

D'accord, je vais répondre à toutes les questions contenues dans le document.

Question 1.a

****What are the different ways to acquire data for the purpose of data analysis?****

Les différentes façons d'obtenir des données pour l'analyse incluent :

1. ****Enquêtes et sondages**** : Collecte directe via des questionnaires en ligne ou physiques.
2. ****Bases de données internes**** : Données générées par les opérations internes d'une entreprise.
3. ****Données publiques**** : Données accessibles librement, comme les données gouvernementales ou les ensembles de données ouverts.
4. ****Réseaux sociaux**** : Données extraites des interactions sur les réseaux sociaux.
5. ****Capteurs et IoT**** : Données collectées par des capteurs et des dispositifs connectés.
6. ****Web scraping**** : Extraction de données à partir de sites web.
7. ****API**** : Utilisation d'interfaces de programmation pour accéder à des données externes.

Question 1.b

****Why do you think that manual or face to face surveys are still important?****

Les enquêtes en face à face sont toujours importantes pour plusieurs raisons :

1. ****Interaction humaine**** : Permet une meilleure compréhension des réponses grâce à l'interaction directe.
2. ****Taux de réponse**** : Souvent plus élevé car les enquêteurs peuvent encourager la participation.
3. ****Clarté des questions**** : Les enquêteurs peuvent clarifier les questions ou les réponses immédiatement.
4. ****Données qualitatives**** : Permet de recueillir des informations plus détaillées et nuancées.
5. ****Accessibilité**** : Peut atteindre des populations qui n'ont pas accès à Internet.

Question 2.a

****What are ACID constraints? Which of these constraints were relaxed by NoSQL data stores and why?****

Les contraintes ACID sont :

1. ****Atomicité**** : Chaque transaction est traitée comme une unité indivisible.
2. ****Cohérence**** : Les transactions respectent les règles de validation des données.
3. ****Isolation**** : Les transactions sont isolées les unes des autres.
4. ****Durabilité**** : Une fois une transaction validée, elle est permanente.

Les bases de données NoSQL relaxent souvent la ****cohérence**** et parfois l'****isolation**** pour favoriser la scalabilité et la disponibilité, en suivant le théorème CAP (Cohérence, Disponibilité, Tolérance aux pannes).

Question 2.b

****What are the different types of NoSQL data stores? Briefly explain them.****

Les différents types de bases de données NoSQL sont :

1. ****Bases de données orientées colonnes**** : Stockent les données par colonnes plutôt que par lignes, optimisant les lectures de grandes quantités de données.
2. ****Bases de données orientées documents**** : Stockent les données sous forme de documents (JSON, XML), offrant flexibilité et évolutivité.
3. ****Bases de données orientées graphes**** : Utilisent des nœuds et des relations pour représenter les données, idéales pour les relations complexes.
4. ****Bases de données clé-valeur**** : Stockent les données sous forme de paires clé-valeur, simples

et rapides pour les opérations de lecture/écriture.

Question 3

****What is data visualization? Why data visualization methods are needed? Explain with some examples.****

La visualisation de données est la représentation graphique des données pour en faciliter la compréhension. Elle est nécessaire pour :

1. ****Identifier des tendances**** : Par exemple, un graphique en ligne peut montrer l'évolution des ventes au fil du temps.
2. ****Comparer des données**** : Les diagrammes à barres permettent de comparer facilement des catégories.
3. ****Détecter des anomalies**** : Les nuages de points peuvent révéler des valeurs aberrantes.
4. ****Communiquer des insights**** : Les cartes thermiques peuvent montrer des zones de forte densité de données.

Question 4

****What are the goals of Data Mining?****

Les objectifs du Data Mining incluent :

1. ****Extraction de motifs**** : Identifier des motifs récurrents dans les données.
2. ****Classification**** : Catégoriser les données en groupes prédéfinis.
3. ****Régression**** : Prédire des valeurs continues basées sur des variables d'entrée.
4. ****Clustering**** : Regrouper des données similaires sans catégories prédéfinies.
5. ****Détection d'anomalies**** : Identifier des valeurs ou comportements inhabituels.
6. ****Analyse de règles d'association**** : Découvrir des relations entre variables.

Question 5.a

****What is a classifier? What are the different types of classifier?****

Un classificateur est un algorithme qui attribue des étiquettes ou des catégories à des données d'entrée. Les types de classificateurs incluent :

1. ****Classificateurs linéaires**** : Par exemple, régression logistique.
2. ****Classificateurs basés sur des arbres**** : Arbres de décision, forêts aléatoires.
3. ****Classificateurs basés sur des réseaux de neurones**** : Perceptrons multicouches.
4. ****Classificateurs basés sur des distances**** : K-plus proches voisins.
5. ****Classificateurs bayésiens**** : Naïve Bayes.

