Projet final : Web scrapping

Présentation du projet

sujet 16 : création d'un ensemble de données de jeux vidéo populaires ou tendance plateforme de référence : steam lien de la pateforme :

Membres du groupe:

1. Ibrahima CAMARA
2. Samsidine DIATTA
3. Abdou Karim SOW

Installation et imporatation des librairies à utiliser

```
#Installation de la librairie "requests" pour le téléchargement de pages web
!pip install requests --upgrade --quiet
```

```
     |████████████████████████████████| 62 kB 815 kB/s
ERROR: pip's dependency resolver does not currently take into account all the
google-colab 1.0.0 requires requests~=2.23.0, but you have requests 2.26.0 wh
datascience 0.10.6 requires folium==0.2.1, but you have folium 0.8.3 which is
```

```
#Installation de la librairie "beautifulsoup4" pour l'extraction de données dans ur
!pip install beautifulsoup4 --upgrade --quiet
```

```
     |████████████████████████████████| 97 kB 5.8 MB/s
```

```
#Importation de la librairie "requests"
import requests
```

```
#Importation de la librairie "bs4"
from bs4 import BeautifulSoup
```

```
import pandas as pd
```

Double-cliquez (ou appuyez sur Entrée) pour modifier

Définition des fonctions à utiliser

```
#Création de la fonction permettant de télécharger le contenu d'une page web
def telechargerPage(lien):
  '''
    Objectif : télécharger la page web correspondant au lien indiqué
    Méthode : Utilisation des méthodes de la librarie <requests>
    Entrées : url de la page (string)
    sortie : contenu de la page web (tableau de string)
  '''
  reponse = requests.get(lien)
  if reponse.status_code >= 200 and reponse.status_code < 300 :
    print(f"La page web à été téléchargée avec succés !\n")
    contenu = reponse.text
    return contenu
  else :
    print(f"Echec du téléchargement de la page correspondant à ce lien '{lien}'\n v
    print(f"Status_code : {reponse.status_code}")



# Création d'une procédure permettant d'enregistrer le contenu d'une page dans un f
def ecrireFichier(nomFichier, contenu) :
  '''
    Objectif : écrire le contenu d'une page web correspondant au lien indiqué dans
    Méthode : écriture sur fichier et Utilisation de la fonction de téléchargement
    Entrées : nom du fichier(string) sans son extension et le contenu de la page (s
    sortie : pas de sortie
    Résultat : création d'un fichier html de la page web indiquée à l'instant t
  '''

  with open(nomFichier+".html", 'w', encoding='utf-8') as file :
    file.write(contenu)
  print("l'écriture du fichier s'est déroulé sans accroc !")



# Création d'une fonction permettant de créer  un objet beautifoulSoup
#pour préparer l'extraction des données
def creerObjetBs4(nomFichier) :
  '''
    Objectif : écrire le contenu d'une page web correspondant au lien indiqué dans
    Méthode : écriture sur fichier
    Entrées : nom du fichier html (string) avec son extension ".html"
    sortie : objet bs4.BeautifoulSoup pour une éventuelle extraction de données

  '''
  with open(nomFichier, 'r') as file:
    source = file.read()
    ObjetBs4 = BeautifulSoup(source, 'html.parser')
    return ObjetBs4
```

