2020_regex

July 1, 2022

1 Tech - expressions régulières

Les expressions régulières sont utilisées pour rechercher des motifs dans un texte tel que des mots, des dates, des nombres...

[1]:
```
from jyquickhelper import add_notebook_menu
add_notebook_menu()
```

[1]:
```
<IPython.core.display.HTML object>
```

1.1 Enoncé

Le texte suivant est un poème d'Arthur Rimbaud, Les Voyelles. On veut en extraire tous les mots.

[2]:
```
poeme = """
A noir, E blanc, I rouge, U vert, O bleu, voyelles,
Je dirai quelque jour vos naissances latentes.
A, noir corset velu des mouches éclatantes
Qui bombillent autour des puanteurs cruelles,

Golfe d'ombre; E, candeur des vapeurs et des tentes,
Lance des glaciers fiers, rois blancs, frissons d'ombelles;
I, pourpres, sang craché, rire des lèvres belles
Dans la colère ou les ivresses pénitentes;

U, cycles, vibrements divins des mers virides,
Paix des pâtis semés d'animaux, paix des rides
Que l'alchimie imprime aux grands fronts studieux;

O, suprême clairon plein de strideurs étranges,
Silences traversés des Mondes et des Anges:
¿ O l'Oméga, rayon violet de Ses Yeux!
"""
```

1.1.1 Exercice 1 : utiliser les expression régulières pour extraire tous les mots

En python, il faut utiliser le module re. Il faudra lire le paragraphe sur la syntaxe Regular Expression Syntax. Autres lectures : Expressions régulières.

[3]:
```
def extract_words(text):
    # utiliser les exrp
    pass
```

```
extract_words(poeme)
```

1.1.2 Exercice 2 : utiliser les expression régulières pour extraire tous les mots se terminant par la lettre s

[4]:

1.1.3 Exercice 3 : utiliser les expression régulières pour remplacer tous les “de” en 2

Les fonctions finditer ou sub pourraient vous être utile.

[5]:

1.1.4 Exercice 4 : utiliser les expression régulières pour extraire les lignes des rimes en `elle` ou `elles` ou `aile` ou `ailes`

La fonction finditer pourrait vous être utile.

[6]:

1.2 Réponses

1.2.1 Exercice 1 : utiliser les expression régulières pour extraire tous les mots

Les accents sont traités comme des lettres différentes par les expressions régulières. On peut soit les garder, soit les remplacer. Pour ce faire, on peut lire What is the best way to remove accents (normalize) in a Python unicode string?.

[7]:
```
import unicodedata

def strip_accents(s):
   return ''.join(c for c in unicodedata.normalize('NFD', s)
                  if unicodedata.category(c) != 'Mn')

strip_accents('têtu')
```

[7]:
```
'tetu'
```

[8]:
```
import re

def extract_words(text):
    text_sans_accent = strip_accents(text)
    return re.findall('[A-Za-z]+', text_sans_accent)

mots = extract_words(poeme)
mots[:5]
```

[8]:
```
['A', 'noir', 'E', 'blanc', 'I']
```

1.2.2 Exercice 2 : utiliser les expression régulières pour extraire tous les mots se terminant par la lettre s

On modifie le motif pour qu'il se termine par la lettre *s*. Le caractère `\b` est utilisé pour signifier que cette lettre ne peut se trouver qu'à la fin d'un mot.

[9]:
```
def extract_words_lettre(text, lettre='s'):
    text_sans_accent = strip_accents(text)
    return re.findall('[A-Za-z]+[' + lettre + ']\\b',
                      text_sans_accent)

mots = extract_words_lettre(poeme, 'se')
mots[:5]
```

[9]:
```
['rouge', 'voyelles', 'Je', 'quelque', 'vos']
```

1.2.3 Exercice 3 : utiliser les expression régulières pour remplacer tous les "de" en 2

[10]:
```
re.sub("de\\b", "2", poeme)
```

[10]:
```
"\nA noir, E blanc, I rouge, U vert, O bleu, voyelles,\nJe dirai quelque jour
vos naissances latentes.\nA, noir corset velu des mouches éclatantes\nQui
bombillent autour des puanteurs cruelles,\n\nGolfe d'ombre; E, candeur des
vapeurs et des tentes,\nLance des glaciers fiers, rois blancs, frissons
d'ombelles;\nI, pourpres, sang craché, rire des lèvres belles\nDans la colère ou
les ivresses pénitentes;\n\nU, cycles, vibrements divins des mers virides,\nPaix
des pâtis semés d'animaux, paix des rides\nQue l'alchimie imprime aux grands
fronts studieux;\n\nO, suprême clairon plein 2 strideurs étranges,\nSilences
traversés des Mondes et des Anges:\n¿ O l'Oméga, rayon violet 2 Ses Yeux!\n"
```

1.2.4 Exercice 4 : utiliser les expression régulières pour extraire les lignes des rimes en `elle` ou `elles` ou `aile` ou `ailes`

Un petit essai avant la solution.

[11]:
```
re.findall("(((aile)|(elle))s?\\b)", poeme)
```

[11]:
```
[('elles', 'elle', '', 'elle'),
 ('elles', 'elle', '', 'elle'),
 ('elles', 'elle', '', 'elle'),
 ('elles', 'elle', '', 'elle')]
```

Un autre pour se convaincre…

[12]:
```
for m in re.finditer("(((aile)|(elle))s?\\b)", poeme):
    print('%02d-%02d: %s' % (
        m.start(), m.end(), m.group(0)))
```

```
46-51: elles
182-187: elles
296-301: elles
346-351: elles
```

On mélange. On découpe en ligne d'abord, et on applique le même traitement sur chaque ligne.

[13]:
```
for i, ligne in enumerate(poeme.split('\n')):
    for m in re.finditer("(((aile)|(elle))s?\\b)", ligne):
        print('% 2d: %02d-%02d/%02d: %s' % (
            i + 1, m.start(), m.end(), len(ligne), ligne))
```

```
 2: 45-50/51: A noir, E blanc, I rouge, U vert, O bleu, voyelles,
 5: 39-44/45: Qui bombillent autour des puanteurs cruelles,
 8: 53-58/59: Lance des glaciers fiers, rois blancs, frissons d'ombelles;
 9: 43-48/48: I, pourpres, sang craché, rire des lèvres belles
```

Il ne resterait plus qu'à vérifier que la rime trouvée, le motif, se trouve à la fin de la ligne.

[14]:
```

```