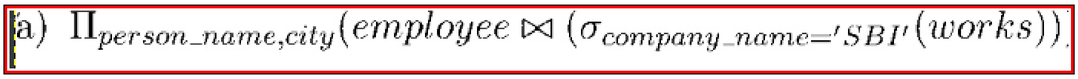
a) $\Pi_{person\_name,city}(employee \bowtie (\sigma_{company\_name='SBI'}(works)))$

b) $\Pi_{person\_name}(employee \bowtie (\sigma_{company\_name='SBI'}(works)))$

c) $\Pi_{person\_name,city}((\sigma_{company\_name='SBI'}(works)))$

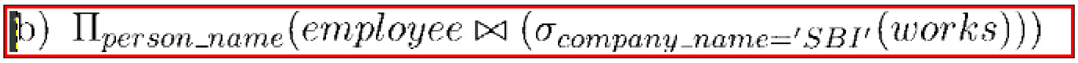
d) $\Pi_{person\_name,city}(employee(\sigma_{company\_name='SBI'}(works)))$

FIGURE 3.45: DBMS 1st run, Assignment 3, Ques 7,All options are separated as single entity



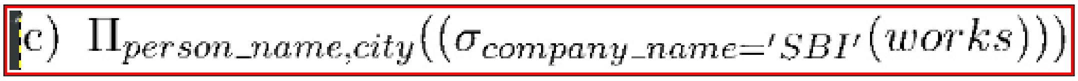
a) $\Pi_{person\_name,city}(employee \bowtie (\sigma_{company\_name='SBI'}(works))$

FIGURE 3.46: Expected ”a)” image



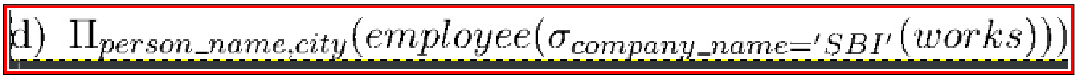
b) $\Pi_{person\_name}(employee \bowtie (\sigma_{company\_name='SBI'}(works)))$

