Home

You can type something here...

Web scraping Via BeautifulSoup

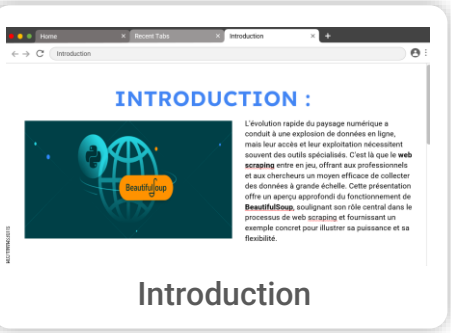
Réalisé par : HADOU Mohamed

Twitter Wikipedia Google SlidesMania

Home | Table de matière

You can type something here...

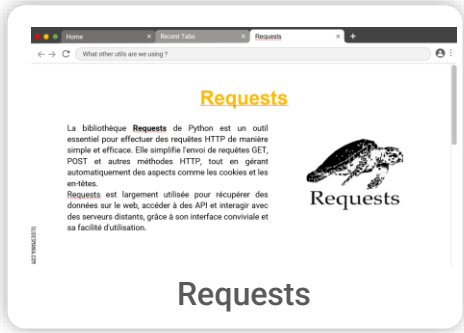
Table de matière



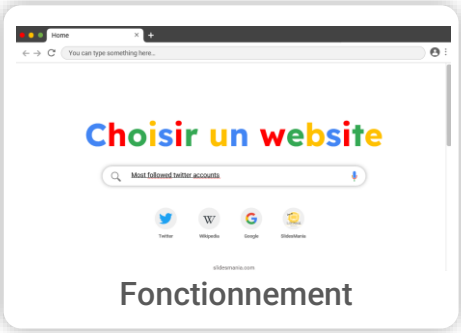
Introduction

Web scraping

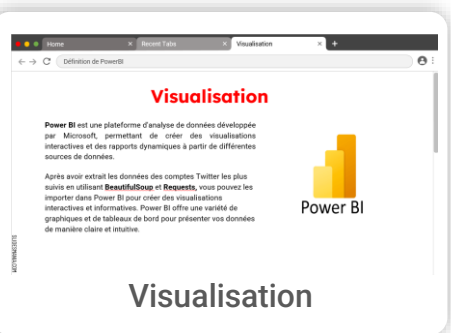
BeautifulSoup



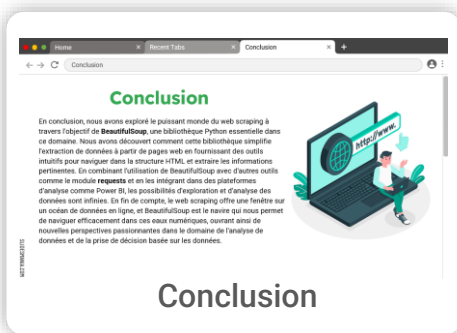
Requests



Fonctionnement

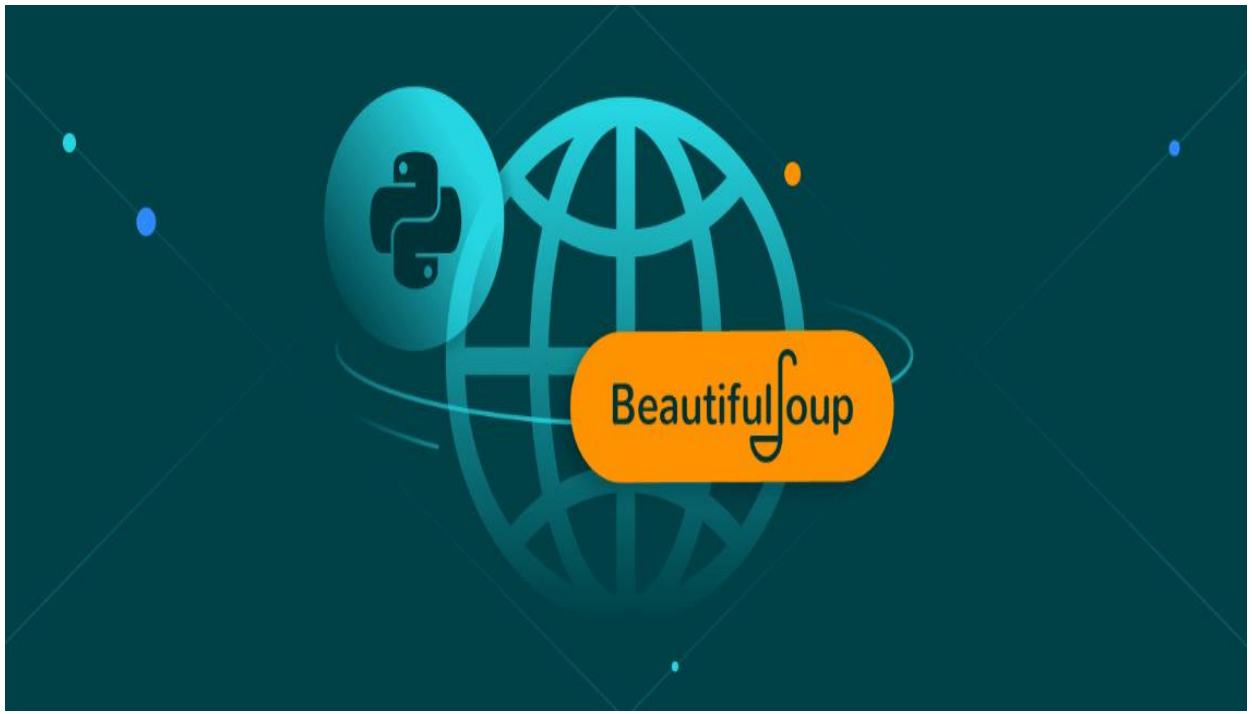


Visualisation



Conclusion

INTRODUCTION :



L'évolution rapide du paysage numérique a conduit à une explosion de données en ligne, mais leur accès et leur exploitation nécessitent souvent des outils spécialisés. C'est là que le **web scraping** entre en jeu, offrant aux professionnels et aux chercheurs un moyen efficace de collecter des données à grande échelle. Cette présentation offre un aperçu approfondi du fonctionnement de **BeautifulSoup**, soulignant son rôle central dans le processus de web scraping et fournissant un exemple concret pour illustrer sa puissance et sa flexibilité.

WEB SCRAPING :

