

WordCount

January 1, 2017

1 Movielens 100k

Ecrivez le code python permettant de calculer la distribution des notes des films proposés sur movielens. Vous utiliserez le dataset “100k” disponible à l’adresse <http://grouplens.org/datasets/movielens/100k/>.

2 Programme

Importation de la librairie csv permettant de lire les données du fichier u.data, de la librairie time permettant de calculer le temps d’exécution

```
In [5]: import csv
import time

start_time = time.clock()
```

Création du dictionnaire “notes” permettant de stocker les résultats

```
In [6]: notes = {
    1: 0,
    2: 0,
    3: 0,
    4: 0,
    5: 0
}
```

Lecture du fichier contenant les notes et stockage des résultats dans le dictionnaire

```
In [7]: with open('ml-100k/u.data', 'r') as csvfile:

    spamreader = csv.reader(csvfile, delimiter='\t')

    for row in spamreader:
        notes[int(row[2])] += 1
```

Affichage des résultats

```
In [8]: print("--- Temps d'execution : %s secondes ---\n" % (time.clock() - start_t
        print("Nombre de films à 1 étoile : " + str(notes[1]))
        print("Nombre de films à 2 étoile : " + str(notes[2]))
        print("Nombre de films à 3 étoile : " + str(notes[3]))
        print("Nombre de films à 4 étoile : " + str(notes[4]))
        print("Nombre de films à 5 étoile : " + str(notes[5]))

--- Temps d'execution : 0.7095688661519272 secondes ---

Nombre de films à 1 étoile : 6110
Nombre de films à 2 étoile : 11370
Nombre de films à 3 étoile : 27145
Nombre de films à 4 étoile : 34174
Nombre de films à 5 étoile : 21201
```