Langages de script
TP n° 5 : Expressions régulières

Avant de commencer

Python gère les jeux de caractères comme UTF-8, qui permettent d'inclure des lettres accentuées. Faites quelques réglages nécessaires de votre environnement de travail :

1. ajoutez la ligne suivante dans votre fichier .bashrc :

   ```
   export LANG="fr_FR.UTF-8"
   ```

2. réglez votre console sur le codage UTF-8. Vérifiez que cela fonctionne en affichant un fichier codé en UTF-8 sur votre console,
3. vérifiez que votre éditeur de texte gère bien UTF-8 et sauvegarde les fichiers dans ce code.

Exercices d'échauffement

Exercice 1 :

Pour chaque motif donné par une expression régulière ci-dessous, dire quelles chaînes correspondent au motif.

Motif	Chaine
A.C	abc
	AZZC
	AC
	ABC

Motif	Chaine
[ABC]	A
	B
	AC
	ABC

Motif	Chaine
[a-k1-5]	A2
	a-3
	e
	1

Motif	Chaine
[^A]	A
	B
	.
	BCD

Motif	Chaine
[^ABC]	AZZ
	A
	B
	XY

Motif	Chaine
AB.|A.C|.BC	AB
	ADC
	AZZC
	AB.

Motif	Chaine
[A|B]	A
	B
	AB
	|

Les prénoms

Nous utiliserons le fichier `prenoms.txt` pour la suite, que vous trouverez sous Didel. Ce fichier provient des données *open data* sur les prénoms en France depuis 1905, disponible également sur le site `http://www.data.gouv.fr/`. Comme le fichier est assez gros, veillez à optimiser le temps de réponse de vos programmes pour que celui-ci soit raisonable.

Exercice 2 :

1. Écrire une fonction `double(s)` qui prend en entrée une chaîne `s` et retourne une chaîne `r` telle que `s = rr` si `r` existe, `None` sinon.

2. Écrire une fonction `contientdouble(s)` qui prend en entrée une chaîne `s` et retourne une chaîne `r` de longueur au moins 2 telle que `s` contient `rr` si un tel `r` existe, `None` sinon.
3. Écrire une expression régulière correspondant aux chaînes de la forme `annee,nombre, id, prenom`, où `annee` est un nombre de 4 chiffres, `nombre` est un entier quelconque, `id` est un entier quelconque, et `prenom` est une chaîne alphanumérique. La chaîne se termine par `\n`. Testez bien votre expression régulière sur tous les cas de figure.
4. Écrire une fonction `extraire_ligne(s)` qui prend en entrée une chaîne de la forme `annee,nombre,id,prenom` (sans espace après les virgules) et retourne un 4-uplet de la forme `(annee, nombre, id, prenom)`. Vous devez utiliser uniquement la fonction `re.match` ainsi que la fonction `str.strip` pour supprimer les blancs (`\t`, `\n`, ...) non désirés.
5. Écrire une fonction `extraire_fichier(f)` qui prend en entrée un nom de fichier dont chaque ligne suit le format de la question précédente, et retourne un dictionnaire associant à chaque couple `(prenom,annee)` le nombre de fois où le prénom a été choisi pendant cette année.
6. Écrire une fonction `popularite(d)` qui prend en entrée un dictionnaire dont les clefs sont des paires (prenom, annee) et les valeurs sont le nombre de fois où le prénom a été choisi cette année, et retourne un dictionnaire dont les clefs sont les prénoms et la valeur associée est le nombre total de fois où le prénom a été choisi, sur toutes les années.
7. Écrire une fonction `liste_prenoms(d)` qui prend en entrée un dictionnaire dont les clefs sont des paires (prenom, annee) et retourne la liste des prénoms sans doublons.

Exercice 3 :

Écrire un script pour répondre aux questions suivantes qui portent sur les données du fichier `prenoms.txt` :

1. Combien y a-t-il de prénoms doubles (au sens de la fonction `double(s)` ci-dessus) ? (N'oubliez pas que la première lettre est majuscule...) ?
2. Combien y a-t-il de prénoms contenant un double (au sens de la fonction `contientdouble(s)` ci-dessus) ?
3. Combien y a-t-il de prénoms qui contiennent 5 fois la même lettre ?
4. Combien y a-t-il de prénoms qui contiennent au moins 4 voyelles consécutives ?
5. Combien y a-t-il de prénoms qui contiennent 4 consonnes consécutives ?
6. Combien y a-t-il de prénoms qui finissent par 4 consonnes consécutives ?
7. Quelle est la proportion de prénoms composés (contenant un espace ou un trait d'union) ?
8. Quelle proportion de la population née entre 1905 et 2015 a un prénom composé ?
9. Quel est le prénom le plus populaire qui commence par une voyelle ? On compte pour chaque prénom la somme des fois qu'il a été choisi sur l'ensemble des années.
10. Pour chaque lettre, quel est le prénom le plus populaire qui commence par cette lettre ?