Βάσεις Δεδομένων και Εξόρυξη Δεδομένων

Ενδεικτικές ερωτήσεις-θέματα για την εξέταση της θεωρίας

1. Τι ονομάζεται Σύστημα Διαχείρισης Βάσεων Δεδομένων (database management system); Αναφέρετε τα πλεονεκτήματα των Συστημάτων Διαχείρισης Βάσεων Δεδομένων και μερικές σύγχρονές εφαρμογές τους.

2. Τι ονομάζεται Εξόρυξη Δεδομένων; Τι πλεονέκτημα έχει; Αναφέρετε μερικά κίνητρα για την εξόρυξη δεδομένων. Περιγράψτε σχηματικά την διαδικασία της εξόρυξης δεδομένων (ή τα βήματά της).

3. Ποιες είναι οι κατηγορίες χρηστών ενός Συστήματος Διαχείρισης Βάσεων Δεδομένων και τι δικαιώματα έχουν; Σε ποια επίπεδα του συστήματος αλληλεπιδρά ο απλός χρήστης και σε ποια ο διαχειριστής και πως;

4. Ποια είναι τα τρία επίπεδα της αρχιτεκτονικής ενός Συστήματος Διαχείρισης Βάσεων Δεδομένων; Για ποιον σκοπό γίνεται ο διαχωρισμός αυτός σε επίπεδα; Περιγράψτε τι περιλαμβάνει το κάθε επίπεδο.

5. Τι ονομάζεται Σχήμα μίας Βάσης Δεδομένων; Ποια είδη σχημάτων υπάρχουν; Τι ονομάζεται Στιγμιότυπο μίας Βάσης Δεδομένων; Πως εξασφαλίζεται η ανεξαρτησία των δεδομένων σε ένα Σύστημα Διαχείρισης Βάσεων Δεδομένων; Ποιοι είναι οι δύο τύποι ανεξαρτησίας;

6. Από τι αποτελείται μία Γλώσσα Ερωτημάτων (query language) και για ποιο σκοπό; Ποιες είναι οι Γλώσσες 4^{ης} Γενιάς και ποια χαρακτηριστικά έχουν;

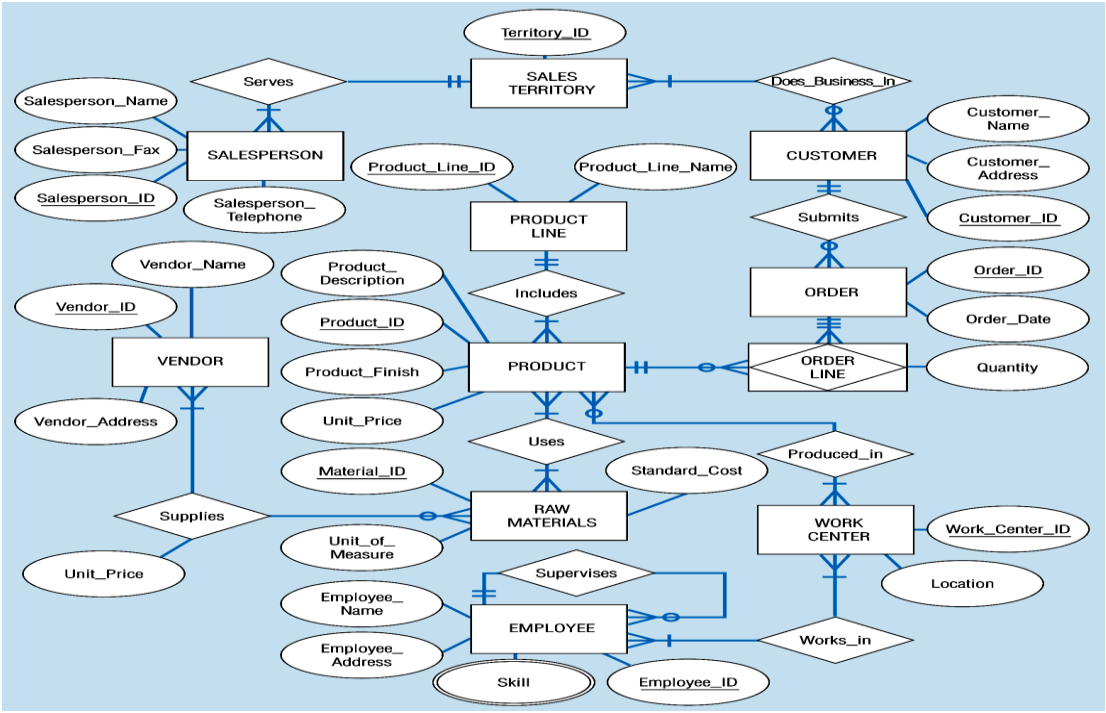
7. Στο μοντέλο ER τι ονομάζεται Οντότητα (Entity); Τι ονομάζεται ισχυρή οντότητα (strong entity) και τι ιδιότητες έχει; Τι ονομάζεται ασθενής οντότητα (weak entity) και τι ιδιότητες έχει; Τι είναι η συσχετιστική οντότητα και ποιες είναι οι προϋποθέσεις για τη δημιουργία της;