Le web scraping, une technique informatique qui consiste à extraire automatiquement des données à partir de pages web. Cela implique l'utilisation de scripts ou de programmes pour parcourir et récupérer des données structurées ou non structurées à partir du contenu HTML ou XML d'un site web, telles que du texte, des images ou des liens. Le web scraping est largement utilisé dans divers domaines, notamment la collecte de données pour la recherche, l'analyse de marché, la surveillance des prix, et bien d'autres encore..

BeautifulSoup :

BeautifulSoup est une bibliothèque Python largement utilisée pour extraire des données à partir de documents HTML et XML. Elle simplifie le processus d'analyse et de manipulation de ces documents en fournissant des fonctionnalités pour parcourir et rechercher leur structure, ainsi que pour extraire des informations spécifiques. BeautifulSoup agit comme un parseur, permettant aux développeurs d'accéder facilement aux éléments du document, tels que les balises, les attributs et le texte, et de les manipuler selon leurs besoins. Cette bibliothèque est utilisée pour extraire des données à partir de pages web de manière efficace et flexible.

Requests

La bibliothèque **Requests** de Python est un outil essentiel pour effectuer des requêtes HTTP de manière simple et efficace. Elle simplifie l'envoi de requêtes GET, POST et autres méthodes HTTP, tout en gérant automatiquement des aspects comme les cookies et les en-têtes.

