

TP de Langages de script n° 7 : manipulation de fichiers textuels et expressions régulières

I) Avant de commencer

a) Votre source principale

Si vous ne connaissez pas encore ce site, naviguez vers <http://docs.python.org/py3k/> pour regarder toute la documentation officielle de PYTHON. En particulier, nous conseillons fortement de parcourir l'excellent *Tutorial* pour vos révisions et pour l'approfondissement. Et quand vous cherchez les renseignements sur un module et ses fonctions, rendez-vous dans *Library Reference*.

b) Quelques réglages

PYTHON gère bien le codage le plus moderne UTF-8, et il serait dommage de ne pas en profiter. Faites quelques réglages nécessaires de votre environnement de travail :

- Ajoutez `export LANG="fr_FR.UTF-8"` dans votre fichier `.bashrc`, redémarrer la session de travail.
- Réglez votre console sur le codage UTF-8. Vérifier que ça marche en affichant un fichier codé en UTF-8 (par exemple `vers-queneau.txt`) sur votre console.
- Vérifiez que votre éditeur de texte gère bien UTF-8 et sauvegarde les fichiers dans ce codage. Si nécessaire faites des réglages.

II) Les exercices

Exercice 1 : La disparition

Écrire un script qui, prenant en entrée un nom de fichier, en crée une nouvelle version dans laquelle toutes les occurrences du caractère `e` ont été supprimées.

Exercice 2 : Cent mille milliards de poèmes de Raymond Queneau

Le fichier `vers-queneau.txt` contient une séquence de 14 groupements de 10 vers séparés par des lignes vides. Le but de cet exercice est de générer un poème en prenant un vers de chaque groupe.

1. Écrire un script PYTHON chargeant le fichier `vers-queneau.txt` dans un dictionnaire associant aux entiers i de 0 à 13, la liste des dix vers du i -ième groupement.
2. Modifier le script pour générer un poème de 14 vers en prenant aléatoirement un parmi les dix vers dans chaque groupement. Utiliser pour cela le module `random` de la bibliothèque standard de PYTHON (<http://docs.python.org/py3k/library/random.html>). Ne pas oublier de signer Raymond Queneau à la fin du poème.

Exercice 3 : Expressions régulières : prise en main

Pour cet exercice, on va utiliser le module `re` (<http://docs.python.org/py3k/library/re.html>, voir aussi <http://docs.python.org/py3k/howto/regex.html>).

Vous pouvez tester vos expressions sur des petites chaînes d'abord, mais aussi sur le texte de Flaubert qui se trouve dans le fichier `educ.txt`.

1. Générez les listes suivantes : toutes les occurrences des mots qui contiennent la lettre `k` ou `w` ; toutes les années du 19-ième siècle ; toutes les formes du verbe `aimer`. Modifiez vos expressions pour ne chercher que les mots en fin de ligne.
2. Reprenez de l'exercice 1 du TP 3, et ré-écrivez en utilisant le module `re` une fonction `decoupe(phrase)` qui prend en argument la chaîne `phrase` et renvoie une liste des mots de cette chaîne (ne contenant pas le mot vide). Les mots peuvent être délimités par des espaces ou par tout signe de ponctuation (virgule, point, etc.).
3. Affichez les numéros de lignes qui contiennent les années de 19-ième siècle.
4. (optionnel) Ecrivez une fonction qui trouve tous les mots qui contiennent 2 fois les mêmes trois lettres consécutives (comme `barbarie` ou `entrent`). Ecrivez une fonction qui trouve les phrases qui contiennent 2 fois le même mot.
5. Affichez les numéros de lignes et les positions dans ces lignes de toutes les formes du verbe `aimer`.

Exercice 4 : Modifications de l'Éducation Sentimentale

On continue de travailler sur le fichier `educ.txt`.

1. Ce texte est formé de chapitres annoncés par une ligne contenant `Chapitre N` où `N` est le numéro du chapitre. Écrire un script qui charge le fichier en mémoire et crée un fichier pour chaque chapitre, intitulé `chapitre-i.txt` (où `i` est le numéro du chapitre).
2. Modifier le script pour qu'il remplace partout le prénom "Frédéric" par "Gustave".
3. Modifier le script pour qu'il remplace tous les mots commençant par "aim" suivi d'un suffixe par le mot "détest" suivi du même suffixe.
4. Modifier le script pour qu'il extraie les dialogues de chaque chapitre dans un fichier nommé `chapitre-dialogues-i.txt`.