8. Στο μοντέλο ER τι ονομάζεται Χαρακτηριστικό (Attribute) μίας οντότητας και τι είναι το πεδίο ορισμού του; Τι ονομάζεται Σύνθετο Χαρακτηριστικό και τι ιδιότητες έχει; Τι ονομάζεται Χαρακτηριστικό Πολλαπλών Τιμών και τι ιδιότητες έχει; Τι ονομάζεται Υπολογιζόμενο Χαρακτηριστικό και τι ιδιότητες έχει;

9. Στο μοντέλο ER τι ονομάζεται Συσχέτιση (Relationship); Τι είναι η συνδετικότητα μίας συσχέτισης και ποια είδη υπάρχουν; Τι είναι ο πληθικός αριθμός μίας συσχέτισης; Πότε μία οντότητα συμμετέχει προαιρετικά και πότε υποχρεωτικά σε μία συσχέτιση; Τι ονομάζεται Βαθμός μίας συσχέτισης;

10. Στο εκτεταμένο μοντέλο ER συχνά παρουσιάζονται συγγενείς μεταξύ τους οντότητες. Τι χαρακτηριστικά έχουν; Με την ιεραρχία γενίκευσης και εξειδίκευσης σε υπερκατηγορίες και υποκατηγορίες τι χαρακτηριστικά περιλαμβάνει η κάθε οντότητα; Ποιοι είναι οι περιορισμοί πληρότητας στο μοντέλο ER; Ποιοι είναι οι περιορισμοί επικάλυψης στο μοντέλο ER;

11. Δίνεται το παρακάτω ER διάγραμμα μίας βάσης δεδομένων. Να αναγνωρίσετε και να καταγράψετε τις Οντότητές του, τα Χαρακτηριστικά της κάθε οντότητας και τις Συσχετίσεις που υπάρχουν, καθώς και κάθε ιδιότητά τους:



12. Στο Σχεσιακό Μοντέλο τι ονομάζεται Σχέση (Relation), από τι αποτελείται και τι ιδιότητες έχει; Μπορούμε να έχουμε σύνθετα χαρακτηριστικά ή χαρακτηριστικά πολλαπλών τιμών σε μία Σχέση; Τι είναι τα πρωτεύοντα κλειδιά και τι τα ξένα κλειδιά;

13. Στο Σχεσιακό Μοντέλο πως μετατρέπεται μία οντότητα σε πίνακα; Πως γίνεται η μετατροπή όταν έχει ένα σύνθετο χαρακτηριστικό (composite attribute); Πως γίνεται η μετατροπή όταν έχει ένα χαρακτηριστικό πολλαπλών τιμών (multivalued attribute); Πως γίνεται η μετατροπή μίας ασθενούς οντότητας (weak entity);

14. Στο Σχεσιακό Μοντέλο πως μετατρέπονται οι συσχετίσεις $2^{ου}$ βαθμού; Αναφέρετε τη διαδικασία και για τις τρεις περιπτώσεις συσχετίσεων: 1:1, 1:N, N:M.

15. Στο Σχεσιακό Μοντέλο πως μετατρέπονται οι συσχετίσεις $1^{ου}$ βαθμού; Αναφέρετε τη διαδικασία και για τις τρεις περιπτώσεις συσχετίσεων: 1:1, 1:N, N:M. Πως μετατρέπονται οι συσχετίσεις $3^{ου}$ ή μεγαλυτέρου βαθμού;

16. Ποια είναι τα πλεονεκτήματα της γλώσσας SQL και ποια τα χαρακτηριστικά της; Σε ποιες υπο-γλώσσες διαιρείται και τι περιλαμβάνει η κάθε μία; Ποιοι είναι οι βασικοί τύποι δεδομένων στη γλώσσα SQL;