Requests est largement utilisée pour récupérer des données sur le web, accéder à des API et interagir avec des serveurs distants, grâce à son interface conviviale et sa facilité d'utilisation.

Home

You can type something here...

Choisir un website

Most followed twitter accounts

Twitter

Wikipedia

Google

SlidesMania

Fonctionnement

[1]:
```
pip install bs4
```

```
Collecting bs4
  Obtaining dependency information for bs4 from https://files.pythonhosted.org/packages/51/bb/bf7aab772a159614954d84aa832c129624ba6c32faa559dfb200a534e50b/bs4-0.0.2-py2.py3-none-any.whl.metadata
  Downloading bs4-0.0.2-py2.py3-none-any.whl.metadata (411 bytes)
Requirement already satisfied: beautifulsoup4 in c:\users\pc\anaconda3\lib\site-packages (from bs4) (4.12.2)
Requirement already satisfied: soupsieve>1.2 in c:\users\pc\anaconda3\lib\site-packages (from beautifulsoup4->bs4) (2.4)
Downloading bs4-0.0.2-py2.py3-none-any.whl (1.2 kB)
Installing collected packages: bs4
Successfully installed bs4-0.0.2
Note: you may need to restart the kernel to use updated packages.
```

[2]:
```
pip install requests
```

```
Requirement already satisfied: requests in c:\users\pc\anaconda3\lib\site-packages (2.31.0)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\pc\anaconda3\lib\site-packages (from requests) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in c:\users\pc\anaconda3\lib\site-packages (from requests) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\pc\anaconda3\lib\site-packages (from requests) (1.26.16)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\pc\anaconda3\lib\site-packages (from requests) (2023.7.22)
Note: you may need to restart the kernel to use updated packages.
```

[3]:
```
pip install lxml
```

```
Requirement already satisfied: lxml in c:\users\pc\anaconda3\lib\site-packages (4.9.3)
Note: you may need to restart the kernel to use updated packages.
```

Home | Recent Tabs | Fonctionnement

How to scrape a website using BeautifulSoup ?

Fonctionnement

[56]:
```
from bs4 import BeautifulSoup # this module helps in web scrapping.
import requests # this module helps us to download a web page
import csv
```

[57]:
```
website='https://en.wikipedia.org/wiki/List_of_most-followed_Twitter_accounts'
result = requests.get(website)
content= result.text

soup = BeautifulSoup(content,'lxml')
print(soup.prettify())
```

```
  <title>
   List of most-followed Twitter accounts - Wikipedia
  </title>
  <script>
   (function(){var className="client-js vector-feature-language-in-header-enabled vector-feature-language-in-main-page-header-disabled vector-feature-
sticky-header-disabled vector-feature-page-tools-pinned-disabled vector-feature-toc-pinned-clientpref-1 vector-feature-main-menu-pinned-disabled vecto
r-feature-limited-width-clientpref-1 vector-feature-limited-width-content-enabled vector-feature-custom-font-size-clientpref-0 vector-feature-client-p
references-disabled vector-feature-client-prefs-pinned-disabled vector-feature-night-mode-disabled skin-theme-clientpref-day vector-toc-available";var
cookie=document.cookie.match(/(?:^|; )enwikimwclientpreferences=([^;]+)/);if(cookie){cookie[1].split('%2C').forEach(function(pref){className=classNam
e.replace(new RegExp('(^| )'+pref.replace(/-clientpref-\w+$|[^\w-]+/g,'')+'-clientpref-\\w+( |$)'),'$1'+pref+'$2');});}document.documentElement.classN
ame=className;}());RLCONF={"wgBreakFrames":false,"wgSeparatorTransformTable":["",""],
"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":["","January","February","March","April","May","June","July","August","Sept
ember","October","November","December"],"wgRequestId":"016d8bb0-59ce-4b75-a0a7-59f52f148b38","wgCanonicalNamespace":"","wgCanonicalSpecialPageName":fa
lse,"wgNamespaceNumber":0,"wgPageName":"List_of_most-followed_Twitter_accounts","wgTitle":"List of most-followed Twitter accounts","wgCurRevisionId":1
220309248,"wgRevisionId":1220309248,"wgArticleId":52247588,"wgIsArticle":true,"wgIsRedirect":false,"wgAction":"view","wgUserName":null,"wgUserGroups":
["*"],"wgCategories":["CS1 German-language sources (de)","Articles with short description","Short description is different from Wikidata","Use mdy dat
es from July 2022","Dynamic lists","Articles containing potentially dated statements from April 2024","All articles containing potentially dated state
ments","Lists of Internet-related superlatives","Twitter accounts","Twitter-related lists",
```