FIGURE 3.47: Expected ”b)” image



c) $\Pi_{person\_name,city}((\sigma_{company\_name='SBI'}(works)))$

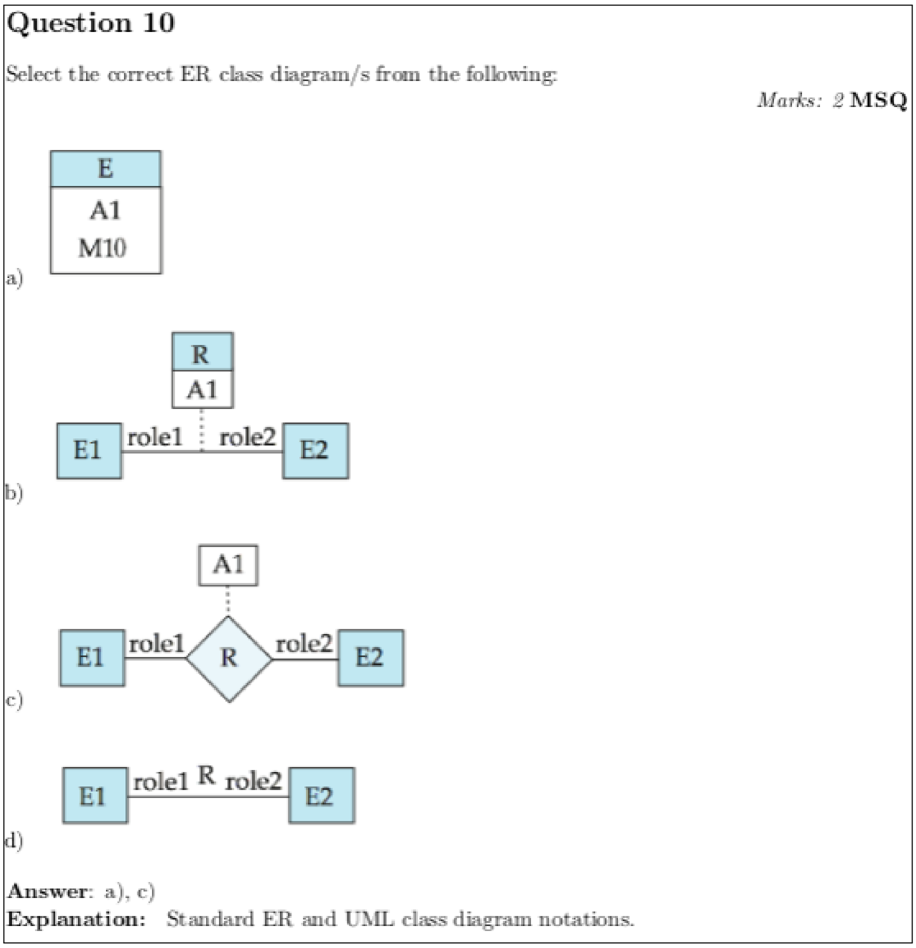
FIGURE 3.48: Expected ”c)” image



d) $\Pi_{person\_name,city}(employee(\sigma_{company\_name='SBI'}(works)))$

FIGURE 3.49: Expected ”d)” image

Some options are not properly formatted. Due to this question part and option "a)" part are segmented in same image. This is mainly due to sudden layout changes between option and insertion of un-formatted diagram.



Question 10

Select the correct ER class diagram/s from the following:

Marks: 2 **MSQ**

a)

b)

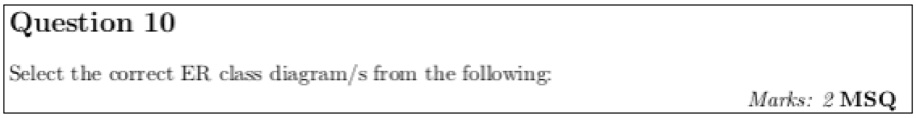
c)

d)

Answer: a), c)
Explanation: Standard ER and UML class diagram notations.

FIGURE 3.50: DBMS 1st run, Assignment 3, Ques 10

We have taken figure 3.50 as input and expecting figure 3.51 as possible question output.



Question 10

Select the correct ER class diagram/s from the following:

Marks: 2 **MSQ**

FIGURE 3.51: DBMS 1st run, Assignment 3, Ques 10, Expecting question as separated entity

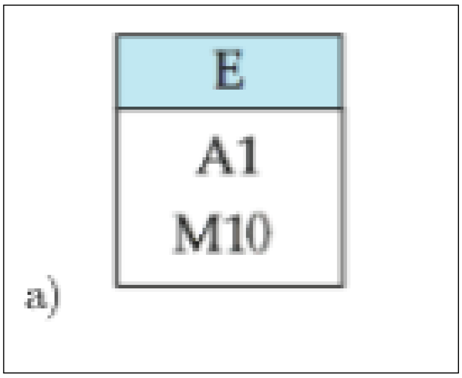
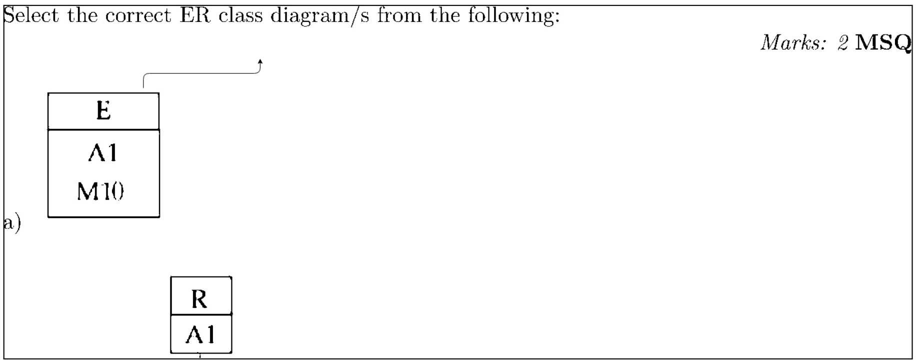


FIGURE 3.52: DBMS 1st run, Assignment 3, Ques 10, Expecting option as separate entity


Select the correct ER class diagram/s from the following:

Marks: 2 **MSQ**

FIGURE 3.53: DBMS 1st run, Assignment 3, Ques 10, But get question and ”a)” option as single entity