17. Να περιγράψτε τι ακριβώς κάνουν οι ακόλουθες εντολές της γλώσσας SQL καθώς και τα αποτελέσματά τους στη βάση δεδομένων:

```
CREATE TABLE Γνωστική_Περιοχή
        (κωδικός INTEGER NOT NULL,
        τίτλος CHAR(50) NOT NULL,
        αριθμός_συνδρομητών INTEGER DEFAULT 0,
        PRIMARY KEY (κωδικός),
        UNIQUE (τίτλος),
        CHECK (αριθμός_συνδρομητών >= 0));
```

```
CREATE TABLE Πρακτικά_Συνεδρίου
        (κωδικός INTEGER NOT NULL,
        συνέδριο VARCHAR(100) NOT NULL,
        ημερομηνία DATE NOT NULL,
        χώρα CHAR(20),
```

```
        κωδικός_εκδοτικού_οίκου INTEGER NOT NULL,
        PRIMARY KEY (κωδικός),
        FOREIGN KEY (κωδικός_εκδοτικού_οίκου) REFERENCES  Εκδοτικός_Οίκος (κωδικός),
        ON DELETE RESTRICTED,
        ON UPDATE CASCADE);
ALTER TABLE Πρακτικά_Συνεδρίου ADD COLUMN Πόλη CHAR(10);
        ALTER TABLE  Συνδρομητής ADD COLUMN ημερομηνία_γέννησης DATE NOT NULL;
ALTER TABLE Πρακτικά_Συνεδρίου DROP COLUMN Πόλη;
        ALTER TABLE  Συνδρομητής DROP COLUMN ημερομηνία_γέννησης;
ALTER TABLE Πρακτικά_Συνεδρίου MODIFY χώρα CHAR(15);
```

18. Δίνεται το παρακάτω σχήμα μίας βάσης δεδομένων με συνδρομητικά δεδομένα:

Συνδρομητής (*κωδικός*, *όνομα*, *οδός*, *αριθμός*, *ΤΚ*, *πόλη*, *χώρα*, *ΑΠΚ*)
Τηλέφωνο_Συνδρομητή (*κωδικός_συνδρομητή*, *αριθμός_τηλεφώνου*)
Γνωστική_Περιοχή (*κωδικός*, *τίτλος*, *αριθμός_συνδρομητών*)
Συνδρομή (*κωδικός_συνδρομητή*, *κωδικός_γνωστικής_περιοχής*, *από*, *έως*)
Συγγραφέας (*κωδικός*, *όνομα*, *οδός*, *αριθμός*, *ΤΚ*, *πόλη*, *χώρα*, *σύνολο_άρθρων*)
Τηλέφωνο_Συγγραφέα (*κωδικός_συγγραφέα*, *αριθμός_τηλεφώνου*)
Άρθρο (*κωδικός*, *τίτλος*, *PDF*, *κωδικός_γνωστικής_περιοχής*, *κωδικός_συνεδρίου*, *κωδικός_περιοδικού*, *αρχική_σελίδα_πρακτικών*, *τελική_σελίδα_πρακτικών*, *τεύχος*, *τόμος*, *αρχική_σελίδα_περιοδικού*, *τελική_σελίδα_περιοδικού*)
Συγγραφή_Άρθρου (*κωδικός_συγγραφέα*, *κωδικός_άρθρου*)
Εκδοτικός_Οίκος (*κωδικός*, *όνομα*, *οδός*, *αριθμός*, *ΤΚ*, *πόλη*, *χώρα*)
Τηλέφωνο_Εκδοτικού_Οίκου (*κωδικός_εκδοτικού_οίκου*, *αριθμός_τηλεφώνου*)
Περιοδικό (*κωδικός*, *τίτλος*, *κωδικός_εκδοτικού_οίκου*)
Πρακτικά_Συνεδρίου (*κωδικός*, *συνέδριο*, *πόλη*, *χώρα*, *ημερομηνία*, *κωδικός_εκδοτικού_οίκου*)

