La Rochelle Université

Master 2 - Informatique

Rapport : web scraping

Réalisation par:
Yassine El-AZAMI
Mamadou Cellou BAH

La Rochelle Université

Run the program under `main.py`

Requirement:

- ➔ shutil
- ➔ justext
- ➔ langid
- ➔ boilerpipe
- ➔ bs4
- ➔ cleaneval_tool

1. Taille totale des données en lignes, moyenne par fichier et écart-type

➔ Just Text

Lignes:

Nombre totals de ligne : 1671
Nombre de ligne moyenne par fichier : 0.98642266824085
Ecart type : 16.570125747752794

Caractères:

Nombre totals de ligne : 1633418
Nombre de ligne moyenne par fichier : 964.2373081463991
Ecart type : 2975.52272768191

Pourcentages des silences et bruits
Silence: 49.1145 %
Bruit : 15.2302 %

➔ BoilerPipe

Lignes:

Nombre totals de ligne : 24129

Nombre de ligne moyenne par fichier : 14.243801652892563
Ecart type : 11.524984409683489

Caractères
Nombre totals de ligne : 4017523
Nombre de ligne moyenne par fichier : 2371.6192443919717
Ecart type : 1290.8966555556162
Pourcentages des silences et bruits
Silence: 98.7013 %
Bruit : 3.0697 %

➔ BeautifulSoup
Lignes
Nombre totals de ligne : 4149
Nombre de ligne moyenne par fichier : 2.449232585596222
Ecart type : 16.381844126995613

Caractères
Nombre totals de ligne : 227538
Nombre de ligne moyenne par fichier : 134.31995277449823
Ecart type : 2969.7794200940525
Pourcentages des silences et bruits
Silence: 2.7745 %
Bruit : 26.3282 %

➔ Just Text avec utilisation de langid
Lignes
Nombre totals de ligne : 1777
Nombre de ligne moyenne par fichier : 1.0489964580873672

Ecart type : 16.621311762991215

Caractères

Nombre totals de ligne : 1660579

Nombre de ligne moyenne par fichier : 980.2709563164109

Ecart type : 2965.41840703931

Pourcentages des silences et bruits

Silence: 48.7013 %

Bruit : 15.5844 %

➔ Just Text avec utilisation de True Lg

Lignes

Nombre totals de ligne : 6478

Nombre de ligne moyenne par fichier : 3.824085005903188

Ecart type : 17.200343891567798

Caractères

Nombre totals de ligne : 4048132

Nombre de ligne moyenne par fichier : 2389.688311688312

Ecart type : 2183.009353950022

Pourcentages des silences et bruits

Silence: 97.9929 %

Bruit : 2.2432 %