Début de scrapping de la plateforme steam

```
# consignation dans des variables, des noms et adresse de la plateformes steam
lienWeb = "https://store.steampowered.com/stats/?l=french"
nomSiteWeb = "Steam"

# téléchargement du contenu de la page steam dans la variable page
page = telechargerPage(lienWeb)
```

```
    La page web à été téléchargée avec succés !
```

```
# écriture du contenu de la page web dans un fichier html nommé <base file>
ecrireFichier("base_file", page)
```

```
    l'écriture du fichier s'est déroulé sans accroc !
```

```
# Création d'un objet BeautifulSoup à partir du fichier html pour une extraction de
bs4_objet = creerObjetBs4("base_file.html")

# Recensement de tous les cent premiers liens des jeux tendances et populaires sur
popular_game_link = bs4_objet.find_all(class_="gameLink")

popular_game_link
```

```
    [<a class="gameLink" href="https://store.steampowered.com/app/730/CounterStri
     <a class="gameLink" href="https://store.steampowered.com/app/570/Dota_2/" on
     <a class="gameLink" href="https://store.steampowered.com/app/578080/PUBG_BAT
     <a class="gameLink" href="https://store.steampowered.com/app/1172470/Apex_Le
     <a class="gameLink" href="https://store.steampowered.com/app/1063730/New_Wor
     <a class="gameLink" href="https://store.steampowered.com/app/440/Team_Fortre
     <a class="gameLink" href="https://store.steampowered.com/app/271590/Grand_Th
     <a class="gameLink" href="https://store.steampowered.com/app/1623660/MIR4/" 
     <a class="gameLink" href="https://store.steampowered.com/app/1203220/NARAKA_
     <a class="gameLink" href="https://store.steampowered.com/app/252490/Rust/" o
     <a class="gameLink" href="https://store.steampowered.com/app/431960/Wallpape
     <a class="gameLink" href="https://store.steampowered.com/app/230410/Warframe
     <a class="gameLink" href="https://store.steampowered.com/app/346110/ARK_Surv
     <a class="gameLink" href="https://store.steampowered.com/app/359550/Tom_Clan
     <a class="gameLink" href="https://store.steampowered.com/app/381210/Dead_by_
     <a class="gameLink" href="https://store.steampowered.com/app/238960/Path_of_
     <a class="gameLink" href="https://store.steampowered.com/app/1085660/Destiny
     <a class="gameLink" href="https://store.steampowered.com/app/1466860/Age_of_
     <a class="gameLink" href="https://store.steampowered.com/app/289070/Sid_Meie
     <a class="gameLink" href="https://store.steampowered.com/app/105600/Terraria
     <a class="gameLink" href="https://store.steampowered.com/app/236390/War_Thun
     <a class="gameLink" href="https://store.steampowered.com/app/1238810/Battlef
     <a class="gameLink" href="https://store.steampowered.com/app/1569040/Footbal
     <a class="gameLink" href="https://store.steampowered.com/app/304930/Unturned
     <a class="gameLink" href="https://store.steampowered.com/app/227300/Euro_Tru
     <a class="gameLink" href="https://store.steampowered.com/app/1468810/_Tale_o
     <a class="gameLink" href="https://store.steampowered.com/app/218620/PAYDAY_2
     <a class="gameLink" href="https://store.steampowered.com/app/1793660/Bta_tec
     <a class="gameLink" href="https://store.steampowered.com/app/444200/World_of
     <a class="gameLink" href="https://store.steampowered.com/app/1263850/Footbal
     <a class="gameLink" href="https://store.steampowered.com/app/1782210/Crab_Ga
     <a class="gameLink" href="https://store.steampowered.com/app/322330/Dont_Sta
     <a class="gameLink" href="https://store.steampowered.com/app/394360/Hearts_o
```

```
 <a class="gameLink" href="https://store.steampowered.com/app/221100/DayZ/" o
 <a class="gameLink" href="https://store.steampowered.com/app/548430/Deep_Roc
 <a class="gameLink" href="https://store.steampowered.com/app/1454400/Cookie_
 <a class="gameLink" href="https://store.steampowered.com/app/1506830/FIFA_22
 <a class="gameLink" href="https://store.steampowered.com/app/582010/Monster_
 <a class="gameLink" href="https://store.steampowered.com/app/413150/Stardew_
 <a class="gameLink" href="https://store.steampowered.com/app/251570/7_Days_t
 <a class="gameLink" href="https://store.steampowered.com/app/292030/The_Witc
 <a class="gameLink" href="https://store.steampowered.com/app/550/Left_4_Dead
 <a class="gameLink" href="https://store.steampowered.com/app/252950/Rocket_L
 <a class="gameLink" href="https://store.steampowered.com/app/4000/Garrys_Mod
 <a class="gameLink" href="https://steamcommunity.com/app/1329410">雀魂麻将(Mah
 <a class="gameLink" href="https://store.steampowered.com/app/1174180/Red_Dea
 <a class="gameLink" href="https://store.steampowered.com/app/594570/Total_Wa
 <a class="gameLink" href="https://store.steampowered.com/app/39210/FINAL_FAN
 <a class="gameLink" href="https://store.steampowered.com/app/377160/Fallout_
 <a class="gameLink" href="https://steamcommunity.com/app/480">Spacewar</a>,
 <a class="gameLink" href="https://store.steampowered.com/app/924970/Back_4_B
 <a class="gameLink" href="https://store.steampowered.com/app/489830/The_Elde
 <a class="gameLink" href="https://store.steampowered.com/app/582660/Black_De
 <a class="gameLink" href="https://store.steampowered.com/app/787860/Farming_
 <a class="gameLink" href="https://store.steampowered.com/app/255710/Cities_S
 <a class="gameLink" href="https://store.steampowered.com/app/261550/Mount__B
 <a class="gameLink" href="https://store.steampowered.com/app/294100/RimWorld
 <a class="gameLink" href="https://store.steampowered.com/app/8930/Sid_Meiers_
 <a class="gameLink" href="https://store.steampowered.com/app/813780/Age_of_E
```

```
#Création d'une liste composée de ces liens téléchargés pour lancer le crawling web
Gameslink = []
for i in range (100):
  link = (popular_game_link[i]["href"])
  Gameslink.append(link)


# Création de fichiers pour enregistrer les contenus web des jeux du top 100
jeux = []
compteur = 0
for link in Gameslink :
  jeux.append(telechargerPage(link))
  ecrireFichier("jeu"+str(compteur), telechargerPage(link))
  compteur += 1
```

```
    La page web à été téléchargée avec succés !

    La page web à été téléchargée avec succés !

    l'écriture du fichier s'est déroulé sans accroc !
    La page web à été téléchargée avec succés !

    La page web à été téléchargée avec succés !

    l'écriture du fichier s'est déroulé sans accroc !
    La page web à été téléchargée avec succés !

    La page web à été téléchargée avec succés !

    l'écriture du fichier s'est déroulé sans accroc !
    La page web à été téléchargée avec succés !
```

```
La page web à été téléchargée avec succés !

l'écriture du fichier s'est déroulé sans accroc !
La page web à été téléchargée avec succés !

La page web à été téléchargée avec succés !

l'écriture du fichier s'est déroulé sans accroc !
La page web à été téléchargée avec succés !

La page web à été téléchargée avec succés !

l'écriture du fichier s'est déroulé sans accroc !
La page web à été téléchargée avec succés !

La page web à été téléchargée avec succés !

l'écriture du fichier s'est déroulé sans accroc !
La page web à été téléchargée avec succés !

La page web à été téléchargée avec succés !

l'écriture du fichier s'est déroulé sans accroc !
La page web à été téléchargée avec succés !

La page web à été téléchargée avec succés !

l'écriture du fichier s'est déroulé sans accroc !
La page web à été téléchargée avec succés !

La page web à été téléchargée avec succés !

l'écriture du fichier s'est déroulé sans accroc !
La page web à été téléchargée avec succés !

La page web à été téléchargée avec succés !

l'écriture du fichier s'est déroulé sans accroc !
La page web à été téléchargée avec succés !

La page web à été téléchargée avec succés !
```

```
#transformation de ces contenus en objets BeautifulSoup pour l'extraction des info
bs4_games = []
i = 0
for game in jeux :
 bs4_games.append( creerObjetBs4("jeu"+str(i)+".html"))
 i += 1
```