Για τα παρακάτω ερωτήματα SQL να απαντήσετε τι ακριβώς κάνουν και τι επιστρέφουν περιγράφοντάς τα:

```
SELECT  Συνδρομητής.κωδικός, Συνδρομητής.όνομα, Γνωστική_Περιοχή.τίτλος
FROM  Συνδρομητής, Γνωστική_Περιοχή, Συνδρομή
WHERE  Συνδρομή.κωδικός_συνδρομητή = Συνδρομητής.κωδικός
AND     Γνωστική_Περιοχή.κωδικός = Συνδρομή.κωδικός_γνωστικής_περιοχής;
```

```
SELECT  συνέδριο
FROM   Πρακτικά_Συνεδρίου
UNION
SELECT  τίτλος
FROM   Περιοδικό;
```

```
SELECT  τίτλος, AVG (αριθμός_συνδρομητών)
FROM   Γνωστική Περιοχή
GROUP BY  τίτλος ;
```

```
SELECT  τίτλος
FROM   Γνωστική_Περιοχή
WHERE  κωδικός  NOT IN
        ( SELECT  κωδικός_γνωστικής_περιοχής FROM   Άρθρο );
```

```
SELECT  τίτλος
FROM  Γνωστική_Περιοχή
WHERE  αριθμός_συνδρομητών > ALL
        ( SELECT  αριθμός_συνδρομητών  FROM   Γνωστική_Περιοχή );
```

19. Δίνεται το σχήμα μίας βάσης δεδομένων με συνδρομητικά δεδομένα όπως παραπάνω. Για τα παρακάτω ερωτήματα SQL να απαντήσετε τι ακριβώς κάνουν και τι επιστρέφουν περιγράφοντάς τα:

```
SELECT  Άρθρο.τίτλος, Συγγραφέας.όνομα
FROM     Άρθρο, Συγγραφέας, Συγγραφή_Άρθρου
WHERE  Συγγραφή_Άρθρου.κωδικός_άρθρου = Άρθρο.κωδικός
AND  Συγγραφέας.κωδικός = Συγγραφή_Άρθρου.κωδικός_συγγραφέα
ORDER BY  Άρθρο.τίτλος;
```

```
SELECT  όνομα
FROM   Συνδρομητής
INTERSECT
SELECT  όνομα
FROM   Συγγραφέας;
```

```
SELECT  *
FROM   Συνδρομητής
WHERE  όνομα  LIKE  '%όπουλος';
```

```
SELECT  κωδικός_συγγραφέα
FROM   Συγγραφή_Άρθρου AS ΣΑ
WHERE EXISTS
   (SELECT * FROM   Άρθρο
         WHERE  κωδικός = ΣΑ.κωδικός_άρθρου  AND
         κωδικός_συνεδρίου  IS NOT NULL AND κωδικός_περιοδικού  IS NOT NULL );
```

```
SELECT κωδικός_συγγραφέα, COUNT(κωδικός_άρθρου)  AS  αριθμός_άρθρων
FROM   Συγγραφή_Άρθρου
GROUP BY  κωδικός_συγγραφέα
HAVING  αριθμός_άρθρων > ( SELECT  COUNT(*)/10  FROM   Συγγραφή_Άρθρου );
```

20. Δίνεται το σχήμα μίας βάσης δεδομένων με συνδρομητικά δεδομένα όπως παραπάνω. Για τα παρακάτω ερωτήματα SQL να απαντήσετε τι ακριβώς κάνουν και τι επιστρέφουν περιγράφοντάς τα:

```
SELECT  Άρθρο.τίτλος, Γνωστική Περιοχή.τίτλος
FROM     Άρθρο  JOIN  Γνωστική Περιοχή  ON
         Γνωστική_Περιοχή.κωδικός = Άρθρο.κωδικός_γνωστικής_περιοχής;
```

```
SELECT  Συγγραφέας.όνομα,
COUNT (Συγγραφή_Άρθρου.κωδικός_άρθρου)
FROM  Συγγραφέας, Συγγραφή_Άρθρου
WHERE  Συγγραφή_Άρθρου.κωδικός_συγγραφέα =  Συγγραφέας.κωδικός
GROUP  BY  Συγγραφέας.όνομα;
```

```
SELECT  τίτλος
FROM   Γνωστική_Περιοχή
WHERE  αριθμός_συνδρομητών  >
         ( SELECT AVG(αριθμός_συνδρομητών) FROM Γνωστική_Περιοχή );
```

```
SELECT  τίτλος
FROM   Γνωστική_Περιοχή
WHERE  αριθμός_συνδρομητών > SOME
         ( SELECT  αριθμός_συνδρομητών  FROM   Γνωστική_Περιοχή );
```

```
SELECT  τίτλος,  AVG (αριθμός_ συνδρομητών)  AS μέσος_αριθμός_συνδρομητών
FROM   Γνωστική Περιοχή
GROUP  BY  τίτλος;
```

