

TD1 Web Scrapping sur le texte

Introduction :

L'enjeu de ce TD est de programmer en python des scripts pour scraper du texte (nettoyer des pages html) en utilisant des modules python.

L'utilisation de différents outils nous aide à faire une comparaison entre la théorie et la pratique sur l'analyse des fichiers.

Il nous a été donné un ensemble de fichier : un dossier clean et un dossier html contenant les fichiers à scraper.

Les fichiers présents dans le dossier clean sont très important car il constitue la référence pour établir notre analyse.

Ainsi, nous pouvons nous faire une idée en fonction des outils quel est celui qui est le plus proche de la référence

Exercice 1 :

L'enjeu de l'exercice 1 c'est :

- Connaître le nombre de ligne des fichiers clean
- Connaître la moyenne du nombre de ligne
- Connaître l'écart type

Avec nos différents tests, nous avons remarqué un énorme écart quand on ne renseignait pas la langue du fichier qu'on analysait. De plus, just text nous impose de mettre une langue sinon il ne fonctionne pas. Nous avons choisi "English", mais ce n'est absolument pas fiable car on aurait pu choisir une autre langue et obtenir des résultats différents.

Concernant le nombre de caractères, nous constatons une grande disparité entre l'outil just text et l'outil BeautifulSoup.

Just text:

```
for filename in os.listdir(directory):
    count += 1
    nblin = 0
    completeName = os.path.join(save_path, filename)
    cf = open(completeName, "w+")
    f = open(directory+"/"+filename, "r", encoding="utf-8", errors='ignore')

    paragraphs = justext.justext(f.read(), justext.get_stoplist("English"))

    for paragraph in paragraphs:
        if not paragraph.is_boilerplate:
            cf.write("<p>"+paragraph.text+"</p>"+'\n')
            nblin+=1
    data.append(nblin)
    #data.append(len(open(completeName).readlines()))

print("La moyenne : " + str(statistics.mean(data)))
print("l'ecart-type : " + str(statistics.stdev(data)))
```

Run: JT

/usr/bin/python3.7 /home/romaric/PycharmProjects/scrapping/JT.py
La moyenne : 4.919126328217238
l'ecart-type : 11.069984115648595
Process finished with exit code 0

```
filelist = os.listdir(save_path+"/")
somme = 0
for file in filelist:
    chars = 0
    lines = 0
    fchar = open(save_path+"/"+ file+'.r')
    for line in fchar.readlines():
        lines +=1
        for char in line:
            chars +=1
    data.append(chars)

print("La moyenne : " + str(statistics.mean(data)))
print("l'ecart-type : " + str(statistics.stdev(data)))

#163000 et 964
```

Run: JTchar

/usr/bin/python3.7 /home/romaric/PycharmProjects/scrapping/JTchar.py
1947151
La moyenne : 941.5623791102514
l'ecart-type : 2394.8850607615545
Process finished with exit code 0

Fichiers du dossier Clean

1. Taille des lignes

The screenshot shows the PyCharm IDE interface. On the left, the Project view displays a directory structure for a project named 'scrapping'. The 'JT.py' file is selected. The main editor window shows the code for 'statsligne.py':

```
1 import os
2 import statistics
3
4 path = "./Corpus_detourage/clean/"
5
6 data = []
7
8 filelist = os.listdir(path)
9
10 for file in filelist:
11     data.append(len(open(path+file).readlines()))
12
13 print("La moyenne : " + str(statistics.mean(data)))
14 print("l'écart-type : " + str(statistics.stdev(data)))
```

Below the editor, the Run window shows the execution of 'statsligne.py' using Python 3.7. The output is:

```
/usr/bin/python3.7 /home/romaric/PycharmProjects/scrapping/stats.py
La moyenne : 13.68529826209488
l'écart-type : 10.889377753346462
Process finished with exit code 0
```

2. Taille des caractères

The screenshot shows the PyCharm IDE interface. On the left, the Project view displays a directory structure for a project named 'scrapping'. The 'stats.py' file is selected. The main editor window shows the code for 'stats.py':

```
1 import os
2 import statistics
3
4 path = "./Corpus_detourage/clean/"
5
6 data = []
7
8 filelist = os.listdir(path)
9
10 for file in filelist:
11     f = open(path + file, 'r')
12     for line in f.readlines():
13         data.append(len(line))
14
15 print("La moyenne : " + str(statistics.mean(data)))
16 print("l'écart-type : " + str(statistics.stdev(data)))
```

Below the editor, the Run window shows the execution of 'stats.py' using Python 3.7. The output is:

```
/usr/bin/python3.7 /home/romaric/PycharmProjects/scrapping/stats.py
La moyenne : 171.21842394288853
l'écart-type : 195.46096518732966
```

Fin exercice 1 :

Comparaison des différentes méthodes avec les fichiers de référence

Méthode JustText

```

60
61 path = "./Corpus_detourage/clean/"
62 data = []
63 filelist = os.listdir(path)
64 for file in filelist:
65     data.append(len(open(path + file).readlines()))
66
67 return data
68
69 def compare(cleanStats, jtStats):
70     comparaison = []
71     for i in range(len(cleanStats)):
72         comparaison.append(abs(cleanStats[i] - jtStats[i]))
73     return comparaison
74
75 print(compare(cleanStats(), data))
76
compare()

