# Aminata oumar n’diaye 22108

# Résumé du cours datamining

Lors du premier cours nous avons abordé le concept de Datamining ainsi que son objectif.

Lors du deuxième cours, nous avons fait un td

1-definir le DM

1. Data Mining : Technique d’analyse pour extraire des connaissances à partir de grandes quantités de données.

2- relation entre :

DM et Big data :

Le Big Data fournit la matière première (données massives), tandis que le data mining est l’outil permettant de transformer ces données en connaissances exploitables.

Le data mining, appliqué au Big Data, nécessite des outils avancés pour gérer la complexité et la taille des données, comme des algorithmes d’apprentissage automatique ou des infrastructures de traitement distribuées .

3 –DM et Analyse de donnée :

Le data mining est une approche spécialisée et automatisée incluse dans le processus plus large de l'analyse de données.

4-DM et IA :

le data mining est un outil clé qui fournit des bases pour le développement et l'entraînement des systèmes IA. L’IA, en retour, pousse le data mining vers des approches plus performantes et dynamiques.

5-DM et IR :

L’IR est orientée recherche (trouver les bons documents ou informations), tandis que le data mining est orienté découverte (extraire des connaissances à partir des données).

6- HDD : Un HDD (Hard Disk Drive) est un type de disque dur utilisé pour le stockage de données dans les ordinateurs et autres appareils électroniques.

Caractéristiques principales :

1. Structure :

Composé de plateaux métalliques recouverts de matériau magnétique.

Une tête de lecture/écriture se déplace au-dessus des plateaux pour accéder aux données.

2. Fonctionnement :

Les données sont stockées de manière magnétique sur les plateaux.

Les plateaux tournent à haute vitesse, et la tête lit ou écrit les données en fonction des instructions.

3. Capacité :

Généralement, les HDD offrent une grande capacité de stockage (jusqu’à plusieurs téraoctets) à un coût relativement faible.

4. Performance :

Plus lent que les SSD (Solid State Drive), car il repose sur des composants mécaniques.

Vitesse moyenne de rotation : 5400 RPM ou 7200 RPM (rotations par minute), affectant la rapidité d'accès.

5. Durabilité :

Sensible aux chocs et aux vibrations à cause de ses composants mécaniques.

Avantages :

Grande capacité de stockage pour un prix abordable.

Idéal pour stocker de gros volumes de données (documents, films, archives).

Inconvénients :

Moins rapide que les SSD.

Plus lourd et moins résistant aux chocs.

Consomme plus d’énergie en fonctionnement.

Applications courantes :

Stockage dans les ordinateurs de bureau ou portables.

Utilisé dans les serveurs pour les archives de données.

Solutions de stockage externe pour sauvegardes.

Lors du troisième cours, nous avons étudié les trois catégories de langages de programmation :

1. Langages de bas niveau :

Proches du matériel informatique, comme le langage machine et l'assembleur.

Ils offrent un contrôle précis sur le matériel, mais sont difficiles à comprendre et à utiliser pour les humains.

2. Langages de moyen niveau :

Intermédiaires entre les langages de bas et haut niveau, comme le langage C.

Ils permettent un certain contrôle sur le matériel tout en restant relativement lisibles.

3. Langages de haut niveau :

Conçus pour être compréhensibles et simples à utiliser, comme Python, Java ou PHP.

Apres on as fait des exercice sur les listes , les boucle…..

Voici le code Python pour chacun des trois exercices :

1. Calculatrice Simple

def calculatrice():

try:

num1 = float(input("Entrez le premier nombre : "))

num2 = float(input("Entrez le deuxième nombre : "))

operation = input("Choisissez une opération (+, -, \*, /) : ")

if operation == "+":

resultat = num1 + num2

elif operation == "-":

resultat = num1 - num2

elif operation == "\*":

resultat = num1 \* num2

elif operation == "/":

if num2 != 0:

resultat = num1 / num2

else:

return "Erreur : Division par zéro."

else:

return "Opération non valide."

return f"Résultat : {resultat}"

except ValueError:

return "Erreur : Veuillez entrer des nombres valides."

print(calculatrice())

Le quatrième cours on as parlé de l’importance de pyhon dans la science de données :

I 1. Langage Polyvalent : Python est facile à apprendre et offre une syntaxe intuitive, idéale pour les débutants comme pour les experts.