Fonctionnement

[58]:
```
res = soup.find(id = "content")
print(res)
```

```
<main class="mw-body" id="content" role="main">
<header class="mw-body-header vector-page-titlebar">
<nav aria-label="Contents" class="vector-toc-landmark" role="navigation">
<div class="vector-dropdown vector-page-titlebar-toc vector-button-flush-left" id="vector-page-titlebar-toc">
<input aria-haspopup="true" aria-label="Toggle the table of contents" class="vector-dropdown-checkbox" data-event-name="ui.dropdown-vector-page-titleb
ar-toc" id="vector-page-titlebar-toc-checkbox" role="button" type="checkbox"/>
<label aria-hidden="true" class="vector-dropdown-label cdx-button cdx-button--fake-button cdx-button--fake-button--enabled cdx-button--weight-quiet cd
x-button--icon-only" for="vector-page-titlebar-toc-checkbox" id="vector-page-titlebar-toc-label"><span class="vector-icon mw-ui-icon-listBullet mw-ui-
icon-wikimedia-listBullet"></span>
<span class="vector-dropdown-label-text">Toggle the table of contents</span>
</label>
<div class="vector-dropdown-content">
<div class="vector-unpinned-container" id="vector-page-titlebar-toc-unpinned-container">
</div>
</div>
</div>
</nav>
<h1 class="firstHeading mw-first-heading" id="firstHeading"><span class="mw-page-title-main">List of most-followed Twitter accounts</span></h1>
```

[59]:
```
table = soup.find('table', class_='wikitable sortable')
print(table)
```

```
<table class="wikitable sortable">
<caption>
</caption>
<tbody><tr>
<th>Rank
</th>
<th>Username
</th>
<th>Owner
</th>
<th>Followers (millions)
</th>
```

Fonctionnement

[61]:
```
data = []
for row in table.find_all('tr')[1:]:
    columns = row.find_all('td')

    # Vérifier si la ligne a toutes les colonnes attendues
    if len(columns) >= 6: # Au moins 6 colonnes sont attendues
        liste = {}
        liste['Username'] = columns[0].text.strip()
        liste['Owner'] = columns[1].text.strip()
        liste['Followers'] = columns[2].text.strip()
        liste['Description'] = columns[3].text.strip()
        # Vérifier si la colonne "Brand action" contient une action de marque
        brand_action = columns[4].find('img', alt='Yes')
        if brand_action:
            liste['Brand Action'] = 'Yes'
        else:
            liste['Brand Action'] = 'No'

        # Extraire le nom du pays à partir du lien dans la colonne "Country"
        country_link = columns[5].find('a')
        if country_link:
            liste['Country'] = country_link['title']
        else:
            liste['Country'] = '' # Si le lien vers le pays n'existe pas, laisser la valeur vide

        data.append(liste)

# Exporter les données vers un fichier CSV
with open('twitter_accounts.csv', 'w', newline='') as f:
    w = csv.DictWriter(f, ['Username', 'Owner', 'Followers', 'Description','Brand Action', 'Country'])
    w.writeheader()
    w.writerows(data)

print("Les données ont été enregistrées avec succès dans le fichier CSV!")
```

```
Les données ont été enregistrées avec succès dans le fichier CSV!
```

