Les Pandas, les Poneys et la Persistance des données

Ici nous apprenons à utiliser plusieurs librairies de manipulation et de mise en forme des données.

Liens utiles :

- Notebook à partir de PyCharm :
  - https://www.jetbrains.com/help/pycharm/using-ipython-notebook-with-product.html
- Pandas :
  - http://www.python-simple.com/python-pandas/panda-intro.php
- Pony :
  - https://docs.ponyorm.com/firststeps.html

Pour installer les librairies pandas et pony :

```
$ pip3 install pandas
```

```
$ pip3 install pony
```

Les notebooks Jupyter

Ce travail sera réalisé à l'aide de "notebooks" fonctionnant sur l'interpréteur "jupyter". Les notebooks permettent d'écrire et d'exécuter des scripts python à l'aide d'un simple navigateur web. Les résultats d'exécution sont conservés et peuvent être retrouvés d'une session à l'autre.

- Si vous êtes sous Windows ou Mac, utilisez l'environnement des notebooks fourni par Anaconda
- Sur un environnement Unix, Ouvrez un terminal dans votre dossier de travail et tapez :

```
$ jupyter-notebook
```

Ceci ouvre un onglet de l'interpréteur jupyter dans votre navigateur.

- Créez un notebook vierge via le menu new –> `python 3`
- Ou bien cliquez sur le notebook sur lequel vous souhaitez travailler.

Pour utiliser un notebook, voir :

- 1. What is the Jupyter notebook?
- 2. Notebook basics
- 3. Running code
- 4. Working with Markdown cells

- Une vidéo en anglais

Analyser les données : Pandas

L'utilisation de données structurées dans un programme Python nécessite de faire appel à des librairies spécialisées. Nous utiliserons ici la librairie `pandas` qui sert à la mise en forme et à l'analyse des données.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas
```

On considère une série d'enregistrements concernant des ventes réalisées par un exportateur de véhicules miniatures. Pour chaque vente, il entre dans son registre de nombreuses informations :

- nom de la société cliente
- nom et prénom du contact, adresse, téléphone
- nombre d'unités vendues
- prix de vente
- etc...

Ces informations sont stockées dans un fichier au format 'csv' (comma separated values) : `ventes_new.csv`. Téléchargez ce fichier et copiez-le dans votre répertoire de travail.

Dans un premier temps, regardez son contenu avec un editeur de texte (**geany**, **gedit** ou autre...). La première ligne contient les noms des attributs (`NUM_COMMANDE`, `QUANTITE`,...). Les ligne suivantes contiennent les valeurs d'attributs correspondant à une vente donnée. En tout plus de 2000 ventes sont répertoriées dans ce fichier.

Ouvrez-le maintenant à l'aide d'un tableur (par exemple **`localc`**). Les données sont maintenant "rangées" en lignes et colonnes pour faciliter la lecture.

Déplacez le fichier `ventes_new.csv` dans votre répertoire de travail.

Lecture des données

Les données sont au format `csv`, on utilise:

- `pandas.read_csv`. Voir dataframes pandas. Pandas permet également de lire les données au format `xls` et `xlsx` (Excel).

```
with open('ventes_new.csv') as f:
    data = pandas.read_csv(f)
print(data)
```

avec `data` une structure de données de type `DataFrame`

Testez les commandes suivantes :

```
print(len(data))
```

```
print(data.columns)
```

Syntaxe de type dictionnaire :

```
print(data["VILLE"])
```

```
print(data[["VILLE", "PAYS"]])
```

Autre syntaxe :

```
print(data.VILLE)
```

```
print(data.VILLE.head(10))
```

PS : Ça marche aussi avec la syntaxe "dictionnaire":

```
print(data["VILLE"].head(10))
```

pour afficher les lignes

Tout tableau de données possède un index:

```
print(data.index)
```

(il s'agit ici d'une indexation automatique par les entiers)

Les données peuvent être accédées par leur index:

```
print(data.loc[0])
```

Modifier les données

Les prix augmentent de 1 euro :

```
data.PRIX_UNITAIRE += 1
data.MONTANT = data.PRIX_UNITAIRE
data.MONTANT *= data.QUANTITE
print(data.MONTANT)
```

Sélectionner les données

```
selection = data[data.MONTANT > 6000]
```

l'objet `selection` se comporte comme un nouveau dataframe ne contenant que les entrées respectant le critère de sélection.

Pour faciliter l'interprétation du résultat, on n'affiche le résulat que sur un sous-ensemble d'attributs:

```
print(selection[["MONTANT","DATE_COMMANDE","VILLE","PAYS","NOM_CONTACT","PRE
NOM_CONTACT"]])
```

Sélection multi-critères :

```
selection = data[(data.MONTANT > 6000) & (data.PAYS == 'France')]
print(selection[["MONTANT","DATE_COMMANDE","VILLE","PAYS","NOM_CONTACT","PRE
NOM_CONTACT"]])
```

<!-- === Opérateurs d'agrégation == * usage : statistique sur les données * principe : * opérateur d’aggrégation : * tout type de données : comptage (attention aux doublons) <code python> print(data["VILLE"].count()) print(data["VILLE"].drop_duplicates().count()) </code> * données quantitatives (et non qualitatives) : somme, moyenne, ecart-type (count, sum, mean, std, min, max, …) <code python> print(data["MONTANT"].mean()) print(data["MONTANT"].std()) </code> === Affichage et figures === Un histogramme simple <code python> data["MONTANT"].hist(bins=25) plt.show() </code> <code python> data.QUANTITE.hist(by = data.TYPE_PRODUIT, bins=25, figsize = (15,8)) </code> === Calcul par groupes === Pandas offre la possibilité d'organiser et analyser les données par //groupe//. Le découpage en groupe repose sur des valeurs d'attributs (il y a autant de groupes qu'il y a de valeurs différentes pour l'attribut considéré) Par exemple si on prend le type de produit: <code python> groupes_selon_produit = data.groupby('TYPE_PRODUIT') </code> ici l'objet ''groupes_selon_produit'' définit les groupes sur le tableau de données selon la valeur de ''TYPE_PRODUIT''. Pour visualiser les groupes: <code python> print(groupes_selon_produit.groups) </code> On peut ensuite effectuer des mesures et calculs par groupes. Par exemple : <code python> nb_ventes_par_produit = groupes_selon_produit.size() </code> l'objet ''nb_ventes_par_produit'' est une liste indexée par les valeurs d'attributs (ici 'Bateaux', 'Avions' etc...) <code python> print(nb_ventes_par_produit.index) </code> On peut bien sûr l'afficher : <code python> print(nb_ventes_par_produit) </code> Les fonctions sum(), mean(), max(), min() etc... s'appliquent sur des valeurs quantitatives, ici ''MONTANT'' ou ''QUANTITE''. Exemple : le chiffre d'affaires par produit (somme des montants) : <code python> CA_par_produit = groupes_selon_produit.MONTANT.sum() </code> Enfin on peut également effectuer une sélection sur les valeurs calculées (l'équivalent du ''HAVING'' en SQL). Exemples: * les produits générant un chiffre d'affaires > 1000000: <code python> print(CA_par_produit[CA_par_produit > 1000000]) </code> * le produit générant le plus haut chiffre d'affaires: <code python> print(CA_par_produit[CA_par_produit == max(CA_par_produit)]) </code> Les groupes peuvent être définis sur des critères multiples : <code python> groupes_pays_ville = data.groupby(['PAYS', 'VILLE']) </code> === Affichage et figures === <code python> grouped = data.groupby(data.TYPE_PRODUIT) print(grouped.NUM_COMMANDE.count()) plt.figure() grouped.NUM_COMMANDE.count().plot(kind = "bar", figsize = (5,3)) plt.figure() grouped.NUM_COMMANDE.count().plot(kind = "pie", figsize = (5,3)) </code> Pour aller plus loin : *

{{http://www.xavierdupre.fr/app/ensae_teaching_cs/helpsphinx/notebooks/td2a_cenonce_session_1.html|Une introduction très détaillée aux DataFrames (en Français)}} * {{http://synesthesiam.com/posts/an-introduction-to-pandas.html#getting-data-out|Introduction to Pandas (en anglais)}} <note tip> ** A faire ** * Trouvez le nombre de ventes, le nombre de clients référencés (sans doublons), et le nombre de références produits (sans doublons). * Afficher le nombre de client et le chiffre d'affaires (somme des montants) * par pays * par pays puis par état * par pays puis par état puis par ville * Donnez le nombre de ventes en fonction du mois pour l'année 2004 * Donnez le chiffre d'affaires par année et trimestre pour les ventes réalisées aux états unis * Quelle est la catégorie de véhicules la plus vendue? </note> === Tables Pivot === Agrégation des données selon différents attributs/dimensions exemple : on représente les ventes selon (1) la dimension géographique et (2) la dimension temporelle <code python> T = pandas.pivot_table(data, values = 'MONTANT', index = ['PAYS'], columns = ['ANNEE'], aggfunc=np.sum) print(T) </code> <code python> T.plot(kind='bar', subplots = 'True') plt.show() </code> Evolution des ventes au cours de l'année pour la France seulement: <code python> selection = data[data.PAYS == "France"] T2 = pandas.pivot_table(selection, values = 'MONTANT', index = ['ANNEE'], columns = ['VILLE'], aggfunc=np.sum) print(T2) T2.plot(kind='bar', subplots = 'True') plt.show() </code> <note tip> ** A faire ** * Donnez le nombre de ventes (''aggfunc = np.size'') par catégorie pour chaque année et trimestre. Choisissez le graphique le plus adapté pour représenter les données. * Donnez le chiffre d'affaires par pays, pour chaque catégorie de produits. Choisissez le graphique le plus adapté pour représenter les données. </note> -->

Organiser et transformer les données : Pony ORM

La librairie Pony ORM est un gestionnaire de persistance qui permet la mise en correspondance entre les objets d'un programme et les valeurs d'une base de données, pour assurer leur conservation d'une session à l'autre.

Pony effectue toutes les opérations de sauvegarde de manière transparente. La création et la mise à jour des objets persistants s'accompagne automatiquement d'opérations de lecture/écriture vers la base de donnée. Les données sont donc conservées sans appel explicite à des requêtes SQL.

Initialisation