21. Τι ονομάζεται Σύστημα Επεξεργασίας Δοσοληψιών (On-Line Transaction Processing ή OLTP); Ποια είναι τα βασικά του χαρακτηριστικά; Ποια είναι τα προβλήματα που εμφανίζονται σε ένα σύστημα OLTP;

22. Τι ονομάζεται Σύστημα On-Line Analytical Processing ή OLAP; Ποια είναι τα λειτουργικά χαρακτηριστικά των απαιτήσεων ενός συστήματος OLAP; Ποιες είναι οι διαφορές μεταξύ ενός συστήματος OLTP και ενός OLAP;

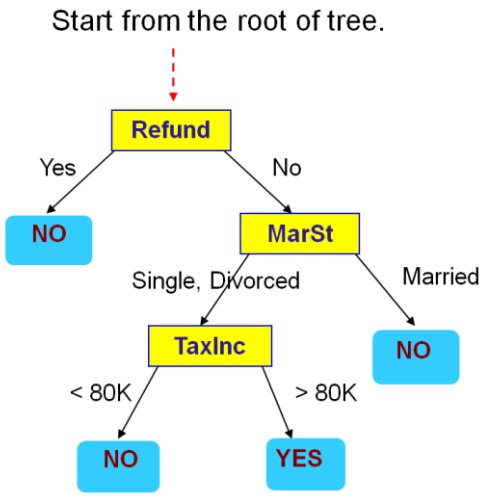
23. Τι αποτελεί μία Αποθήκη Δεδομένων; Ποια είναι τα λειτουργικά χαρακτηριστικά μίας αποθήκης δεδομένων; Ποιες είναι οι βασικές λειτουργίες μίας αποθήκης δεδομένων; Ποια είναι τα βασικά επίπεδα-περιοχές της αρχιτεκτονικής μίας αποθήκης δεδομένων;

24. Τι λέγεται Κύβος Δεδομένων; Πως ορίζονται οι διαστάσεις και οι μετρήσεις του; Τι είναι οι ιεραρχίες των διαστάσεων; Με ποια μοντέλα γίνεται η αναπαράσταση ενός κύβου δεδομένων; Ποια είναι τα βασικά χαρακτηριστικά κάθε μοντέλου;

25. Τι λέγεται Βασικός Κυβοειδής και τι περιέχει; Τι είναι το Πλέγμα όλων των Κυβοειδών; Ποιες είναι οι βασικές συναθροιστικές συναρτήσεις που εφαρμόζονται σε έναν κύβο δεδομένων; Ποιες είναι οι βασικές πράξεις που μπορούμε να κάνουμε σε έναν κύβο δεδομένων;

26. Ποια διαδικασία ονομάζεται Κατηγοριοποίηση (Classification) και τι στόχο έχει; Τι είναι ένα δέντρο απόφασης (Decision Tree) και με ποια γενική διαδικασία παράγεται και εφαρμόζεται;

Δεξιά απεικονίζεται ένα δέντρο απόφασης που έχει παραχθεί από ένα σύνολο εγγραφών. Έχουμε και μία νέα εγγραφή (test data) στην οποία δεν έχει οριστεί η κατηγορία Cheat και για να την προσδιορίσουμε εφαρμόζουμε το μοντέλο του δέντρου απόφασης. Ποια πορεία θα διασχίσει στο δέντρο η νέα εγγραφή και τι τιμή θα προκύψει τελικά στην κατηγορία Cheat;



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

27. Αναφέρετε μερικούς αλγορίθμους παραγωγής δέντρων απόφασης. Περιγράψτε τα βασικά βήματα του αλγορίθμου του Hunt.

28. Από τι εξαρτάται ο διαχωρισμός (split) κατά την παραγωγή ενός δέντρου απόφασης; Πως γίνεται ο διαχωρισμός σε διακριτά (π.χ. λεκτικά) χαρακτηριστικά και πως σε συνεχή;

29. Ποια τρία βασικά μέτρα χρησιμοποιούνται για να επιτευχθεί βέλτιστος διαχωρισμός (split) σε ένα δέντρο απόφασης; Ποια είναι η μέγιστη και ποια η ελάχιστη τιμή που παίρνουν και πότε συμβαίνει αυτό; Το Κέρδος Πληροφορίας (Information Gain) από ποιο βασικό μέτρο ορίζεται, τι μετράει, σε ποιους χαρακτηριστικούς αλγορίθμους εφαρμόζεται και ποιο είναι το βασικό μειονέκτημά του; Με βάση ποιο άλλο μέτρο ξεπεράστηκε το μειονέκτημα αυτό;

