TP1 : CSV, XML, JSON[1]

1. Les données dans notre société

Les données sont devenues un enjeu pour notre société. Elles touchent tous les domaines : la santé, l'éducation, l'industrie, la sécurité, le commerce,... De nouveaux termes sont apparus : ***Big Data***, ***Open Data***, ... de nouveaux métiers sont créés : *Architecte Big Data*, *Data Scientist*, ... de nouvelles disciplines sont enseignées : *global data analytics*, ... de nouvelles technologies sont développées : le ***Cloud Computing***, les bases de données ***NoSQL***, ... et de nouveaux algorithmes sont appliqués : ***MapReduce***, ***Spark***, ...

L'utilisation et la maîtrise du big data suscite beaucoup d'enthousiasme, mais également des inquiétudes, en particulier sur la protection des données à caractère personnel.

Travail n°1 :

A partir des liens ci-dessous, répondre aux questions suivantes :

https://www.lebigdata.fr/definition-big-data
https://www.lebigdata.fr/top-metiers-du-big-data-cloud
https://fr.blog.businessdecision.com/bigdata/2016/05/blockchain-big-data-enjeux-strategiques/
https://blockchainfrance.net/decouvrir-la-blockchain/c-est-quoi-la-blockchain/
https://www.lebigdata.fr/open-data-definition

- Etablir une définition du big data.
- Définir la règle des 3v qui définit les caractéristiques des outils big data.
- Expliquer le métier d'ingénieur big data.
- Indiquer en quoi la solution blockchain permet de protéger les données.
- Lister les 3 critères fondateurs de l'opendata.

[1] d'après les activités proposées par CANTALOUBE J. sur eduscol.education.fr

2. Mise en forme des données

Les données sont principalement représentées sous la forme de tableaux. On parle de données tabulaires.

Exemple :

Nom	Prenom	Age
Perrin	Léo	18
Petit	Loïc	32
Leroux	Pierre	27

Il existe trois formats pour représenter un tableau de données : les formats CSV, XML et JSON.
Ces trois formats sont des fichiers composés d'une suite de caractères où l'on distingue deux types d'information :

- Les données.
- Les caractères permettant de structurer ces données.

a) Le format CSV

Le format *Comma Separated Values (CSV)* structure les données sous la forme de valeurs séparées par des virgules. Ce format est très facile à générer et à manipuler.
Chaque ligne du fichier CSV correspond à une ligne du tableau et chaque valeur séparée par une virgule correspond à une colonne du tableau.

Exemple du tableau précédent au format CSV :

```
Nom,Prenom,Age
Perrin,Léo,18
Petit,Loic,32
Leroux,Pierre,27
```

La première ligne du fichier contient l'entête de la table, à savoir le nom de chacune des colonnes. Les lignes suivantes contiennent les données du tableau, en respectant l'ordre des colonnes. Le séparateur n'est pas forcément une virgule, on peut par exemple utiliser le point-virgule.

b) Le format XML

Le format *eXtensible Markup Language* (XML) est un format basé sur l'utilisation de balises pour structurer les données. Les balises sont utilisées pour encadrer un contenu : il y a une balise ouvrante et une balise fermante.

Exemple du tableau au format XML :

```
<?xml version="1.0" encoding="UTF-8"?>
<liste>
    <personne>
        <NOM>Perrin</NOM>
        <PRENOM>Léo</PRENOM>
        <AGE>18</AGE>
    </personne>
    <personne>
        <NOM>Petit</NOM>
        <PRENOM>Loic</PRENOM>
        <AGE>32</AGE>
    </personne>
    <personne>
        <NOM>Leroux</NOM>
        <PRENOM>Pierre</PRENOM>
        <AGE>27</AGE>
    </personne>
</liste>
```

c) Le format JSON

Le format *JavaScript Object Notation* (JSON) est un format plus récent utilisé pour représenter des objets qui dérive de la notation des objets du langage JavaScript. Un document JSON est essentiellement un ensemble de *paires* constituées d'une étiquette et d'une valeur ou d'une liste de valeurs. Les paires sont placées entres accolades et séparées par des virgules. Les valeurs des listes sont placées entre crochets et séparées par des virgules.