```
from pony import orm
```

```
db = orm.Database()
```

Création du schéma de données

Nous définissons ici trois schémas de classes correspondant aux ensembles d'entités Client, Commande et Produit.

- **Client**(id_client, téléphone, ville, pays)
- **Commande**(num_commande, quantité, montant, mois, année, id_client, code_produit)
- **Produit**(code_produit, type_produit, prix_unitaire)

Les clés étrangères de la table des commande définissent deux relations de un à plusieurs :

- une relation de un à plusieurs entre *un* produit et *des* commandes,
- et une relation de un à plusieurs entre *un* client et *des* commandes.

Dans un modèle ORM, les relations de un à plusieurs se traduisent par des attributs de type liste ou ensemble :

- A un client correspond un ensemble de commandes
- A un produit correspond un ensemble de commandes
- A une commande correspond *un* client et *un* produit

Classe Client

Les classes sont définies ici comme des schémas de données.

La classe Client hérite de la classe générique *Entity*. Les attributs des objets obéissent à une définition parmi quatre définitions possibles :

- attribut clé primaire : `PrimaryKey`
- attribut requis (la valeur doit être renseignée) : `Required`
- attribut facultatif: `Optional`
- relation de un à plusieurs : `Set`

```
class Client(db.Entity):
    id_client = orm.PrimaryKey(str)
    telephone = orm.Required(str)
    ville = orm.Required(str)
    pays = orm.Required(str)
    achats = orm.Set('Commande')
```

Classe Produit

