Wiam Beidede IUP22186

Résume des cours Data mining

- Le data mining est le processus d'analyse de grandes quantités de données pour identifier des motifs, tendances ou relations utiles. Il est utilisé pour extraire des connaissances exploitables dans divers domaines comme la finance, le marketing ou la santé.
- Le Big Data désigne un ensemble massif de données variées, générées à grande vitesse et en volumes importants, qui nécessitent des outils avancés pour leur collecte, stockage, traitement et analyse. Il est utilisé pour découvrir des insights et faciliter la prise de décision dans des domaines tels que la santé, la finance et le marketing.
- L'analyse de données consiste à examiner et interpréter des données pour en extraire des informations utiles, en utilisant des outils statistiques et des techniques analytiques,Elle est essentielle pour la prise de décision et l'identification de tendances.
- L'intelligence artificielle (IA) désigne la simulation de processus cognitifs humains par des machines, en particulier les ordinateurs. Elle inclut des domaines comme l'apprentissage automatique, le traitement du langage naturel et la vision par ordinateur, permettant aux systèmes de résoudre des problèmes, apprendre et s'adapter de manière autonome.

    - Relation entre Big Data et Data Mining : Le Big Data offre une grande quantité de données brutes, tandis que le Data Mining les exploite pour en extraire des connaissances et des informations pertinentes.
    - Relation entre Data Mining et Analyse des données : Le Data Mining est un sous-ensemble de l'analyse des données, spécialisé dans l'identification de motifs ou tendances cachés.
    - Relation entre Data Mining et Intelligence Artificielle : Le Data Mining applique des algorithmes d'IA pour analyser les données et en extraire des informations exploitables, contribuant ainsi à améliorer les performances des systèmes d'IA.
    - Techniques de Data Mining pour l'exploration des données : Les méthodes courantes incluent la classification, le clustering, l'analyse d'association et la régression.

_Type de modèle :

1.Description de concepts / classification : Identifier les caractéristiques des objets.
2.Analyse d'association : Découvrir les relations entre éléments.
3.Classification et prédiction : Catégoriser et prédire.
4.Analyse de clustering : Regrouper selon similitude.
5.Analyse d'évolution et de déviation : Suivre les changements dans le temps.

_Pipeline d'exploration de données :

1.Sélection : Choix des données pertinentes dans un ensemble de données plus large.
2.Prétraitement : Préparation des données sélectionnées pour l'analyse.
3.Transformation : Conversion des données prétraitées dans un format adapté au Data Mining.
4.Data Mining : Analyse des données transformées pour découvrir des motifs.
5.Interprétation/Évaluation : Interprétation et évaluation des motifs découverts pour extraire des
connaissances significative .

_L’importance de Python dans Data mining :

Python est essentiel en Data Mining grâce à sa simplicité, ses nombreuses bibliothèques (Pandas, NumPy, Scikit-learn) et ses capacités d’automatisation, rendant l’analyse des données rapide et efficace.

_Collecte de données avec Python :

Python permet d'extraire des données provenant de diverses sources, notamment :

- Fichiers classiques : Formats comme CSV, TXT, Excel (XLSX), JSON, XML, ou HTML.
- Bases de données : Compatibilité avec les systèmes relationnels (SQL) et non relationnels (NoSQL).
- Données en ligne : Extraction depuis des sites web via le web scraping ou l'utilisation d'APIs.

_Formats populaires et outils Python associés :

- CSV : Format simple et largement compatible (utilisation de pandas.read_csv()).
- TXT : Format brut pour textes simples, souvent avec des délimiteurs personnalisés (via pandas.read_csv()).
- XLSX : Pour des données complexes avec plusieurs feuilles (chargement avec pandas.read_excel()).
- JSON : Idéal pour des données structurées et faciles à manipuler (via pandas.read_json()).
- XML : Utilisé pour des données hiérarchiques et des structures imbriquées (avec pandas.read_xml()).
- HTML : Extraction de tableaux depuis des pages web (fonction pandas.read_html()).

Utiliser Pandas pour lire un fichier CSV:

```
import pandas as pd
df = pd.read_csv('data.csv’)
print(df.head())
```

Utiliser Pandas pour lire un fichier Xml:

```
import pandas as pd
df = pd.read_xml('data.xml’)
print(df.head())
```

Utiliser Pandas pour lire un fichier Html:

```
import pandas as pd
tables= pd.read_html('https://aws.amazon.com’)
print(tables[0])
```

Utiliser Pandas pour lire un fichier JSON:

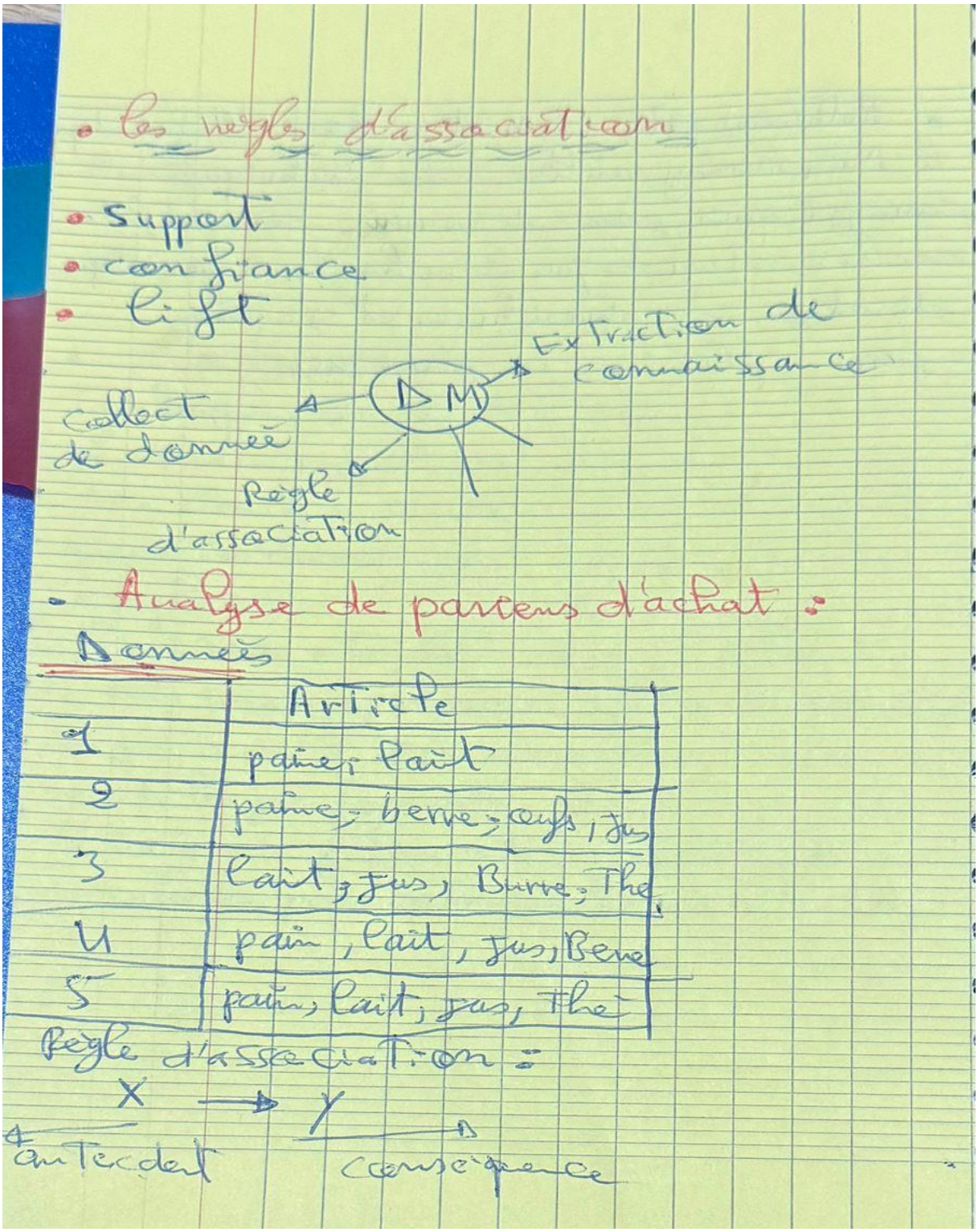
```
importjson
with open(‘data/data.Json ', ‘r’) as f:
data = json.load(f)
print (data)
```

Connexion aux bases de données SQL:

```
import sqlite3
conn = sqlite3.connect('database.db')
query = "SELECT * FROM table_name"
df = pd.read_sql_query(query, conn)
import sqlite3
```

Règle D’association :

L'objectif est de définir une association et de trouver la relation cachée qui existe entre les éléments .



- Support = Soutien = $\frac{\text{nombre Transaction X}}{\text{Total}}$
- Confiance $(X \Rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$

{pain} → {lait}

Element	T	Support	~~80%~~
pain	1, 2, 4, 5	4/5	80%
Lait	1, 3, 4, 5	4/5	80%
bière	2, 3, 4	3/5	60%
Jus	2, 3, 4, 5	4/5	80%
Thé	3, 5	2/5	40%
Œufs	1	1/5	20%
pain, Lait	1, 4, 5	3/5	60%
pain, Lait, Thé	1	1/5	20%

$C = \frac{S(P,L,T)}{S(P,L)} = \frac{0,2}{0,6} = 33,33\%$

? {pain, lait, jus} → {Thé}

$S(P,L,J) = 0,4$ $\Rightarrow C = \frac{0,2}{0,4} = 50\%$

$S(P,L,J,T) = 0,2$