Visualisation

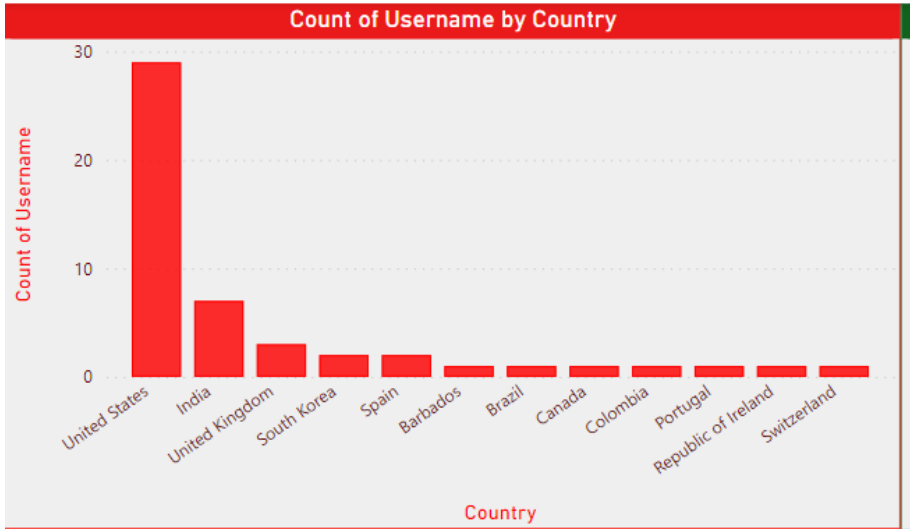
Power BI est une plateforme d'analyse de données développée par Microsoft, permettant de créer des visualisations interactives et des rapports dynamiques à partir de différentes sources de données.

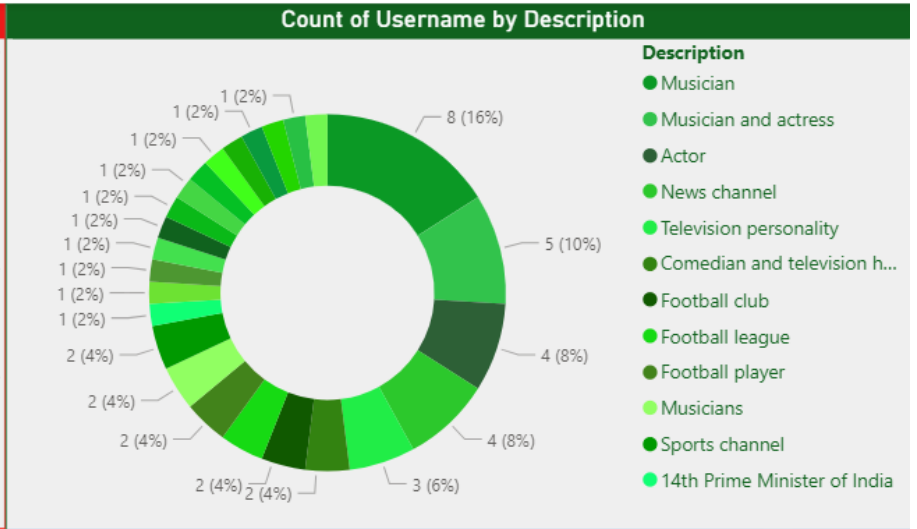
Après avoir extrait les données des comptes Twitter les plus suivis en utilisant **BeautifulSoup** et **Requests,** vous pouvez les importer dans Power BI pour créer des visualisations interactives et informatives. Power BI offre une variété de graphiques et de tableaux de bord pour présenter vos données de manière claire et intuitive.

