

PROG-M1_TD3_2020_MILLOUR

January 21, 2021

1 TD 3 : extraire des informations d'un corpus brut grâce à python et aux expressions régulières

1.1 1/ Quel est le package python qui permet de gérer les expression régulières ? importez-le et affichez l'aide correspondante avec la fonction help()

```
[ ]: # importez le package
import re
# affichez l'aide (ou ouvrez https://docs.python.org/fr/3/library/re.html)
# help(re)
```

1.2 2/ La première fonction que nous explorons est la fonction search, quels sont les paramètres de cette fonction ? Quel est le type d'objet renvoyé ?

```
[ ]: # paramètre 1 :
# paramètre 2 :
# paramètre 3 : (valeur par défaut : )
# valeur de retour :
```

1.3 3/ utiliser la fonction search pour identifier les présidents qui au cours de leurs discours d'investiture ont mentionné :

1. La Nation
2. Les enjeux climatiques (mot clé : "climat")

Utilisez les boucles (parcours de liste) et les fonctions pour rendre votre code lisible.

```
[1]: print("Question 3.1 \n")
mots = ["Nation","climat"]
```

```
fichiers = ["investiture_Chirac1.txt","investiture_Hollande.
↳txt","investiture_Mitterrand1.txt","investiture_Sarkozy.
↳txt","investiture_Chirac2.txt","investiture_Macron.
↳txt","investiture_Mitterrand2.txt","investiture_VGE.txt"]
```

Question 3.1

1.4 **## 4/** La deuxième fonction que nous utilisons est la fonction `re.compile`, à quoi sert-elle ? Quels sont les paramètres de cette fonction ? Quel est le type d'objet renvoyé ?

```
[ ]: # paramètre 1 :
     # valeur de retour :
```

1.5 **5/** Ouvrez le fichier `moi_president_de_la_republique.txt` et stockez son contenu dans une variable `discours`.

1. stockez dans une variable `moi_je` le nombre de fois que M. Hollande utilise un pronom personnel de la première personne du singulier dans cet extrait de discours (“je”, “j” et “moi” inclus) (42). Vous allez avoir besoin de l’expression régulière plus bas, utilisez la fonction `compile` pour la stocker.

- 2.

1.6 stockez dans une variable `anaphore` les phrases commençant par le syntagme « Moi président de la République » et dans une variable `anaphore_count` le nombre d’occurrences (15). Attention la variable `anaphore` doit contenir la phrase entière.

```
[2]: # ouverture du fichier ...
```

1.7 **6/** Ouvrez le fichier `reponse_sarkozy.txt` et stockez son contenu dans une variable `reponse`.

1. Combien de fois Nicolas Sarkozy utilise-t-il un pronom personnel de la première personne du singulier tel que défini ci-dessus ? répondez sans réécrire l’expression régulière correspondante.

2. Identifiez les figures de style employées par Nicolas Sarkozy (voir https://fr.wikipedia.org/wiki/Moi_pr%C3%A9sident_de_la_R%C3%A9publique), et utilisez à la manière de la question 5 des expressions régulières pour les isoler et les afficher.
3. Quelles sont les personnalités citées par Nicolas Sarkozy ?

```
[3]: # ouverture du fichier

## Question 6.1 Combien de fois Nicolas Sarkozy utilise-t-il un pronom personnel
# de la première personne du singulier tel que défini ci-dessus ?
# répondez sans réécrire l'expression régulière correspondante.
print("Question 6.1 \n")

## Question 6.2 Identifiez les figures de style employées par Nicolas Sarkozy
#(voir https://fr.wikipedia.org/wiki/Moi\_pr%C3%A9sident\_de\_la\_R%C3%A9publique),
# et utilisez à la manière de la question 5 des expressions régulières pour les
↪ isoler et les afficher.
print("\nQuestion 6.2 \n")

# Question 6.2 a) répétition

# Question 6.2 b)print("question 6.2 \n")

# Question 6.2 c) répétition + épiphore
```

Question 6.1

Question 6.2

```
[4]: # Question 6.3 Quelles sont les personnalités citées par Nicolas Sarkozy ?
print("\nQuestion 6.3 \n")
```

Question 6.3

1.8 7/ Utilisez la fonction `re.sub` pour enlever les éléments de ponctuation dans les deux textes ci-dessus. Vérifiez que les résultats correspondent à vos attentes et enregistrez les dans `moi_president_de_la_republique_mots.txt` et `reponse_sarkozy_mots.txt`. Combien chacun des textes contient-il de mots ?

```
[5]: # Question 7
print("\nQuestion 7 \n")
```

Question 7

1.9 8/ Utilisez la fonction `re.sub` remplacer les 7 premières occurrences de “Moi” dans le texte `moi_president_de_la_republique.txt` par `Vous` et stockez le résultat dans une variable `vous_president`.

```
[6]: # Question 8
print("\nQuestion 8 \n")
```

Question 8

1.10 9/ La troisième fonction que nous utilisons est la fonction `re.split`, à quoi sert-elle ? Quels sont les paramètres de cette fonction ? Quel est le type d'objet renvoyé ?

```
[ ]: # paramètre 1 :  
      # paramètre 2 :  
      # paramètre 3 : (valeur par défaut : )  
      # paramètre 4 : (valeur par défaut : )  
      # valeur de retour :
```

1.11 10/ Les expressions régulières pour traiter des fichiers csv

- Utilisez la fonction `re.split` pour isoler les différents champs contenus dans le fichier `Gargantext_CSV_Corpus_Sample.csv` (rappel : la fonction permettant de lire un document *ligne par ligne* n'est pas `read()`).
- Parcourez les listes créées par la fonction `split` pour former une liste des années de publication des ouvrages du corpus. résultat attendu : `['2011', '2011', '2012', '2012', '2012', '2012', '2012', '2013', '2013', '2013', '2013', '2013', '2013', '2013', '2013', '2013', '2013', '2013', '2013', '2013', '2013', '2014', '2014', '2014', '2014', '2014']`.

```
'2014', '2014', '2014', '2014', '2014', '2015', '2015', '2015', '2015', '2015',  
'2015', '2015', '2015', '2015', '2015', '2015', '2015', '2016', '2016', '2016']
```

```
[7]: # Question 10.1  
print("\nQuestion 10.1 \n")  
# ouverture du fichier
```

Question 10.1