

TP7 - Web scraping

Octobre 2021

Objectifs : découvrir le module *BeautifulSoup* et apprendre à extraire des informations d'une page html.

Temps : 2 séances (4h00)

1 Avant de commencer

Il est demandé d'utiliser la version 3 de Python

Dans les exercices suivants, vous devrez créer un fichier par exercice (par exemple *exercice1.py* et *exercice2.py*). Le but étant d'aller lire des données sur la carte du site web de Domino's, vous la trouverez au lien suivant :

<https://www.dominos.fr/la-carte/nos-pizzas>

Vous pouvez la télécharger sur votre machine grâce à la commande suivante (sur linux) :

```
wget https://www.dominos.fr/la-carte/nos-pizzas
```

Pour ces exercices, vous utiliserez le module Python *BeautifulSoup*. Le lien vers sa documentation est :

<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

2 Afficher la carte de Domino's (10 points)

Téléchargez la carte de *Domino's* grâce à la commande *wget* et écrivez un algorithme capable d'afficher les noms de toutes leurs pizzas (et uniquement les noms des pizzas !).

3 Afficher quelques détails (10 points)

Après avoir téléchargé les pages nécessaires, affichez les **ingrédients** et le nombre de **kilo calories** (pour 100gr) de la pizza *Bacon Groovy* en **taille médium classique** et de la pizza *Gamberetti*.

Affichez le nom de la pizza la moins calorique.

4 Consignes de rendu

Lorsque votre TP est terminé, récupérez les fichiers de votre programme *.py*, déposez les dans un dossier qui a pour nom "prénom1_nom1_prénom2_nom2.TPX" en remplaçant *prénom* et *nom* par ceux des binômes et en remplaçant le TP *X* par le numéro de TP. Comprimez le sous format *.zip*, *.rar* ou *tar.gz* puis envoyez le à l'adresse **mickael.bettinelli@lcis.grenoble-inp.fr** .