Python Insia TP : les Misérables

Source du TP

Prendre un texte du Projet Gutenberg. Par exemple, les Misérables de Victor Hugo .

Exercices

Exercice 1

Explication : Ouverture de fichiers, `for line in file`, `.split`.

Faire un équivalent de `wc` en comptant les mots, les lettres et les lignes.

Exercice 2

Explication : `.lower`, `.upper`, `.islower`, `.isupper`, `.isalpha`.

Compter l'occurrence de chaque lettre (de `a` à `z`, en convertissant les majuscules en minuscules, et en ignorant les accents).

Exercice 3

Explication : le fichier words.py contient deux générateurs, que l'on peut utiliser en faisant `for word in words(file):`, un qui renvoie les mots de manière brute, et l'autre qui les renvoie en minuscules, nettoyés de leur ponctuation, et avec le mot magique `SEP` pour séparer deux phrases. Utilisation avec `import`.

Explication : `.items`, `.sort`, `cmp`

Compter l'occurrence de chaque mot (tels que renvoyés par `clean_words`), et afficher les 20 mots les plus fréquents, les moins fréquents et le nombre de mots différents.

Exercice 4

Compter tous les couples de mots successifs, avec leurs occurrences. Par exemple, dans la phrase "Le chien ? Il poursuit le chat.", on a les couples suivants : [("SEP", "le"), ("le", "chien") , ("chien", "SEP"), ("SEP", "il"), ("il", "poursuit"), ("poursuit", "le"), ("le", "chat"), ("chat", "SEP")]

Afficher le nombre de couples ainsi que les 10 couples les plus fréquents, et les 10 couples les moins fréquents.

Exercice 5

Parmi les couples précédent, chercher tous les couples dont le premier élément est `SEP`, avec leur occurrence. Par exemple, dans la phrase "Le chien ? Il poursuit le chat.", il y a deux couples qui commencent par `SEP` : ("SEP", "le") et ("SEP", "il"), avec une occurrence chacun. Afficher ces couples avec leur occurrence.

Exercice 6

Explication : module `random`

À partir des données des exercices précédant, créer un générateur de phrases.

Partir de `SEP`, et en utilisant `random`, choisir aléatoirement le premier mot de la phrase parmi les couples commençant par `SEP`, puis ainsi de suite, jusqu'à retomber sur `SEP`.

Par exemple, si dans la liste, on trouve 3 fois le couple ("le", "chat") et une fois le couple ("le", "chien"), il y aura 75% de chances que le mot "chat" soit choisi après le mot "le", et 25% de chances pour que ça soit le mot "chien".

Exercice 7

Modifier l'exercice 6 pour travailler avec des suites de 3 puis 4 mots au lieu de 2 mots. Prendre des extraits de texte plus petit si les performances ne suivent pas.

Suite la semaine prochaine: optimisation de l'exercice 7 avec des décorateurs de profiling.