ESILV LÉONARD DE VINCI | ÉCOLE D'INGÉNIEURS PARIS-LA DÉFENSE

Python for Data Analysis
ESILV S1 2021–2022
[Tutorial 4: Web Scraping]

Regular expression operations

Exercise 1 [Regex]

A ***regular expression*** is a special sequence of characters that helps you match or find other strings or sets of strings, using a specialized syntax held in a pattern.

Python has a built-in package called re, which can be used to work with Regular Expressions.

You can find details here: https://docs.python.org/3/library/re.html#regular-expression-syntax

1. Import the re module :

```
import re
```

2. Write a regexp to extract the score, Cristiano's age, number of goals and the number of selections from the following text:

> Buteur face au Qatar (3-0), samedi en amical, Cristiano Ronaldo (36 ans) a fait tomber un nouveau record. En effet, l'attaquant de Manchester United, meilleur buteur de l'histoire des sélections (112 buts), est devenu le footballeur européen le plus capé avec son équipe nationale. Il compte désormais 181 matchs avec la formation lusitanienne, soit un de plus que Sergio Ramos avec l'Espagne. Le Red Devil n'est plus qu'à cinq longueurs de Bader Ahmed al-Mutawa (Koweït), recordman absolu à l'échelle mondiale.

3. Write a regexp to get prices from the following text:

> Vous cherchez un nouvel ordinateur portable performant, fiable et en réduction? Sur Rakuten, l'ordinateur Apple MacBook Air 2020 est en promotion, avec une remise de près de 230 euros pour un tarif final qui chute sous les 900 euros grâce à la plateforme marchande.

4. Write a regexp to get reduction values, along with promo code:

> - 40% sur tous les réservoirs d'essence et collecteurs d'échappement (achats internet uniquement) avec le code promo : PRINTEMPS
>
> - 40% sur tous les carénages plastique (achats internet uniquement) avec le code promo : CARENAGE40
>
> - 50% sur toutes les rampes d'injection (achats internet uniquement) avec le code promo : INJECTION50

Parsing the HTML with BeautifulSoup

Exercise 2 [BeautifulSoup**]**

BeautifulSoup is a Python library for parsing HTML and XML documents. It is often used for web scraping. BeautifulSoup transforms a complex HTML document into a complex tree of Python objects, such as tag, navigable string, or comment.

Import the BeautifulSoup class from the bs4 module:

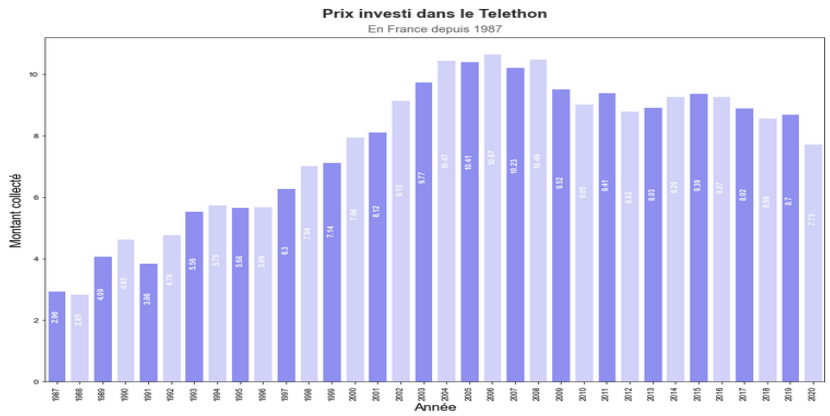
```
from bs4 import BeautifulSoup
```

Use case 1: https://www.afm-telethon.fr/telethon/bref/parrains-resultats-telethon-1379

1. Query the website and return the html into the variable 'page'.
2. Parse the html using beautiful soup and store in variable 'Soup'.
3. Save the Soup variable output into an HTML file and open it using web browser. What do you remark?
4. In 2006, the counter reached its highest level. Extract this information from the webpage.

   PS. Do not use the table.

   Extract the same information using the selector then using the HTML tag.
5. Retrieve the publication date from the webpage, using the CSS classes.
6. Retrieve the number of days remaining until the next telethon from the webpage.
7. Retrieve all hidden inputs from the webpage.
8. Find the number of tables defined in the soup. Retrieve this table from the webpage.
9. Retrieve the items from this table. Create a python dataframe representing this table (do not forget to include the name of the columns as defined in the webpage).
10. Visualize the evolution of the amount collected per year.
11. Mark the maximum and minimum of the evolution with respectively red and green colors.
12. Bonus question: We expect the graph bellow:



Exercise 3 [BeautifulSoup**]**

Use case 2: https://www.infoclimat.fr/observations-meteo/archives/7/septembre/2019/paris-montsouris/07156.html

1. Store the html table in a dataframe.
   Define a function that allows you to retrieve the value of row i and column j of the table
2. Explore the data using visualization libraries. Interpret and conclude.