```
class Produit(db.Entity):
    code_produit = orm.PrimaryKey(str)
    type_produit = orm.Required(str)
    prix_unitaire = orm.Required(float)
    ventes = orm.Set('Commande')
```

Classe Commande

Dans la classe Commande, il n'y a pas de clé étrangère (comme dans le modèle relationnel) mais :

- un attribut de type `Client` qui lie la commande au client qui a effectué la commande

- un attribut de type `Produit` qui lie la commande au produit commandé

```
class Commande(db.Entity):
    num_commande = orm.PrimaryKey(int)
    quantité = orm.Required(int)
    montant = orm.Required(float)
    mois = orm.Required(int)
    année = orm.Required(int)
    client = orm.Required(Client)
    produit = orm.Required(Produit)
```

Pour afficher

La commande `show` est une commande d'affichage à tout faire. Elle permet ici de vérifier le schéma de la classe.

```
orm.show(Client)
```

Association à un gestionnaire de BD

Les schémas de données définis dans les classes peuvent être implémentés dans différents gestionnaires de bases de données.

Nous choisissons ici le gestionnaire sqlite, ce qui évite de définir une connexion un serveur distant. La base de données est ici émulée en mémoire centrale (pour les besoins de l'exercice, les données n'ont pas besoin d'être conservées)

```
db.bind(provider='sqlite', filename=':memory:')
```

Mode debug

Le mode debug permet de voir les échanges avec la base de données.

```
orm.set_sql_debug(True)
```

La commande `generate_mapping` définit l"appariement entre les objets et la base de données. Cela correspond ici à la création de trois tables.

```
db.generate_mapping(create_tables=True)
```

Transfert des données Client

Les données sont lues dans le dataFrame `data` sur les quatre attributs définis et insérées dans la base à l'aide du constructeur de la classe Client.

On vérifie avant l'insertion que le client n'est pas déjà présent dans la base à l'aide du test:

```
    if Client.get(id_client = c.CLIENT) is None:
        ...
```

```
clients = data[["CLIENT", "TELEPHONE", "VILLE", "PAYS"]].drop_duplicates()
for i in range(len(clients)):
    c = clients.iloc[i]
    if Client.get(id_client = c.CLIENT) is None:
        Client(id_client = c.CLIENT, telephone = c.TELEPHONE, ville =
c.VILLE, pays = c.PAYS)
        orm.commit()
```

On remarque que l'initialisation des clients ne porte que sur les attributs élémentaires (la liste des achats n'est pas initialisée explicitement).

Affichage

Pour afficher la liste de tous les clients (et non le schéma de la classe Client), il faut faire appel à la méthode `select()` qui effectue une lecture dans la base avant l'affichage.

```
Client.select().show()
```

On peut également afficher les clients un par un à l'aide leur index (ici le nom du magasin)

```
print(Client["Land of Toys Inc."])
print(Client["Land of Toys Inc."].id_client)
print(Client["Land of Toys Inc."].ville)
print(Client["Land of Toys Inc."].pays)
print(Client["Land of Toys Inc."].achats)
```

On notera que la liste des achats est vide (les commandes n'ont pas encore été saisies)

L'appel à la méthode `select()` permet de sélectionner les clients selon la valeur d'un ou plusieurs attributs. Cette sélection passe par une fonction anonyme `lambda`:

```
requête = Client.select(lambda c : c.pays == "France")
```

Une requête se comporte comme un itérateur sur les objets:

```
for c in requête:
    print(c.id_client, c.ville, c.pays)
```

Transfert des données produits

Les produits sont insérés de la même façon que les clients:

```
produits = data[["CODE_PRODUIT", "TYPE_PRODUIT",
"PRIX_UNITAIRE"]].drop_duplicates()
for i in range(len(produits)):
    p = produits.iloc[i]
    if Produit.get(code_produit = p.CODE_PRODUIT) is None:
        Produit(code_produit = p.CODE_PRODUIT, type_produit =
p.TYPE_PRODUIT, prix_unitaire = p.PRIX_UNITAIRE)
        orm.commit()
```

Affichage du contenu de la classe

