Web Scraping TD1 -DUMAS JEANNE MELICZEK

> Boilerpipe n'a jamais fonctionné chez nous.
>
> Tous les résultats des exercices 1, 2 et 4 sont déjà présents dans le fichier stats.csv sur le git.
>
> L'exercice 5 n'est pas fait.

Exercice 1

Taille totale des données en lignes et caractères, moyenne et écart-type:

- Pour le dossier clean
    - Ligne :
        - Total = 22835
        - Moyenne = 13,48
        - Ecart-type = 9,75
    - Caractère :
        - Total = 3890218
        - Moyenne = 2296,47
        - Ecart-type = 1982,13
- Avec JusText
    - Ligne :
        - Total = 9541
        - Moyenne = 5.63
        - Ecart-type = 13,84
    - Caractère :
        - Total = 1700082
        - Moyenne = 1003,59
        - Ecart-type = 2551,13
        - Moyenne de difference = 1763,90
        - Ecart-type de difference = 1763,90
- Avec BeautifulSoup
    - Ligne :
        - Total = 1044436
        - Moyenne = 616,55
        - Ecart-type = 791,93
    - Caractère :
        - Total = 53067584
        - Moyenne = 31326,79
        - Ecart-type = 37447,22

- Moyenne de difference = 29031,59
- Ecart-type de difference = 37276,04

Exercice 2

> Les données manquantes sont dans stats.csv. (github)

[JUSTEXT-LANGID] Moyenne de difference = 922.6670602125148 [JUSTEXT-LANGID] Ecart-type de difference = 2063.83514170999

[JUSTEXT-TRUELG] Moyenne de difference = 922.6670602125148 [JUSTEXT-TRUELG] Ecart-type de difference = 2063.83514170999

Exercice 3

----------JusText----------

Langue	F-Score	Precision	Recall
English	79.336	82.794	83.561
Russian	1.385	99.990	0.862
Greek	0.863	98.626	0.446
Polish	0.715	99.325	0.367
Chinese	0.367	100.0	4.437
Total	**18.143**	**96.147**	**17.935**

----------JusText_langid----------

Langue	F-Score	Precision	Recall
English	79.336	82.794	83.561
Russian	70.687	70.170	80.311
Greek	86.875	90.338	87.502
Polish	78.944	76.987	86.343
Chinese	8.609	99.948	4.573
Total	**64.890**	**84.047**	**68.458**

----------JusText_trueLg----------

----------JusText_trueLg----------

Langue	F-Score	Precision	Recall
English	79.336	82.794	83.561
Russian	70.687	70.170	80.311
Greek	86.875	90.338	87.502
Polish	78.944	76.987	86.343
Chinese	8.609	99.948	4.573
Total	**64.890**	**84.047**	**68.458**

----------BeautifulSoup----------

Langue	F-Score	Precision	Recall
English	31.357	20.262	95.862
Russian	18.656	10.962	91.373
Greek	28.457	19.794	95.936
Polish	33.791	22.499	93.002
Chinese	6.160	3.520	81.128
Total	**23.684**	**15.407**	**91.460**

----------Unfluff----------

Langue	F-Score	Precision	Recall
English	83.419	88.191	81.862
Russian	1.508	97.052	0.973
Greek	1.316	98.520	0.852
Polish	45.998	84.175	41.447
Chinese	8.41	100.0	4.437
Total	**28.131**	**93.588**	**25.914**

Exercice 4

- Avec Unfluff
    - Ligne :
        - Total = 17712
        - Moyenne = 10.45
        - Ecart-type = 19,26
    - Caractère :
        - Total = 1773796
        - Moyenne = 1047,11
        - Ecart-type = 1731,73
        - Moyenne de difference = 1407,73
        - Ecart-type de difference = 1811,37