Analyse reproductible en Python
exercices séance 2

Sylvain Tenier

mardi 3 janvier 2017

Pour cette séance, réutilisez le dépôt Git créé pour la séance 1. Le travail réalisé doit y être déposé.

Exercice 1.

1. Lancez le serveur Jupyter Notebook.
2. Créez un nouveau notebook. Nommez le fichier "pandas.ipynb".
3. Accédez à l'adresse `http://pandas.pydata.org/pandas-docs/stable/10min.html`. Reproduisez le tutoriel en le traduisant (y compris au niveau du code) en Français jusqu'à la section *Reshaping* (inclus - Vous vous arrêterez avant la partie *Time Series*). Le texte doit être formaté dans des cells de type *markdown*. Le code doit être fonctionnel dans des cellules de type *code*.
4. Si la commande `jupyter` n'est pas reconnue, ajoutez le répertoire `bin` de la distribution Anaconda à la variable PATH.

Exercice 2.

1. Créez un nouveau notebook "ping.ipynb"
2. Créez une série dont, pour chaque ligne :

   l'index est un code pour chaque dominante de l'Esigelec

   la valeur est le nom complet de la dominante
3. Ajoutez une colonne avec le département auquel chaque dominante appartient
4. Créez une dataframe contenant le nom, prénom et code de dominante pour chaque membre de votre PING
5. Créez une dataframe contenant le pourcentage d'élèves par département pour votre PING
6. Enregistrez le résultat dans un fichier csv

Exercice 3.

1. Créez un notebook "routes.ipynb"
2. Récupérez les informations "Open Data" sur les travaux routiers en Seine-Maritime

   Les métadonnées http://opendata76.blob.core.windows.net/se-deplacer/CG76_DR_SGDR_DT_METAD.pdf

   Les données http://www.opendata-27-76.fr/jeux-de-donnees/cg76-routes-derniers-travaux/
3. Partie 1 : Chargement et nettoyage
   - Téléchargez les données au format CSV
   - Chargez les dans une dataframe
   - Vérifiez que `epaisseurdernierstravaux` est bien de type réel.
   - Identifiez et nettoyez les valeurs aberrantes. Justifiez votre stratégie de nettoyage.
4. Partie 2 : Analyse
   - Quels sont les 5 natures de travaux les plus fréquentes ?
   - Pour chaque année, quel est le nombre de travaux réalisés et l'épaisseur moyenne ? Triez le résultat par épaisseur descendante.

Exercice 4.

Parcourez la semaine 2 du MOOC *Introduction to Data Science with Python* sur Coursera.

Une évaluation sur vos acquis vous sera proposée lors de la prochaine séance.