```
Produit.select().show()
```

Affichage d'un produit particulier

```
print (Produit['S10_1678'])
print (Produit['S10_1678'].type_produit)
print (Produit['S10_1678'].prix_unitaire)
print (Produit['S10_1678'].ventes)
```

Transfert des données ventes

Pour créer les commandes, il faut ici définir deux références :

- une référence au client qui a effectué la commande
- une référence au produit commandé

qui sont des objets définis précédemment lors de l'insertion des données client et des donnés produit. Ils correspondent donc à des entrées de leurs classes respectives, indexes par leur identifiant (`id_client` et `code_produit`).

```
ventes = data[["NUM_COMMANDE", "QUANTITE", "MONTANT", "MOIS", "ANNEE",
"CLIENT", "CODE_PRODUIT"]].drop_duplicates()
for i in range(len(ventes)):
    v = ventes.iloc[i]
    if Commande.get(num_commande = int(v.NUM_COMMANDE)) is None:
        client = Client[v.CLIENT]
        produit = Produit[v.CODE_PRODUIT]
        Commande(num_commande = int(v.NUM_COMMANDE),
                    quantité = int(v.QUANTITE),
                    montant = float(v.MONTANT),
                    mois = int(v.MOIS),
```

```
                        année = int(v.ANNEE),
                        client = client,
                        produit = produit)
        orm.commit()
```

Affichage

```
Commande.select().show()
```

```
print(Commande[10118])
print('Montant :', Commande[10118].montant)
print('Quantité :', Commande[10118].quantité)
print('Année :', Commande[10118].année)
print('Mois :', Commande[10118].mois)
print('Client :', Commande[10118].client)
print('Produit :', Commande[10118].produit)
```

Exemples de requête

```
requête = Commande.select(lambda c : c.montant > 10000)
for r in requête:
    print(r.num_commande, r.quantité, r.mois, r.année, r.client, r.produit)
```

Ou plus simplement :

```
requête.show()
```

Autre écriture

```
requête = orm.select(c for c in Commande if c.montant > 10000)
```

Mise à jour automatique des contenus

Maintenant que les commandes on été entrées dans la base, la liste des achats est à présent renseignée pour chaque client de la classe Client:

```
print(Client["Land of Toys Inc."])
print(Client["Land of Toys Inc."].id_client)
print(Client["Land of Toys Inc."].ville)
print(Client["Land of Toys Inc."].pays)
print(Client["Land of Toys Inc."].achats)
```

Et la liste des ventes est de même renseignée pour chaque produit de la classe Produit:

```
print (Produit['S10_1678'])
```

```
print (Produit['S10_1678'].type_produit)
print (Produit['S10_1678'].prix_unitaire)
print (Produit['S10_1678'].ventes)
```

Modifier les valeurs

```
Produit['S12_1108'].prix_unitaire = 100
orm.commit()
```

Supprimer un objet

```
Produit['S12_1108'].delete()
orm.commit()
```

A faire

- Pour chaque client, calculer le montant total des achats
- Pour chaque produit, calculer le montant total des ventes
- Corriger le champ `pays` pour les clients nord-américains : si le pays vaut `"United States"`, le remplacer par `"USA"`
- En profiter pour supprimer les doublons de la classe Client
- Créez un nouveau client
- Faites-lui commander plusieurs produits (n'oubliez pas de définir le numéro de commande!!)
- Vérifiez que les nouvelles commandes apparaissent bien dans la liste des ventes de la table `Produits` . Magique, non?

Si vous avez le temps

- Définissez un modèle ORM pour le schéma de données du TD1.
- Remplissez la base à l'aide des données contenues dans animal.json et équipement.json
- Effectuez quelques requêtes pour vérifier que tout marche bien