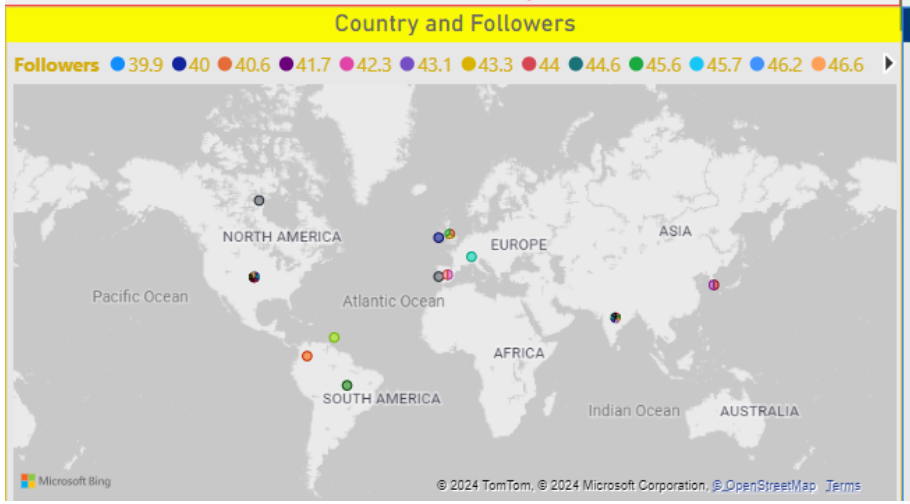
Home | Recent Tabs | Visualisation

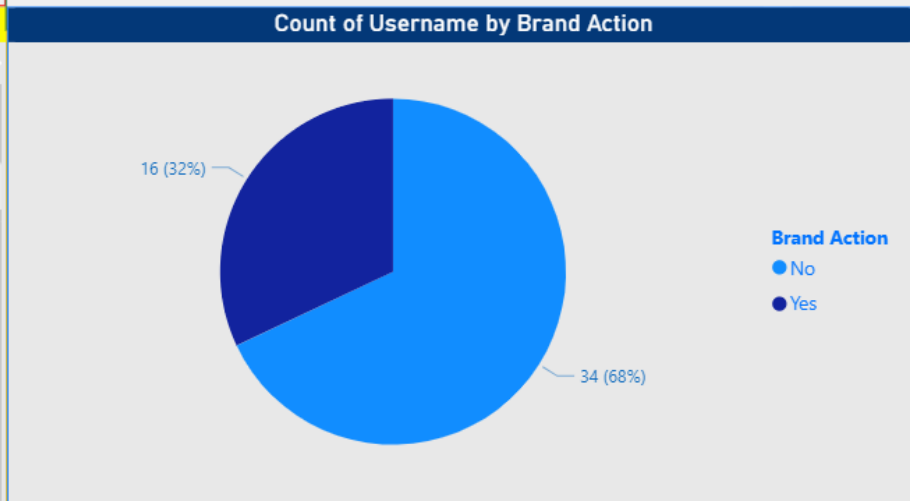
Visualisation en PowerBI

Visualisation









Conclusion

En conclusion, nous avons exploré le puissant monde du web scraping à travers l'objectif de **BeautifulSoup**, une bibliothèque Python essentielle dans ce domaine. Nous avons découvert comment cette bibliothèque simplifie l'extraction de données à partir de pages web en fournissant des outils intuitifs pour naviguer dans la structure HTML et extraire les informations pertinentes. En combinant l'utilisation de BeautifulSoup avec d'autres outils comme le module **requests** et en les intégrant dans des plateformes d'analyse comme Power BI, les possibilités d'exploration et d'analyse des données sont infinies. En fin de compte, le web scraping offre une fenêtre sur un océan de données en ligne, et BeautifulSoup est le navire qui nous permet de naviguer efficacement dans ces eaux numériques, ouvrant ainsi de nouvelles perspectives passionnantes dans le domaine de l'analyse de données et de la prise de décision basée sur les données.

Home

You can type something here...

Merci pour votre attention!

Encadré par : Benlahbib Abdessamad

Twitter Wikipedia Google SlidesMania