```

Run: JTchar

```

/usr/bin/python3.7 /home/romaric/PycharmProjects/scrapping/JTchar.py
[43, 11, 11, 7, 8, 20, 5, 16, 36, 6, 8, 12, 11, 26, 5, 1633, 16, 16, 8, 3990, 9, 9, 3852, 6708, 26, 17, 2612, 5, 8, 11, 5, 29, 5, 12, 5226,

```

Synthèse des résultats de l'exercice 1

-----Résultats pour les lignes bp-----

Le nombre total de lignes est: 25146

La moyenne : 14.844155844155845

l'écart-type : 13.976667273782345

-----Résultats pour les caractères bp -----

Le nombre total de caractères est: 5959156

La moyenne : 3517.801652892562

l'écart-type : 2831.8938456995365

-----Résultats pour les lignes jt-----

Le nombre total de lignes est: 9541

La moyenne : 5.632231404958677

l'écart-type : 13.84164684346534

-----Résultats pour les caractères jt -----

Le nombre total de caractères est: 1700082

La moyenne : 1003.5903187721369

l'écart-type : 2551.1318771297415

-----Résultats pour les lignes bs-----

Le nombre total de lignes est: 1912713

La moyenne : 1129.1103896103896

l'écart-type : 1075.816976980987

-----Résultats pour les caractères bs -----

Le nombre total de caractères est: 60013800

La moyenne : 35427.27272727273

l'écart-type : 39172.801619634585

-----Résultats pour les lignes clean-----

Le nombre total de lignes est: 22835

La moyenne : 13.479929161747343

l'écart-type : 9.748727864409059

-----Résultats pour les caractères clean -----

Le nombre total de caractères est: 3890218
La moyenne : 2296.468713105077
l'écart-type : 1982.1266929040776

---Comparaison avec clean pour les caractères---
moyenne just text différence 1763.8961038961038
ecart type just text différence 2440.818708293154

moyenne bs différence 33130.804014167654
ecart type bs différence 38997.25630766418

Cet exercice, nous fait tirer une première conclusion. Si on veut avoir des résultats précis, il faut a minima rajouter de l'informations pour orienter les outils (spécifier la langue des textes)

Exercice 2 :

L'enjeu de l'exercice 2 est d'affiner nos résultats avec les méthodes JT True et JT LangID.

| | JT | JTLANGID | JtTRUE | BeautifulSoup |
|---------------------|---------|----------|-------------------|--------------------|
| Diff des caractères | 1700082 | 1757079 | 4134593 | 60013800 |
| Moyenne | 1003.59 | 1733 | 930.3760330578513 | 35427 |
| Ecart type | 2551 | 2450 | 2072.85 | 39172.801619634585 |

Résultats avec jt true et jtlangid:

-----Résultats pour les lignes jtLangID-----

Le nombre total de lignes est: 9231
La moyenne : 5.449232585596222
l'écart-type : 13.464203684725257

-----Résultats pour les caractères jtLangID -----

Le nombre total de caractères est: 1757079
La moyenne : 1037.236717827627
l'écart-type : 2568.8982927162047

-----Résultats pour les lignes jtTrueLg-----

Le nombre total de lignes est: 21836
La moyenne : 12.890200708382526

l'écart-type : 17.1619501335756

-----Résultats pour les caractères jtTrueLg -----

Le nombre total de caractères est: 4134593

La moyenne : 2440.727863046045

l'écart-type : 3040.3390445694004

moyenne jtlangid différence 1733.2591499409682

ecart type jtlangid différence 2450.902960915322

moyenne jt true lg différence 930.3760330578513

ecart type jt true lg différence 2072.8594309780797

En utilisant la langue cette fois nous constatons une diminution de la différence avec les fichiers clean.

Exercice 3 :

| OUTILS | el | | | all | | | pl | | |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|-------|
| | F | R | P | F | R | P | F | R | P |
| JT | 17,335 | 44,405 | 11,95 | 22,18 | 48,79 | 15,225 | 21,54 | 47,145 | 15,17 |
| BP | 28,6 | 32,15 | 26,705 | 40,52 | 44,655 | 37,565 | 36,79 | 40 | 35,4 |
| BS | 10,57 | 44,22 | 6,66 | 11,715 | 48,745 | 7,415 | 15,185 | 47,1 | 9,74 |
| langid | 19,455 | 46,445 | 13,69 | 24,07 | 48,305 | 17,02 | 23,055 | 46,55 | 16,63 |
| truelg | 19,455 | 46,445 | 13,69 | 24,07 | 48,305 | 17,02 | 23,055 | 46,55 | 16,63 |

Exercice 4 :

Dans cet exercice, nous nous sommes concentrés sur l'outil unfluff.

L'objectif de cet outil est de récupérer le contenu des fichiers html du dossier html. Il permet de faire une extraction du contenu en fonction des balises.

Nous avons ensuite écrit les résultats dans des fichiers.

-----Résultats pour les lignes uf-----

Le nombre total de lignes est: 17712

La moyenne : 10.455726092089728

l'écart-type : 19.26104104099297

-----Résultats pour les caractères uf -----

Le nombre total de caractères est: 1773796

La moyenne : 1047.1050767414404

l'écart-type : 1731.731114140195

moyenne uf différence 1407.7272727272727

ecart type uf différence 1811.3705693849026