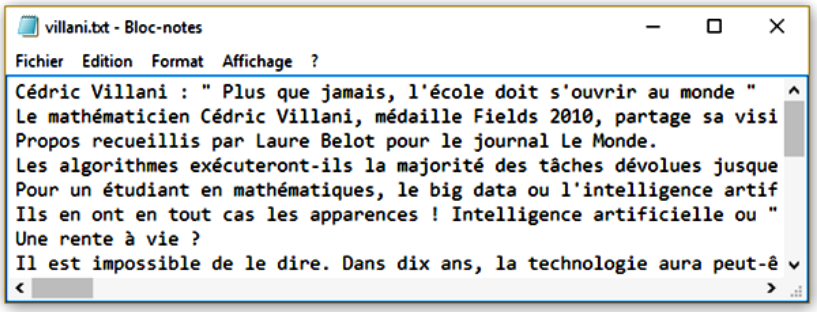
Un TP pour aller plus loin

Analyse des données d'un fichier texte

On souhaite analyser la fréquence d'occurrence des différentes lettres de l'alphabet dans le texte de Cédric Villani figurant dans le fichier « SP_villani.txt ».



💻 fichier disponible SP_villani.txt

1. Le programme Python suivant ouvre le fichier texte (le fichier txt doit se trouver dans le même répertoire que le programme), le « lit », passe les caractères en minuscules, remplace les caractères accentués et supprime les caractères non alphabétiques (ponctuation, espaces, retour à la ligne, chiffres…).
Pour compter l'occurrence de chaque lettre de l'alphabet, il utilise un « dictionnaire » D. Ce dictionnaire permettra l'accès rapide aux données.

```
# remplacer "villani.txt" par l'adresse complète du fichier
# "E:\maths\python\villani.txt"
f = open("villani.txt", 'r')
texte = f.read()

# passage en minuscules
texte = texte.lower()
# remplacement des caractères accentués
Remplacer = {"é":"e","è":"e","ê":"e","ë":"e","à":"a","â":"a","ä":"a","ù":"u",
             "ü":"u","î":"i","ï":"i","ô":"o","ö":"o","ç":"c"}
for carac in Remplacer:
    texte=texte.replace(carac, Remplacer[carac])
# suppression des caractères non alphabétiques
Alphabet = ["a","b","c","d","e","f","g","h","i","j","k","l","m","n","o","p","q",
            "r","s","t","u","v","w","x","y","z"]
for carac in texte:
    if carac not in Alphabet:
        texte = texte.replace(carac,"")

# occurence des lettres de l'alphabet
D={}
for lettre in Alphabet:
    D[lettre] = texte.count(lettre)
print(D)
f.close()
```

Un dictionnaire D est une série de « valeurs » associées à des « clés ». Le programme précédent fournit le dictionnaire suivant. À la clé « 'a' » correspond la valeur 377. On note D['a'] = 377.

```
{'a': 377, 'b': 33, 'c': 194, 'd': 176, 'e': 927, 'f': 53, 'g': 60, 'h'
: 53, 'i': 397, 'j': 25, 'k': 1, 'l': 280, 'm': 177, 'n': 411, 'o': 333
, 'p': 130, 'q': 73, 'r': 364, 's': 426, 't': 407, 'u': 321, 'v': 80, '
w': 0, 'x': 25, 'y': 17, 'z': 8}
```

a. Implémenter le programme. À quoi correspondent les valeurs du dictionnaire D ?

b. À la ligne 21 du programme, le dictionnaire est initialement vide. Quel est le rôle de la boucle des lignes 22 et 23 ?
2. On modifie la fin du programme comme ci-dessous.

```
20 # occurence des lettres de l'alphabet
21 D={}
22 for lettre in Alphabet:
23     D[lettre] = texte.count(lettre)
24
25 N = sum(D.values())
26 for a in D:
27     D[a] = D[a] / N
28
29 print(D)
30 f.close()
```

a. À quoi correspond la variable N ?
b. Que calcule la ligne 27 du programme ?
3. Modifier le programme comme ci-dessous pour construire un graphique statistique (on fait appel aux modules numpy et matplotlib.pyplot).

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3
```

```
23 # occurence des lettres de l'alphabet
24 D={}
25 for lettre in Alphabet:
26     D[lettre] = texte.count(lettre)
27
28 N = sum(D.values())
29 for a in D:
30     D[a] = D[a] / N
31
32 plt.bar(np.arange(len(D)) + 0.1, [D[a] for a in Alphabet], 0.5, color = 'r')
33 plt.xticks(np.arange(len(D)) + 0.1, [a for a in Alphabet])
34 plt.show()
35
36 f.close()
```

a. De quel type de graphique s'agit-il ?
b. Analyser le graphique obtenu.
4. Utiliser le programme pour analyser le fichier « SP_big_data.txt » correspondant à l'article, en anglais, « Big data » de l'encyclopédie Wikipedia. Comparer avec les résultats précédents.

💻 fichiers disponibles
SP_big_data.txt
SP_analyse_texte.py