30. Ποια είναι τα κριτήρια τερματισμού (κοινά και πρόωρης διακοπής) της παραγωγής ενός δέντρου απόφασης; Ποια είναι τα πλεονεκτήματα των δέντρων απόφασης σε σύγκριση με άλλες μεθόδους κατηγοριοποίησης; Ποιος είναι ο κανόνας του Occam (Occam's Razor);

31. Ποια διαδικασία ονομάζεται εξόρυξη Κανόνων Συσχέτισης (Association Rules) και τι στόχο έχει; Πως ορίζεται ένας κανόνας συσχέτισης; Πως ορίζονται τα μέτρα Support και Confidence ενός κανόνα συσχέτισης και τι εκφράζουν; Να υπολογίσετε

το support και το confidence των παρακάτω κανόνων που βρίσκονται αριστερά με βάση της εγγραφές/συναλλαγές που βρίσκονται δεξιά. Τι παρατηρείτε;

{Milk,Diaper} → {Beer}
{Milk,Beer} → {Diaper}
{Milk} → {Diaper,Beer}

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

32. Γιατί η παραγωγή των συχνών στοιχειοσυνόλων δεν μπορεί να γίνει με την brute-force μέθοδο; Ποιες είναι οι στρατηγικές παραγωγής συχνών στοιχειοσυνόλων; Ποιος είναι ο κανόνας Apriori και σε ποια ιδιότητα του support βασίζεται; Που βοηθάει η χρήση ενός δέντρου κατακερματισμού (Hash Tree);

33. Εφαρμόστε τον αλγόριθμο Apriori για την παραγωγή συχνών στοιχειοσυνόλων με ένα, δύο και τρία στοιχεία στο παρακάτω σύνολο εγγραφών/συναλλαγών αν το όριο για το ελάχιστο support είναι 2 (δηλαδή support>1):

Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Κάνοντας δυαδική διαμέριση στο μοναδικό συχνό στοιχειοσύνολο με τρία στοιχεία που παράγεται, καταγράψτε τους αντίστοιχους κανόνες συσχέτισης.

34. Τι ονομάζεται Μέγιστο Στοιχειοσύνολο (Maximal Itemset); Τι ονομάζεται Κλειστό Στοιχειοσύνολο (Closed Itemset); Ποια σχέση συνόλων συνδέει τα συχνά στοιχειοσύνολα, τα κλειστά και τα μέγιστα; Ποια δομή χρησιμοποιεί ο αλγόριθμος FP-growth και ποια βασική στρατηγική εφαρμόζει; Ποια δομή χρησιμοποιεί ο αλγόριθμος ECLAT, ποιο βασικό πλεονέκτημα και ποιο μειονέκτημα έχει;

35. Ποια διαδικασία ονομάζεται Ομαδοποίηση (Clustering) και τι στόχο έχει; Ποιοι δύο βασικοί τύποι ομαδοποίησης υπάρχουν; Ποιοι είναι οι βασικοί τύποι ομάδων και πως ορίζεται η έννοια της ομάδας στην κάθε περίπτωση;

36. Περιγράψτε τα βήματα του αλγορίθμου ομαδοποίησης K-means. Πώς επιλέγονται τα αρχικά centroids και πως υπολογίζονται σε κάθε επανάληψη; Ποια μέτρα εγγύτητας συνήθως εφαρμόζονται ώστε ο αλγόριθμος να συγκλίνει; Ποια είναι η χρονική πολυπλοκότητα του αλγορίθμου αυτού;

37. Τι είναι και τι παράγει η Ιεραρχική Ομαδοποίηση (Hierarchical Clustering); Ποια είναι τα πλεονεκτήματά της; Ποιοι βασικοί τύποι ιεραρχικής ομαδοποίησης υπάρχουν; Περιγράψτε τα βήματα του Αλγορίθμου Ομαδοποίησης Επικόλλησης. Με ποιους τρόπους (απλή αναφορά) μπορεί να οριστεί η ομοιότητα μεταξύ ομάδων στον αλγόριθμο αυτό; Ποια είναι η χωρική και η χρονική πολυπλοκότητά του;

38. Που βασίζεται ο αλγόριθμος DBSCAN; Τι λέγεται σημείο πυρήνας (core point), τι λέγεται συνοριακό σημείο (border point) και τι σημείο θορύβου (noise point); Περιγράψτε τα βήματα του αλγορίθμου DBSCAN. Ποια είναι τα πλεονεκτήματα και ποια τα μειονεκτήματά του;