Exemple du tableau au format JSON :

```
{
  "liste": {
    "personne": [
      {
        "NOM": "Perrin",
        "PRENOM": "Léo",
        "AGE": 18
      },
      {
        "NOM": "Petit",
        "PRENOM": "Loic",
        "AGE": 32
      },
      {
        "NOM": "Leroux",
        "PRENOM": "Pierre",
        "AGE": 27
      }
    ]
  }
}
```

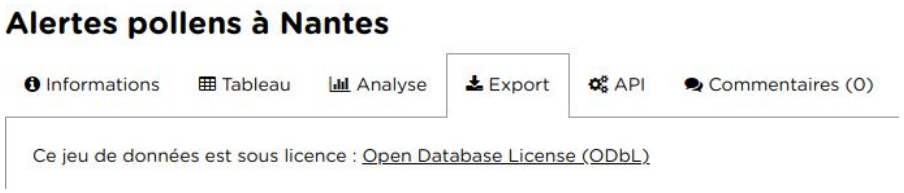
Le premier ensemble possède une paire dont la clé est ***liste*** et dont la valeur est une liste de trois éléments. Chacun de ces trois éléments est un ensemble avec trois paires représentant respectivement le nom, le prénom et l'âge de chaque personne. Remarquez que les nombres ne sont pas placés entre guillemets, contrairement aux mots et aux clés qui doivent être entre guillemets.

Travail n°2 :
A partir du tableau suivant créer trois fichiers au format CSV, XML et JSON.

Nom	Prix	code
Banane	5,99	77
Pomme	2,99	99
Poire	7,99	170

3. Traitement des données avec Python

Python possède de nombreuses bibliothèques dédiées au traitement des données. L'objectif de l'activité sera de réaliser un programme pour chaque format de données : CSV, XML et JSON.
Les fichiers de données, que nous allons utiliser, sont accessibles sur l'open data de Nantes à l'adresse suivante : https://data.nantes.fr/donnees/detail/alertes-pollens-a-nantes/



Ces données représentent le niveau de pollens à Nantes en fonction des essences d'arbres ou de plantes. Ces données sont actualisées tous les jours.

Table : Alertes_pollens_nantes

Nom de l'attribut	Code de l'attribut	Type	Description	Valeurs possibles
Date	Date	DATE	Date	2018-02-14T00:00:00.000Z
Nom	Nom	STRING	Nom de l'espèce allergisante	Flouve, Houlque, Noisetier, Ray grass, Graminée, Vulpin, Chêne, Frêne, Bouleau, Dactyle, Armoise, Plantain, Fromental, Fléole, Saule, Aulne
Type	Type	STRING	Type de l'espèce	Herbacée, Arbre
Sous-type	Sous-type	STRING	Sous-type de l'espèce	Graminée, null
Etat	Etat	INTEGER	Etat de l'alerte (1 pour "pas d'émission", 2 pour 'émission en cours", 3 pour "plus d'émission" et 4 pour "non observable")	1, 2, 4

Travail n°3 :
Télécharger puis renommer les fichiers CSV, JSON sous la forme Alertes_pollens_nantes.csv

d) Traitement d'un fichier CSV

La bibliothèque CSV permet la manipulation des fichiers CSV : voir https://docs.python.org/fr/3/library/csv.html#module-csv