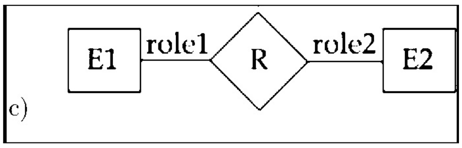


FIGURE 3.54: DBMS 1st run, Assignment 3, Ques 10, Expected ”c)” option as single entity

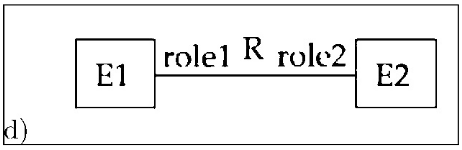


FIGURE 3.55: Expected ”d)” option as single entity

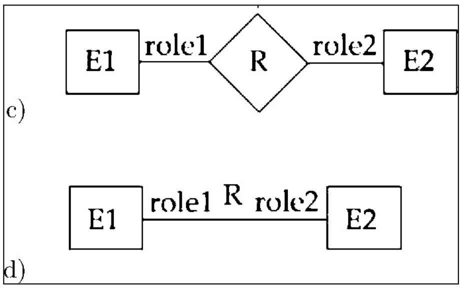
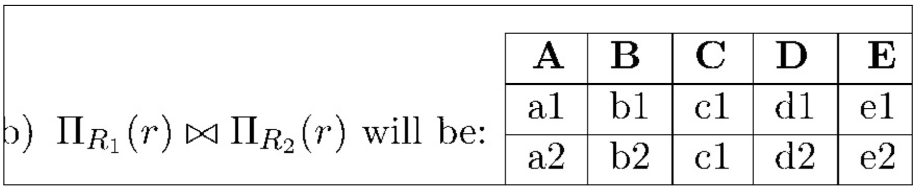


FIGURE 3.56: DBMS 1st run, Assignment 3, Ques 10, But get ”c)” and ”d)” option as single entity

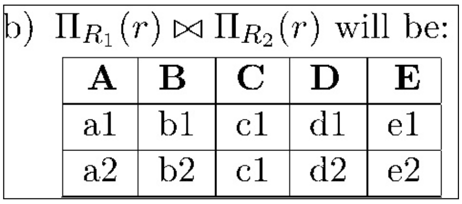
Some question have diagram in option along with text, but these diagram should start from next line.So, next next option can be easily segmented.



b) $\Pi_{R_1}(r) \bowtie \Pi_{R_2}(r)$ will be:

A	B	C	D	E
a1	b1	c1	d1	e1
a2	b2	c1	d2	e2

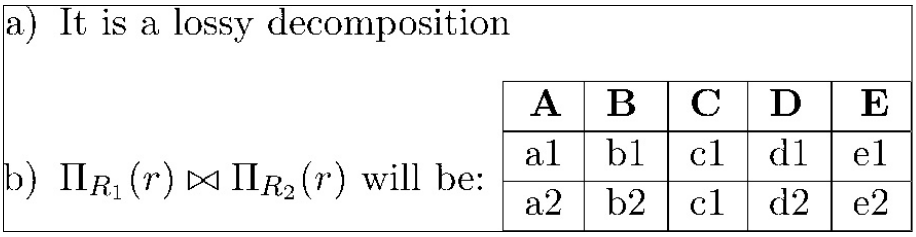
FIGURE 3.57: DBMS 1st run, Assignment 4, Ques 21, Given



b) $\Pi_{R_1}(r) \bowtie \Pi_{R_2}(r)$ will be:

A	B	C	D	E
a1	b1	c1	d1	e1
a2	b2	c1	d2	e2

FIGURE 3.58: DBMS 1st run, Assignment 3, Ques 10, Expected



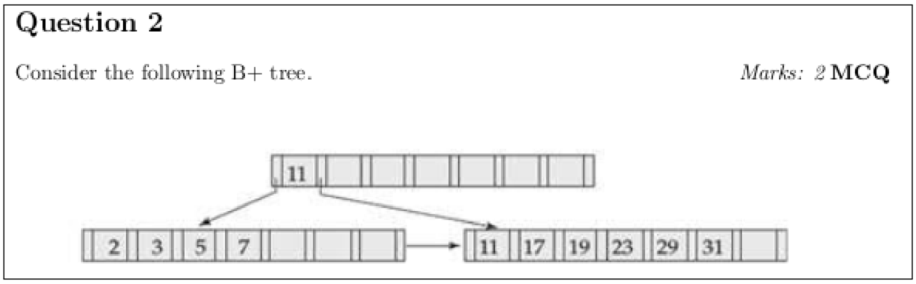
a) It is a lossy decomposition

b) $\Pi_{R_1}(r) \bowtie \Pi_{R_2}(r)$ will be:

A	B	C	D	E
a1	b1	c1	d1	e1
a2	b2	c1	d2	e2

FIGURE 3.59: DBMS 1st run, Assignment 3, Ques 10, Current wrong output

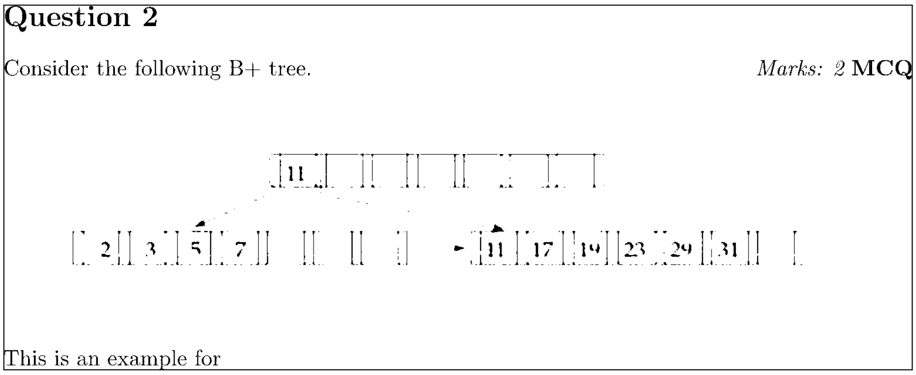
Diagram and figure should have only black and white color combination. Suppose, here we have included gray image but after binarization it becomes blurry and unable to recognize. Some of sample questions in DBMS 1st run, assignment 6 are 2,3,4.



Question 2

Consider the following B+ tree. *Marks: 2* **MCQ**

FIGURE 3.60: DBMS 1st run, Assignment 6, Ques 2, Actual image



Question 2

Consider the following B+ tree. *Marks: 2* **MCQ**

This is an example for

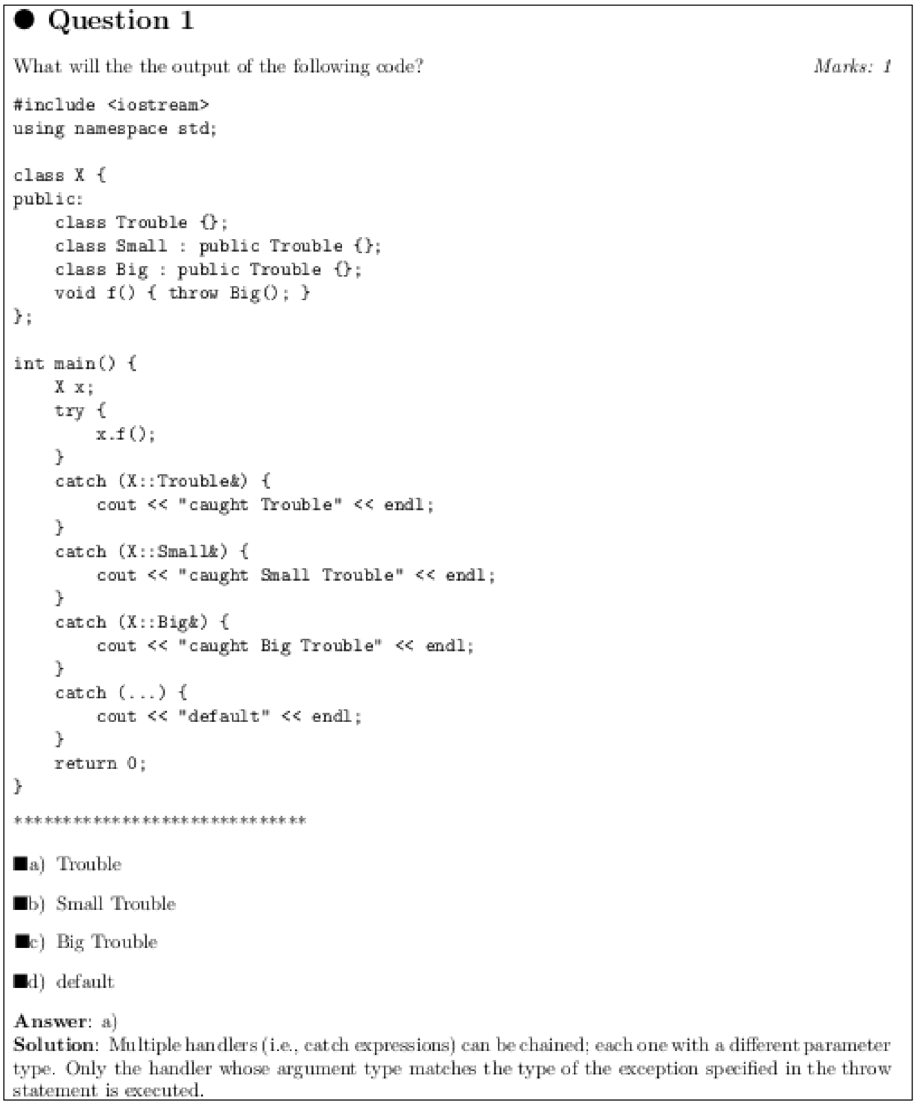
FIGURE 3.61: DBMS 1st run, Assignment 6, Ques 2, Blurry image unable to recognize