2. Écosystème Riche : Bibliothèques populaires comme Pandas, NumPy, Matplotlib, Scikit-learn et TensorFlow pour l’analyse, la visualisation et l’apprentissage automatique.

3. Communauté Active : Large support communautaire avec des forums, des tutoriels et des ressources pour résoudre les problèmes.

4. Intégration Facile : S'intègre avec d'autres technologies et langages pour un traitement des données à grande échelle.

5. Utilisation Industrielle : Standard dans les entreprises pour l’analyse des données, le Big Data et l’intelligence artificielle.

6. Puissance et Flexibilité : Gestion efficace des tâches complexes comme la manipulation de données massives, la modélisation et la visualisation interactive.

Ensemble du pipeline d'exploration de données

Collecte de données → proprecessing→transformation→datamining→évaluation ,interprétation→validation

Utilisation de Pandas pour la collecte depuis des fichiers (CSV, Excel).

La collecte de données est essentielle pour obtenir des insights et prendre des décisions éclairées.

- Sources possibles : fichiers, bases de données, sites web, APIs.

- Objectif : Apprendre à utiliser des outils Python pour collecter des données efficacement

Sources principales de données

Fichiers traditionnels : CSV, TXT, XLSX, JSON, XML, HTML

Bases de données : SQL, NoSQL.

Web : Sites web, APIs.

Outils nécessaires :

Frameworks de scraping web.

Connecteurs de bases de données.

Librairies de manipulation de fichiers.

Outils Python pour la collecte de données :

Pandas : Manipulation et analyse des données.

Lecture des fichiers CSV, Excel, SQL.

BeautifulSoup : Analyse HTML et XML pour le scraping web.

Requests : Récupération de données via APIs et services web.

Connecteurs SQL : mysql-connector-python, psycopg2, sqlite3.

Yahoo Finance : Extraction de données financières (prix, historiques).

Pourquoi utiliser différents formats de fichiers ?

Avantages : Compatibilité avec divers outils et applications.

Formats populaires et outils Python associés : CSV : Format simple et largement supporté (pandas.read\_csv()).

TXT : Format brut pour textes simples (pandas.read\_csv() avec délimiteurs personnalisés).

XLSX : Données complexes avec feuilles multiples (pandas.read\_excel())

Les règles d’association :

Les trois fonctions pour le règle d’association

Support

Confiance

Left

On a pris un exemple d’analyse de panier d’achat

|  |  |
| --- | --- |
| Transaction | Article |
| 1 | Pain, lait |
| 2 | Pain, beurre, eouf, jus |
| 3 | Lait, jus, beure, thé |
| 4 | Pain, lait, jus, beurre |
| 5 | Pain, lait, jus, thé |

{pain, lait} →{beurre}

{pain} →{lait}

{lait} →{pain}

Support=nombre de transaction de x/total

|  |  |  |  |
| --- | --- | --- | --- |
| élément | transaction | support | % |
| pain | 1,2,4 ,5 | 4/5 | 80% |
| lait | 1,3,4,5 | 4/5 | 80% |
| beurre | 2,3,4 | 3/5 | 60% |
| jus | 2,3,4,5 | 4/5 | 80% |
| thé | 3,5 | 2/5 | 40% |
| eouf | 2 | 1/5 | 20% |
| Pain , lait | 1,4,5 | 3/5 | 60% |
| Pain ,lait ,thé | 5 | 1/5 | 20% |

Confiance (x→Y)= S(xvy)/s(x)

C(lvp)=0.6/0.8=75%