Traitement des données en tables

Lycée Philippe de Girard Avignon – 1ère NSI

1 Introduction

Comme expliqué dans l'activité introductive, la majorité des données dans le monde sont stockées dans des bases de données. L'introduction de l'étude des bases de données est au programme de terminale NSI, notamment via un langage spécifique aux bases de données : le langage SQL (`Structured Query Language`). En première, il s'agit de manipuler des données sans passer par des bases de données, mais en effectuant certaines opérations sur ces données qui seront revues en terminale avec SQL.

Ces manipulations seront faites à l'aide de tables et en Python, qui fournit un ensemble d'outils facilitant la manipulation de données en tables. Un exemple de table se trouve ci-contre.

Nom	Prénom	Date de naissance
Durand	Jean-Pierre	23/05/1985
Dupont	Christophe	15/12/1967
Terta	Henry	12/06/1978

Utilité des données en tables et limites des tableurs De nombreuses données peuvent être représentées grâce à des tables : bulletin d'élève, feuille d'appel, mais aussi score de tennis, relevé de compte bancaire, etc.

Jusqu'à présent, si vous avez déjà manipulé des données, vous avez certainement utilisé des tableurs (`Excel` ou `Calc`) pour effectuer ces manipulations. En effet, ces logiciels sont faits pour manipuler des données temporairement sans nécessairement avoir besoin de robustesse et de sécurité sur celles-ci.

Manipuler des données sous forme de tables permet de combler certaines limites inhérentes aux tableurs :

— Les tableurs sont des fichiers souvent difficilement protégeables via des accès authentifiés.

— Les tableurs n'ont que très peu d'outils permettant d'assurer la cohérence des données (types de données représentatifs des données représentées et identiques pour une même colonne, présence de descripteurs, suppression de doublons).

2 Vocabulaire

P-uplet Un p-uplet (ou n-uplet, ou tuple) est une ligne d'une table. Cela correspond donc à un ensemble de valeurs décrivant un même objet.

Descripteur Un descripteur est l'étiquette d'une colonne d'une table. Il s'agit de ce que chaque donnée du p-uplet représente pour un objet donné. On utilise aussi le terme **Attribut**.

Domaine de valeurs Un domaine de valeurs est l'ensemble des valeurs que peut prendre une donnée. On attribue généralement un domaine de valeurs à une colonne, et il s'agit souvent d'un ensemble de valeurs issu d'un type de base (ex : l'âge a comme domaine de valeurs les entiers naturels, sous-ensemble des entiers).

Indexation Indexer une table consiste à attribuer une valeur unique (généralement entier naturel) à chaque p-uplet de cette table, c'est-à-dire à attribuer des indices à chaque p-uplet (comme pour un tableau en Python).

Doublon Un doublon est un p-uplet dont les valeurs sont toutes strictement identiques à un autre p-uplet.

Cohérence La cohérence d'une table est le fait que cette table représente convenablement les données qu'elle contient. Les doublons ou des domaines de valeurs différents sur une colonne rendent une table non cohérente.

Tri d'une table Trier une table consiste à ordonner les p-uplets selon une ou plusieurs colonnes. Si plusieurs colonnes sont utilisées, le tri est ordonné (d'abord selon une colonne, puis selon une autre).

Sélection (ou recherche) dans une table La sélection consiste à créer la sous-table d'une table contenant les p-uplets qui respectent un critère exprimé sous forme d'expression boolénne (voir l'activité introductive).

Fusion (ou jointure) de deux tables La fusion consiste à créer une table à partir de deux tables ayant un descripteur commun. La table possède alors les colonnes des deux tables fusionnées. (voir l'activité introductive).

3 Format CSV

Il existe un format de données (donc de fichier) permettant de gérer le traitement des données en tables. Il s'agit du format CSV (`Comma-Separated Values`, en français "Valeurs Séparées par des Virgules"). Les fichiers CSV sont des fichiers textes contenant les données déjà tabularisées (qui ont déjà la forme d'un tableau) et respectant certaines règles :

— Chaque ligne du fichier correspond à un p-uplet.

— Sur une ligne, le caractère ',' (virgule) permet de séparer les colonnes. Il arrive parfois que le séparateur soit différent : point-virgule, deux-points, tabulation, espace, etc. En effet, notamment en français, la virgule est déjà le séparateur des nombres décimaux, ce qui n'est pas le cas en anglais.

Exemple de l'introduction en CSV :

```
nom,prenom,date_naissance
Durand,Jean-Pierre,23/05/1985
Dupont,Christophe,15/12/1967
Terta,Henry,12/06/1978
```

Il est possible depuis tout tableur d'exporter les données en CSV, ou au contraire d'importer un fichier CSV.