For comprehension question, added paragraph/passage to each question for testing. These give correct segmentation. But, there would be problem in numbering question. Question may be misnumbered. Example: Comprehension have 5A, 5B and 5C as three question of comprehension. But, in output we would get question 6,7 and 8 respectively.

All these are some of problem faced during execution of task. Some previous problem get solved and few are still remaining which are mentioned.

3.5.4 Possible Improvement

We add special symbols and try to segment them. All option have a unique square marker. Fill in the blanks have empty circle marker. Comprehension have triangle marker. Normal question have filled circle as marker. Possible symbols may be following:



● **Question 1**

What will the the output of the following code? *Marks: 1*

```
#include <iostream>
using namespace std;

class X {
public:
    class Trouble {};
    class Small : public Trouble {};
    class Big : public Trouble {};
    void f() { throw Big(); }
};

int main() {
    X x;
    try {
        x.f();
    }
    catch (X::Trouble&) {
        cout << "caught Trouble" << endl;
    }
    catch (X::Small&) {
        cout << "caught Small Trouble" << endl;
    }
    catch (X::Big&) {
        cout << "caught Big Trouble" << endl;
    }
    catch (...) {
        cout << "default" << endl;
    }
    return 0;
}
******************************
```

■a) Trouble

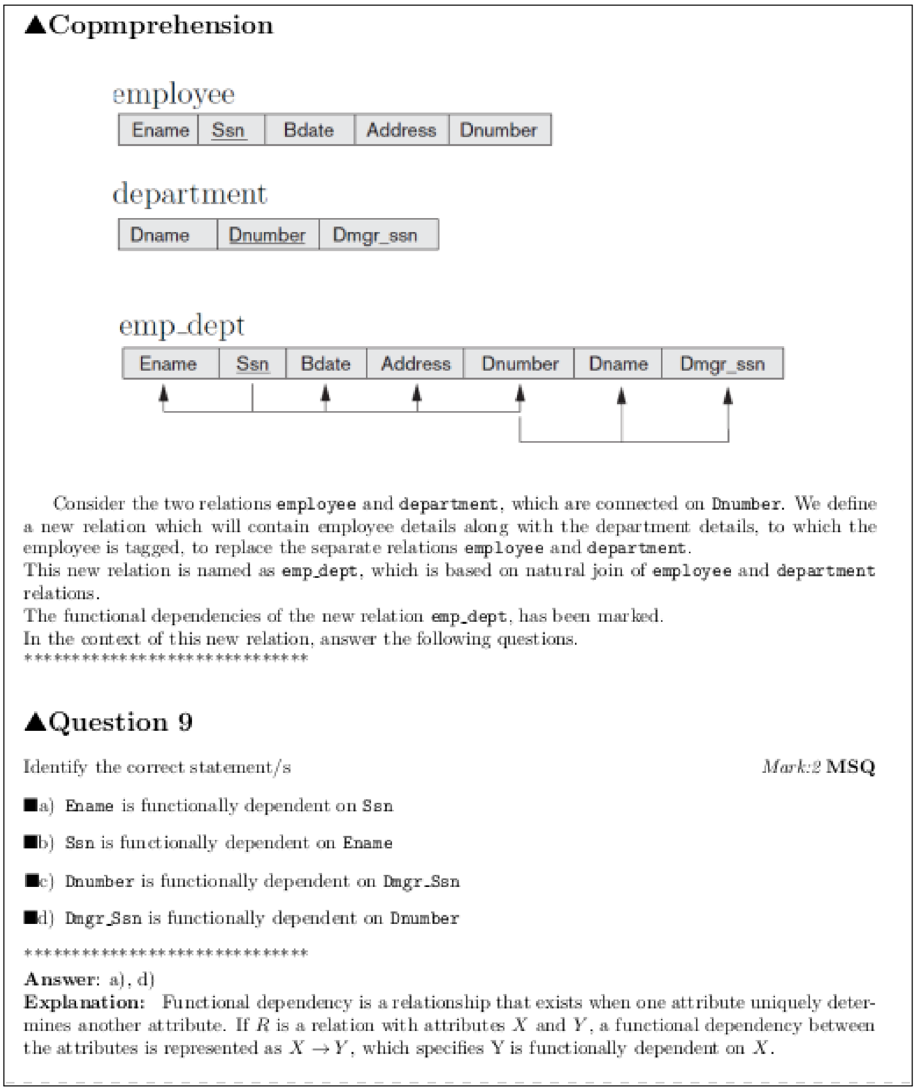
■b) Small Trouble

■c) Big Trouble

■d) default

Answer: a)
Solution: Multiple handlers (i.e., catch expressions) can be chained; each one with a different parameter type. Only the handler whose argument type matches the type of the exception specified in the throw statement is executed.

Figure 3.62: Normal question, Filled circle marker



▲**Copmprehension**

employee

Ename	Ssn	Bdate	Address	Dnumber

department

Dname	Dnumber	Dmgr_ssn

emp_dept

Ename	Ssn	Bdate	Address	Dnumber	Dname	Dmgr_ssn

Consider the two relations `employee` and `department`, which are connected on `Dnumber`. We define a new relation which will contain employee details along with the department details, to which the employee is tagged, to replace the separate relations `employee` and `department`.
This new relation is named as `emp_dept`, which is based on natural join of `employee` and `department` relations.
The functional dependencies of the new relation `emp_dept`, has been marked.
In the context of this new relation, answer the following questions.
****

▲**Question 9**

Identify the correct statement/s *Mark:2* **MSQ**

■a) `Ename` is functionally dependent on `Ssn`

■b) `Ssn` is functionally dependent on `Ename`

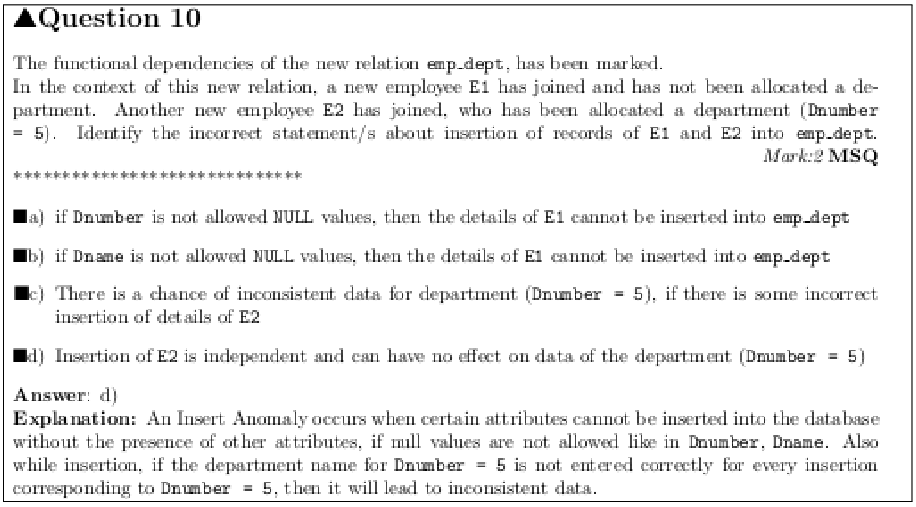
■c) `Dnumber` is functionally dependent on `Dmgr_Ssn`