Question 5.b

****How do you evaluate and compare the efficiency of a classifiers?****

L'efficacité des classificateurs peut être évaluée par :

1. ****Précision**** : Proportion de vrais positifs parmi les prédictions positives.
2. ****Rappel**** : Proportion de vrais positifs parmi les réels positifs.
3. ****F1-score**** : Moyenne harmonique de la précision et du rappel.
4. ****Matrice de confusion**** : Tableau montrant les vrais positifs, faux positifs, vrais négatifs, et faux négatifs.
5. ****Courbe ROC et AUC**** : Mesure de la capacité du classificateur à distinguer les classes.

Question 5.c

****What are the different clustering methods that you used in your practical sessions? What are their advantages and limitations?****

Les méthodes de clustering couramment utilisées incluent :

1. **K-means** : Avantages : Simple et rapide. Limitations : Sensible à l'initialisation, nécessite de connaître le nombre de clusters.
2. **DBSCAN** : Avantages : Peut trouver des clusters de forme arbitraire. Limitations : Sensible aux paramètres de distance.
3. **Clustering hiérarchique** : Avantages : Ne nécessite pas de spécifier le nombre de clusters. Limitations : Peut être coûteux en calcul.
4. **Clustering par partitionnement spectral** : Avantages : Utilise des informations sur la structure des données. Limitations : Complexité computationnelle élevée.

Question 6

Consider a CSV file containing the following columns: PhotographId, City, Year, and ViewCount. Write a Python program to:

1. **Find the most viewed and least viewed photograph**
2. **Find the city with the maximum and least number of photographs**
3. **Find the year with the highest number of photographs**
4. **For every city, calculate the average number of views for photographs in the year 2018**

Voici un exemple de programme Python utilisant `pandas` :

```
```python
import pandas as pd

Lire le fichier CSV
df = pd.read_csv('photographs.csv')

1. Photo la plus vue et la moins vue
most_viewed = df.loc[df['ViewCount'].idxmax()]
least_viewed = df.loc[df['ViewCount'].idxmin()]

2. Ville avec le plus grand et le plus petit nombre de photos
city_counts = df['City'].value_counts()
max_city = city_counts.idxmax()
min_city = city_counts.idxmin()

3. Année avec le plus grand nombre de photos
year_counts = df['Year'].value_counts()
max_year = year_counts.idxmax()

4. Nombre moyen de vues par ville en 2018
avg_views_2018 = df[df['Year'] == 2018].groupby('City')['ViewCount'].mean()

print("Most viewed photograph:", most_viewed)
print("Least viewed photograph:", least_viewed)
print("City with most photographs:", max_city)
print("City with least photographs:", min_city)
print("Year with most photographs:", max_year)
print("Average views per city in 2018:\n", avg_views_2018)
```
```

Question 7.a

****What is an artificial neural network?****

Un réseau de neurones artificiel est un modèle informatique inspiré du cerveau humain, composé de couches de neurones interconnectés. Chaque neurone reçoit des entrées, les pondère, applique une fonction d'activation, et transmet le résultat aux neurones suivants. Les réseaux de neurones sont utilisés pour des tâches comme la reconnaissance d'images, le traitement du langage naturel, et la prédiction.

Question 7.b

****Why do you think reinforcement learning is relevant for internal and outdoor navigation by robots?****

L'apprentissage par renforcement est pertinent pour la navigation des robots car il permet aux robots d'apprendre par essai et erreur, en recevant des récompenses pour des actions qui les rapprochent de leur objectif. Cela est particulièrement utile dans des environnements dynamiques ou inconnus, où les robots doivent s'adapter et prendre des décisions en temps réel pour éviter les obstacles et atteindre leurs destinations.

Question 8

****Find all possible association rules from this table. What do you conclude about this picture?****

Pour trouver les règles d'association, nous devons identifier les combinaisons fréquentes de hashtags utilisés par les utilisateurs. Voici quelques règles possibles :

1. ****{#Architecture} → {#StreetArt}**** : Les utilisateurs qui utilisent #Architecture utilisent souvent #StreetArt.
2. ****{#Nature} → {#Paris, #StreetArt}**** : Les utilisateurs qui utilisent #Nature utilisent souvent #Paris et #StreetArt.
3. ****{#Paris} → {#StreetArt}**** : Les utilisateurs qui utilisent #Paris utilisent souvent #StreetArt.

****Conclusion sur l'image**** : L'image semble être liée à l'architecture et à l'art de rue, probablement située à Paris, avec des éléments de nature et de motifs fractals.

Si vous avez besoin de plus de détails ou d'explications supplémentaires, n'hésitez pas à demander !