Lucie Schmidt, Julien Boissinot

Compte-rendu

Exercice 1 :

1.1 Taille totale en nombre de lignes

```
-----infos lignes clean-----
nbLignes : 29136
MoyenneLigne : 13.68529826209488
EcartTypeLignes : 10.889377753346407
----------------------------
-----infos lignes BS-----
nbLignes : 9545
MoyenneLigne : 5.64792899408284
EcartTypeLignes : 8.096499026003768
------------------------
```

```
-----infos lignes clean-----
nbLignes : 29136
MoyenneLigne : 13.68529826209488
EcartTypeLignes : 10.889377753346407
----------------------------
-----infos lignes JT-----
nbLignes : 9542
MoyenneLigne : 5.642814902424601
EcartTypeLignes : 13.851671127282081
------------------------
```

Nous avons seulement effectué les vérifications sur “JusText” et “BeautifulSoup” car “BoilerPipe” ne fonctionnait pas sur notre machine.
Les résultats sont très éloignés du résultat attendu (9500 contre 29000) si on ne définit pas les langues de chaque document à analyser. Cependant, pour “JustText”, nous avons dû renseigner “english” comme langue par défaut car nous sommes obligés de renseigner une langue pour analyser. On constate que ce détail augmente considérablement l’écart-type entre le fichiers JT et clean (écart type de 13 pour JT contre 8 pour BS).

1.1 Taille totale en nombre de caractères

```
-----infos caracters clean-----
nbWords : 4988620
MoyenneWord : 2343.175199624237
EcartTypeWords : 2433.4980743068854
----------------------------
-----infos caracters BS-----
nbWords : 14698774
MoyenneWord : 8697.499408284024
EcartTypeWords : 6719.761079485178
------------------------
```

```
-----infos Carac clean-----
nbCaracs : 4988620
MoyenneCarac : 2343.175199624237
EcartTypeCaracs : 2433.4980743068854
----------------------------
-----infos Carac JT-----
nbCaracs : 1697609
MoyenneCarac : 1003.9083382613838
EcartTypeCaracs : 2547.4988944132997
-------------------------
```

Concernant le nombre de caractères, nous sommes encore loin du nombre attendu. Par ailleurs on remarque que l'écart-type est très grand ce qui implique que les résultats sont très variables d'un fichier à l'autre que ce soit pour "JustText" ou "BeautifulSoup". De plus, on voit que "JusText" possède un nombre de caractère très faible comparé à "BeautifulSoup". En effet, comme la langue par défaut est l'anglais, plusieurs fichiers ne retournent aucune données après analyse.

Différence totale en nombre de caractères

```
-----infos différence carac JT-----
nbDiff : 2984026
MoyenneDiff : 1765.6958579881657
EcartTypeDiff : 2440.5327759960705
--------------------------------
```

```
-----infos différence carac BS-----
nbDiff : 10826559
MoyenneDiff : 6406.247928994083
EcartTypeDiff : 6333.370609209954
--------------------------------
```

Avec une différence de 1765 caractères en moyenne et avec un écart-type de 2440 pour JT, on peut assumer que la première analyse n'est pas concluante.

Exercice 2 :

	JT	BS	JT_langid	JT_trueLg
Différence (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

Dans cet exercice, on utilise maintenant la gestion de la langue avec l'algorithme “JustText” afin de réaliser un traitement de meilleure qualité. La colonne “JT_langid” correspond à l’analyse de langue en utilisant “langid” et le colonne “JT_trueLg” correspond à l’analyse en utilisant les langues effective.

On remarque que faire l’analyse en utilisant la bonne langue permet une analyse plus performante (la différence de caractère diminue de 1 422 291). L’écart-type diminue par la même occasion.

Cependant, on observe aucune différence entre “JT_trueLg” et “JT_trueLg” car le module “languid” à toujours trouvé bonne langue qui correspond au fichier. On peut nuancer notre résultat car nous avons été obligé de renseigner la langue “anglais” si le document est en “chinois” car cette langue n’est pas prise en compte dans le module.

Exercice 3 :

	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