■d) `Dmgr_Ssn` is functionally dependent on `Dnumber`

****

Answer: a), d)
Explanation: Functional dependency is a relationship that exists when one attribute uniquely determines another attribute. If R is a relation with attributes X and Y, a functional dependency between the attributes is represented as $X \rightarrow Y$, which specifies Y is functionally dependent on X.

FIGURE 3.63: Comprehension question, Filled triangle marker



▲Question 10

The functional dependencies of the new relation `emp_dept`, has been marked.
In the context of this new relation, a new employee `E1` has joined and has not been allocated a department. Another new employee `E2` has joined, who has been allocated a department (`Dnumber = 5`). Identify the incorrect statement/s about insertion of records of `E1` and `E2` into `emp_dept`.
Mark:2 **MSQ**

****

■a) if `Dnumber` is not allowed `NULL` values, then the details of `E1` cannot be inserted into `emp_dept`

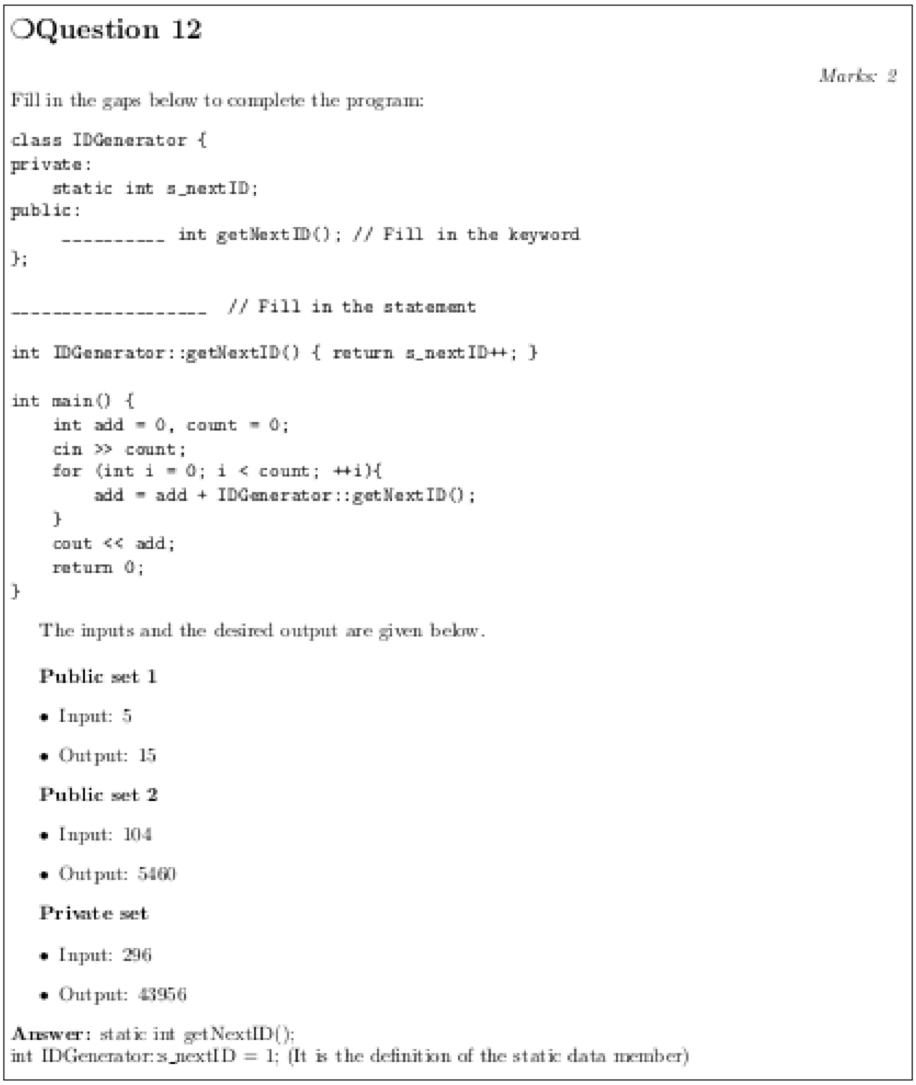
■b) if `Dname` is not allowed `NULL` values, then the details of `E1` cannot be inserted into `emp_dept`

