

Nom: Khatou Sadatt
ID: 22154

Introduction Sur le DM

- **Def (DM)**: il consiste à analyser de grands ensembles de données pour y trouver des motifs et tendances utiles.
- **Def (le Big Data)**: c'est une grande quantité de données complexes qui nécessitent des outils spéciaux pour être stockés, traités et analysés.
- **la relation entre les deux**: Le big data fournit les données et les outils, tandis que data mining utilise des techniques pour extraire des connaissances.
- **la relation entre DM et analyse de données**:
L'analyse de données utilise des méthodes statistiques pour comprendre les données, tandis que le data mining extrait des modèles cachés à l'aide d'algorithmes avancés.
- **la relation entre DM et IA**
Le DM utilise des techniques d'IA, comme l'apprentissage automatique pour extraire des modèles cachés et améliorer l'analyse de données.
- **les méthodes et techniques de DM pour l'exploration des données**:
 - 1 - classification
 - 2 - clustering
 - 3 - association

Python

les langage du programmation

Interprétée

n'a pas besoin de l'exécuter à nouveau pour avoir les modifications

Compilé

on a besoin d'exécuter chaque fois on a changé le code

Python: c'est un langage de programmation interprétée et très avancé, de haut niveau et. se tend adapté à l'analyse de donnée et ou data mining

Base de programmation

{ Variables
Expressions
Instructions
structures de contrôle
boucles
Fonctions

P.O.O

{ Les classes
Les objets
encapsulation
Polymorphisme
héritage

• Déclaration de Variable :

x = 51 # entier

y = 3.14 # flottant

x = "Alice" # caractère

NB : on a changé la valeur de x

- en python on a pas besoin de spécifier le type du char il sera automatiquement identifié

NB:

les entier prend ~~pas~~ 4 octet

„ float \Rightarrow 4 octet

„ string \Rightarrow Chaque caractère un octet

les structure composée

list: liste \Rightarrow modifiable

tuple: tuple \Rightarrow Immodifiable

dict: dictionnaire \Rightarrow dictionnaire

Set: ensemble: ses element peut pas être repeter

NB: `random.randint(a,b)`: choisir par hasard des nombre entre a et b
`for in range(c)`

* les fonction python utilisée dans list, tuple, dict, set

append: ajoute un element à la liste

insert: „ „ selon l'ordre

extend: includes a list in another one

clear(): delete all elements

Vue d'ensemble du pipeline

Collection
des données

→ Preprocessing

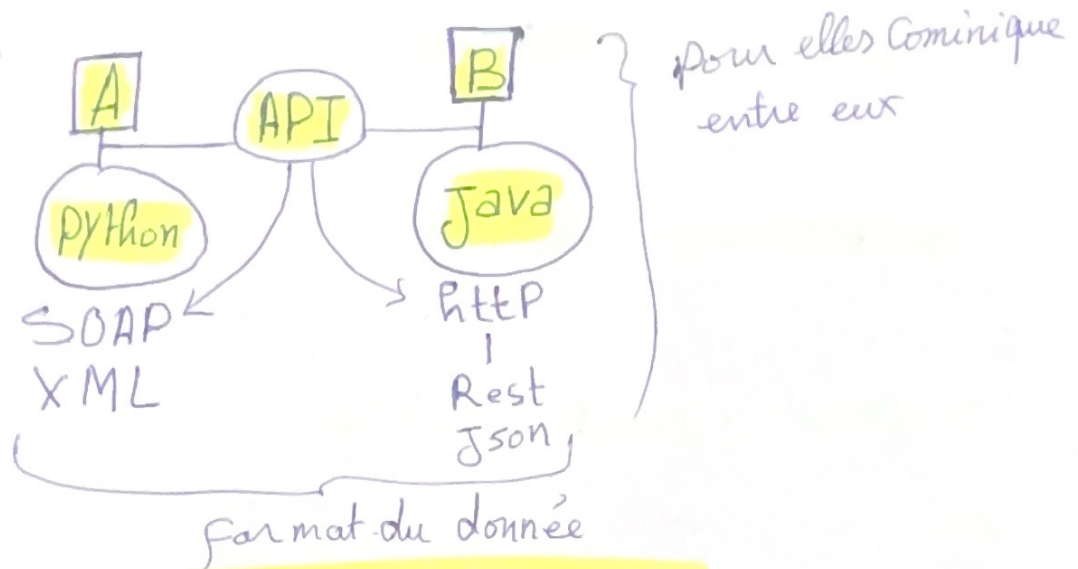
→ transformation

→ Datamining

→ exhaaction

NB: deux App l'un crée par python et l'autre crée par Java, elles peuvent pas communiquer entre eux

exemple



Regles d'association

Son objectif de règles une association est de trouver la relation cachée qui existe entre les éléments

les mesures de règles une association:

Support
Confiance
Lift

$$S = \frac{\text{nombre de transactions}}{T}$$

$$\text{Confiance} \Rightarrow C(x \rightarrow y)$$

$$C(x \rightarrow y) = \frac{S(xy)}{S(x)}$$

Element	T	Support	Frequent
x	1, 2, 3, 4	$\frac{4}{5}$	80%
y	1, 2	$\frac{2}{5}$	40%
z	2	$\frac{1}{5}$	20%
p	3	$\frac{1}{5}$	20%

* Pourquoi utiliser différents format de fichier ?

□ CSV : format simple et largement supporté.

→ (Pandas.read_csv())

□ Txt : format brut pour text simple

→ (Pandas.read_text())

□ XLSX : données complex avec feuille multiples

→ (Pandas.read_excel())

□ JSON : données structure facile à lire

→ (Pandas.read_json())

□ XML : données hiérarchique

→ (Pandas.read_xml())

□ HTML : Extraction de tableaux web

(Pandas.read_html())

Pandas : est utilisé pour la manipulation et l'analyse de données