TP -Webscrapping

Membres de groupe:
Mohamed Kohen - Ahmed Kharroubi - Yassine Achour

Exercice 1 : Utilisation d'outils de détourage

1- Taille totale des données en nombre de lignes, moyenne (et écart-type) des différences par fichier

Les tableaux ci dessous représentent les statistiques des fichiers scrappé avec les outils JT et BeautifulSoup en comparaison avec le dossier Clean.

Clean/	
nombre de lignes	29136
Moyenne ligne	13.69
Ecart type	10.89

Justext	
Nombre de lignes	9542
Moyenne ligne	5.65
Ecart type	13.85

Beautifulsoup	
Nombre de lignes	9545
Moyenne ligne	5.65
Ecart type	8.1

Nous constatons que les résultat sont éloignés des résultats attendus si nous ne définissons pas de langue précise.
Avec JusText, nous sommes obligé de définir un langage par défaut, nous avons dû renseigner l'anglais pour ce dernier.
On remarque que ce dernier, augmente l'écart type entre le fichiers JT et clean.
13 pour JT contre 8 pour BS.

2- Taille totale des données en nombre de caractères, moyenne (et écart-type) des différences par fichier

2-1 Différence totale en nombre de caractères

Clean/	
Nombre de caractères	4988620
Moyenne caractères	2343.175
Ecart type caractères	2433.498

Beautifulsoup	
Nombre de caractères	14698774
Moyenne caractères	8697.4994
Ecart type caractères	6719.76

Justext	
Nombre de lignes	4988620
Moyenne ligne	2343.175
Ecart type	2433.498

En observant le nombre de caractères, on s'aperçoit qu'on est loin des nombres réels, pareil pour l'écart type ce qui iñplique nos résultats.
le nombre de caractères sur **Justext** est trop bas par rapport à **beautifulsoup** est c'est dut au choix de la langue par défaut anglais ce qui influence sur nos données de retour.

JusText	
nb Diff	2984026
moyenne diff	1765.69585
Ecart type Diff	2440.53277

Beautifulsoup	
Nombre de Différence	10826559
moyenne de différence	6406.247928

Écart type Différence	6333.37060920

Exercice 2 : Guider le scrapping avec la reconnaissance de langue

Dans cette partie de l'exercice nous utilisons la bibliothèque langid pour avoir un rendu plus fiable.
La colonne “JT_langid” correspond à l’analyse de langue en utilisant l’outil de recherche de langue “langid”.
la colonne “JT_trueLg” correspond à l’analyse en utilisant les langues effectives des fichiers (avec fichier de vérité terrain “doc_lg.json”).

Nous remarquons qu’une bonne analyse nécessite de définir la bonne langue.
On remarque dans le tableau ci dessous que la différence de caractères diminue ce qui implique automatiquement la diminution de l’écart type.

	JusText	Beautifulsoup	JusText_langid	Justext_trueLg
Différence de caractères	2984026	10826559	1561735	1561735
Moyenne de la différence	1765.7	6406.2	924.1	924.1
Écart-type de la différence	2440.5	6333.4	2065.3	2065.3

Exercice 3 : Evaluation Intrinsèque

	el			en			pl			ru			zh			all		
	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P
JT	0.85	0.43	98	76	80.7	79	0.7	0.3	99	1.38	0.8	99	8.4	4.4	100	23	24	93
JT_langid	84,5	84,9	88,1	76	80.7	78.9	76.4	83.2	74.7	69.8	79.5	68,6	8,4	4,43	100	60,3	63.4	83.2
JT_truelg	84,5	84,9	88,1	76	80.7	78.9	76.4	83.2	74.7	69.8	79.5	68,6	8,4	4,43	100	60,3	63.4	83.2
BS	50,5	93	36.7	48	91.2	35.9	48	89	35	34	89	23	7.6	33	5.1	36.9	77.1	26.7

On peut remarquer que l'analyse avec les outils langid et truelg sont identiques au contraire de JusText et Beautifulsoup, on conclut que l'étude de la langue des documents est primordiale afin de récupérer les données textuelles.

Exercice 4

Pour cette partie du td nous avons choisi d'explorer l'outil **Unfluff** qui nous a permis de récupérer les contenues des pages en les séparant en plusieurs parties.

	el			en			pl			ru			zh			all		
	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P	F	R	P
JT_truelg	84,5	84,9	88,1	76	80.7	78.9	76.4	83.2	74.7	69.8	79.5	68,6	8,4	4,43	100	60,3	63.4	83.2
Unfluff	2	1	82	86	86	89	71	73	76	3	1	81	18	10.8	82	42	39	83

Les résultats ci-dessus montrent que l'outil Unfluff est moins performant que JusText par contre pour les fichiers en langue anglaise et chinoise est plus performant que les autres.