■c) There is a chance of inconsistent data for department (`Dnumber = 5`), if there is some incorrect insertion of details of `E2`

■d) Insertion of `E2` is independent and can have no effect on data of the department (`Dnumber = 5`)

Answer: d)
Explanation: An Insert Anomaly occurs when certain attributes cannot be inserted into the database without the presence of other attributes, if null values are not allowed like in `Dnumber`, `Dname`. Also while insertion, if the department name for `Dnumber = 5` is not entered correctly for every insertion corresponding to `Dnumber = 5`, then it will lead to inconsistent data.

Figure 3.64: Comprehension next question, Filled triangle marker



○**Question 12**

Marks: 2

Fill in the gaps below to complete the program:

```
class IDGenerator {
private:
    static int s_nextID;
public:
     __________ int getNextID(); // Fill in the keyword
};

___________________  // Fill in the statement

int IDGenerator::getNextID() { return s_nextID++; }

int main() {
    int add = 0, count = 0;
    cin >> count;
    for (int i = 0; i < count; ++i){
        add = add + IDGenerator::getNextID();
    }
    cout << add;
    return 0;
}
```

The inputs and the desired output are given below.

Public set 1

- Input: 5
- Output: 15

Public set 2

- Input: 104
- Output: 5460

Private set

- Input: 296
- Output: 43956

Answer: static int getNextID();
int IDGenerator:s_nextID = 1; (It is the definition of the static data member)

Figure 3.65: Fill in the blanks question, Empty circle marker, It has not any option

3.5.5 Conclusion

Some of changes in latex generated pdf are possible, which is feasible. But proposed change may not be possible. So, we have to add particular special symbol which can be recognised by OCR.

Bibliography

Anjo Anjewierden. Aidas: Incremental logical structure discovery in pdf documents. In *icdar*, page 0374. IEEE, 2001.

Hui Chao and Jian Fan. Layout and content extraction for pdf documents. In *International Workshop on Document Analysis Systems*, pages 213–224. Springer, 2004.

Hui Chao, Giordano Beretta, and Henry Sang. Pdf document layout study with page elements and bounding boxes. In *Workshop on document layout interpretation and its applications (DLIA2001)*, 2001.

Jian Fan. Text extraction via an edge-bounded averaging and a parametric character model. In *Document Recognition and Retrieval X*, volume 5010, pages 8–20. International Society for Optics and Photonics, 2003.

Karim Hadjar, Maurizio Rigamonti, Denis Lalanne, and Rolf Ingold. Xed: a new tool for extracting hidden structures from electronic documents. In *Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop on*, pages 212–224. IEEE, 2004.

Patrick Haffner, Léon Bottou, Paul G Howard, and Yann LeCun. Djvu: Analyzing and compressing scanned documents for internet distribution. In *Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on*, pages 625–628. IEEE, 1999.

Tamir Hassan and Robert Baumgartner. Table recognition and understanding from pdf files. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 1143–1147. IEEE, 2007.

William S Lovegrove and David F Brailsford. Document analysis of pdf files: methods, results and implications. *Electronic Publishing–Origination, Dissemination and Design*, 8(3):207–220, 1995.

Cartic Ramakrishnan, Abhishek Patnia, Eduard Hovy, and Gully APC Burns. Layout-aware text extraction from full-text pdf of scientific articles. *Source code for biology and medicine*, 7(1):7, 2012.

Philip N Smith and David F Brailsford. Towards structured, block-based pdf. *Electronic Publishing–Origination, Dissemination and Design*, 8(3):153–165, 1995.