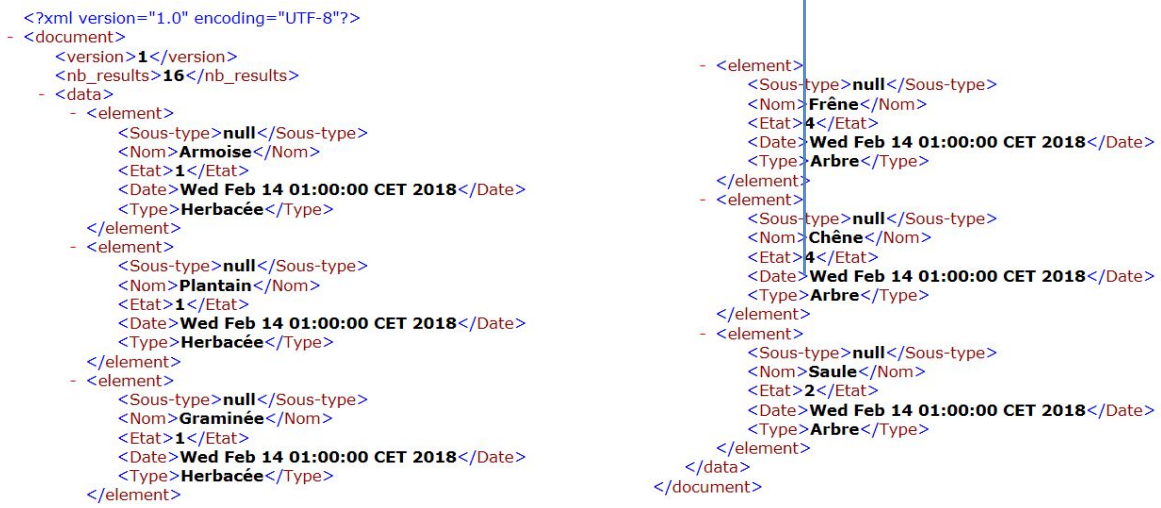
Exemple :

```
import csv
with open('Alertes_pollens_Nantes.csv', newline='') as csvfile:
   spamreader = csv.reader(csvfile, delimiter=' ')
   for row in spamreader:
       print(', '.join(row))
```

Travail n°4 :
Réalisez un programme permettant d'afficher le nom de toutes les espèces allergisantes.
Vous utiliserez obligatoirement un algorithme avec une structure itérative (boucle for).

e) Traitement d'un fichier XML

Le fichier XML que nous souhaitons analyser est de la forme :



Pour traiter des données XML, Python dispose de bibliothèques comme par exemple ***LXML*** : http://apprendre-python.com/page-xml-python-xpath

Le XML est un format standard du net, notamment pour les flux RSS. De nombreux fils d'actualité (météo, bourse, informations,...) sont disponibles en ligne au format RSS. Ce format est de plus en plus supplanté par le format JSON.

f) Traitement d'un fichier JSON

Le format JSON est aussi un standard du net et son utilisation est devenue prépondérante par rapport au format XML.
Le fichier JSON que nous souhaitons analyser est de la forme :

Structure de base du JSON est une paire clef-valeur (key-value) : `"Nom": "Armoise"`
On distingue les valeurs atomiques et les valeurs complexes (construites)

- Valeurs atomiques :
  - chaînes de caractères (entourées par les classiques guillemets) `"version": "1"`
  - nombres (entiers, flottants) `"nb_results": 16`
  - valeurs booléennes (true ou false)
- Valeur complexes :
  - Tableau `"data" : [{…},{…},{…},]`
  - Objet `"date" : {"$date":"2018-02-14T00 :00 :00.000Z"}`

Pour travailler avec le format JSON, Python va utiliser la bibliothèque json : https://www.w3schools.com/python/python_json.asp

Exemple :

```
import json
with open('Alertes_pollens_Nantes.json', 'r') as f:
    datas = json.load(f)
print(datas)
```

Travail n°6 :
Recopiez le code ci-dessus. Testez et analysez le code à l'aide de la documentation Python.
Réalisez ensuite un programme permettant d'afficher le nom de toutes les espèces allergisantes ainsi que leur type (Herbacée ou Arbre).
Vous utiliserez obligatoirement un algorithme avec une structure itérative (boucle for).