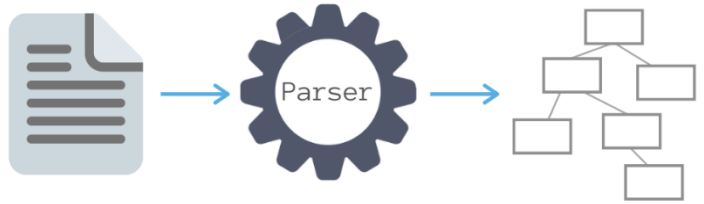
Traitement des chaînes de caractères

Résumé



Résumé

- Une chaîne de caractère peut être **convertie en liste** pour être traitée, soit :
  - Lettre par lettre par énumération
  - Selon un découpage explicite avec .split().
- Python dispose de différents outils pour spécifiques pour les chaînes de caractères :
  - La conversion de casse (.lower(), .upper())
  - La recherche avec .find() et remplacer avec .replace()
  - Des test sur la nature des caractères : .isnumeric(), .isalpha() et .isalnum().

Résumé

- Néanmoins c'est avec les expression rationnel que l'on trouve les outils qui permettent de traiter les motifs complexes :
  - La librairie **re** avec les fonctions findall(), sub(),…
- Dans une expressions rationnelle on construit un motif à partir d'indications sur :
  - Le types et la répétition de groupes de caractères
  - D'opérations logiques
  - De positionnement par rapport à la ligne.

Se sont des outils de base du traitement de données.

Exemple vu précédemment :

Parsing HTML :

```
from re import sub
s = 'Un texte <strong>HTML<strong/> avec des balises'
s += ' et même<script type="text/javascript">'
s += ' var i = 5 ;</script> du javascript dedans.'
s1 = sub('<.*>', '', s)
s2 = sub('<[a-z\\/"=\s]*>', '', s)
s3 = sub('<[^>]*>', '', s)
print(s1)
print(s2)
print(s3)
```

```
Un texte  du javascript dedans.
Un texte HTML avec des balises et même var i = 5 ; du javascript dedans.
Un texte HTML avec des balises et même var i = 5 ; du javascript dedans.
```

Exercice : Parsing (X)HTML

Dans cette exercice on appelle balise tout bloc de caractères compris entre un unique caractère < et un unique caractère >, inclus.

Par exemple :

- est une balise
- X n'en est pas une, c'est un texte qui contient deux balises.

L'objectif est de trouver les balises et de les interpréter ou les remplacer par une autre forme de codage de l'information.

Exercice : Parsing (X)HTML

1. Proposez une fonction *balise2dico(t)* qui renvoie un dictionnaire des balises issues d'une chaîne de caractères t, indexées par le numéro du 1er caractère de la balise dans le texte.

   Par exemple :

```
balise2dico('Bonjour chers <strong>camarades</strong> et amis')
```

```
{14: '<strong>', 31: '</strong>'}
```

   *Petite subtilité : faire en sorte que le dictionnaire puisse contenir les balises doublonées.

Exercice : Parsing (X)HTML

2. Proposez une fonction *interpreter(t)* qui supprime les balise d'un texte mais, lorsqu'un bloc de texte est compris entre une balise <h*> (l'* représentant un chiffre), fait passer le bloc en majuscules.

3. Proposez une fonction une fonction *xml2csv(t)* qui prend en argument des balises xml dont ont suppose connue la structure et renvoie un tableau de données sous format csv.

   Par exemple :

```
s2 = "<individu><taille>159</taille><poids>57</poids></individu>"
s2 += "<individu><taille>168</taille><poids>71</poids></individu>"

xml2csv(s2)
```

```
' 159,57; 168,71;'
```