جامعة هواري بومدين
للعلوم والتكنولوجيا
U S T H B

Programmation PYTHON

Cours 8

Nassim ZELLAL

2020/2021

Analyse statistique de données textuelles

- Qu'est-ce qu'une cooccurrence ?
- Les cooccurrences sont des unités textuelles pour le traitement statistique des textes.
- Co-présence/apparition simultanée et statistiquement significative de deux ou plusieurs unités linguistiques/éléments linguistiques/éléments dans une fenêtre textuelle précise, c'est-à-dire dans la même fenêtre contextuelle.
- Exemples : train rapide, train à vapeur, train luxueux, train complet, train de banlieue, rater le train, voiture verte, voiture électrique, voiture d'occasion, etc.

zip()

```
pays = ['Algérie', 'Mexique', 'Suisse']
indicatifs = [213, 52, 41, 33, 64]
#zip
res = zip(pays, indicatifs)
# mettre l'objet zip dans une liste
res_list = list(res)
print(res_list,"\n")
# unzip avec *
pays,indicatifs=zip(*res_list)
print(pays)
print(indicatifs)
```

Zip

```
[('Algérie', 213), ('Mexique', 52), ('Suisse', 41)]

('Algérie', 'Mexique', 'Suisse')
(213, 52, 41)
```

Calcul de cooccurrences - exercice

- Écrire un script permettant de calculer les cooccurrences du fichier "text-b.txt" encodé en UTF-8. Ce fichier est le premier argument passé à votre script.
- Le deuxième argument est la longueur de la cooccurrence, qui peut aller de 2 à *n* tokens.
- Le troisième argument est la fréquence de la cooccurrence, qui peut aller de 1 à *n*.
- Les deux derniers arguments sont la longueur du premier et du dernier token de la cooccurrence.
- Exemples de sortie :
  - **direction générale 4** (cette cooccurrence a été extraite via les arguments ➔ text-b.txt 2 4 9 8
  - **services sociaux des chemins de fer 3** (cette cooccurrence a été extraite via les arguments ➔ text-b.txt 6 3 8 3

Manipulation de répertoires - listdir(path)

- La méthode listdir(path), appartenant au module « os », retourne la liste des noms des entrées d'un répertoire. Autrement dit, cette méthode retourne la liste des fichiers et des répertoires situés dans le répertoire cible.
- La méthode listdir(path) retourne uniquement le premier niveau et prend comme argument un « path » (chemin).
- Exemple: tester la méthode listdir() sur le répertoire « **TEST** ».
- import os,sys
- print(**os.listdir(sys.argv[1])**)
- **> ['file1.txt', 'file2.txt', 'TEST2']**

Manipulation de répertoires - isdir()

- La méthode isdir(path), appartenant au module « path » du module « os », permet de vérifier si un chemin (path) est un répertoire existant.

- Exemple: vérifier si « **TEST** » est un répertoire.

- import os,sys
- print(os.**path.isdir(sys.argv[1])**)
- **> True**
- Remarque : avec la méthode **isfile(path)** du module « path », on vérifie si un chemin est un fichier.

Exercice 1

- Écrire un script permettant de lister tout le contenu du répertoire « **COURS8** ».
- Vous devez obtenir la sortie suivante :

- **COURS8/REP/REP1/file1.txt**
- **COURS8/REP/REP1/REP2/REP3/file2.txt**
- **COURS8/REP/REP1/REP2/REP3/REP4/file3.txt**
- **COURS8/REP/REP1/REP2/REP3/REP4**
- **COURS8/REP/REP1/REP2/REP3**
- **COURS8/REP/REP1/REP2**
- **COURS8/REP/REP1**
- **COURS8/REP**

Exercice 2

- Faire un script Python qui prend en entrée les fichiers "dico1.txt" et "dico2.txt".

- Le résultat généré par ce script est un fichier contenant :

  - les entrées qui sont dans dico1.txt et dico2.txt ;

  - les entrées qui sont dans dico1.txt et qui ne sont pas dans dico2.txt.

Exercice 3

- Reprendre votre solution du précédent exercice pour :
  - générer un compteur démarrant à 1 pour chaque catégorie dans le fichier de sortie;
  - prendre chaque entier maximal et calculer sa factorielle, en l'affichant sur l'invite de commandes.