detail jeu

```
# consignation des infos de chaque jeu du top 100 dans un liste
Games_detail =[]
for i in range (100):
  detail = bs4_games[i](id="genresAndManufacturer", recusive = False)
  Games_detail.append(BeautifulSoup(str(detail)))
```

description

```
# consignation de la description de chaque jeu du top 100 dans un liste
Games_desc =[]
for i in range (100):
  desc = bs4_games[i](class_="game_area_description", recusive = False)
  Games_desc.append(BeautifulSoup(str(desc)))
```

configuration

```
# consignation des infos de la configuration minimale requise pour chaque jeu du to
syswinreq = []
sysmacreq = []
syslinreq = []
for i in range (100):
  win = bs4_games[i].find_all("div",{"data-os":"win"}, recusive = False)
  mac = bs4_games[i].find_all("div",{"data-os":"mac"}, recusive = False)
  linux = bs4_games[i].find_all("div",{"data-os":"linux"}, recusive = False)
  syswinreq.append(BeautifulSoup(str(win)))
  sysmacreq.append(BeautifulSoup(str(mac)))
  syslinreq.append(BeautifulSoup(str(linux)))


# création du jeu de données
jeuDeDonnees = []
par1 = []
par2 = []
par3 = []

for i in range(100) :
  par1.append(str(Games_detail[i].text).strip())
  par2.append(str(Games_desc[i].text.strip()))
  par3.append((str(syswinreq[i].text.strip() + sysmacreq[i].text.strip() + syslinre
  jeuDeDonnees.append({"Détails" : par1[i],
                  "Description" : par2[i],
                  "Configuration minimale requise" : par3[i]})



# Traitement du jeu de données avec pandas
jeuDeDonnees = pd.DataFrame(jeuDeDonnees)


jeuDeDonnees.head()
```

	Détails	Description	Configuration minimale requise
0	[\nTitle: Counter-Strike: Global Offensive\nGe...	[\nAbout This Game\n\t\t\t\t\t\t\tCounter-Stri...	[\n\t\t\t\t\t\tWindows\t\t\t\t\t, \n\n\nMinimu...
1	[\nTitle: Dota 2\nGenre: Action, Free to Play,...	[\nReviews\n“A modern multiplayer masterpiece....	[\n\t\t\t\t\t\tWindows\t\t\t\t\t, \n\n\nMinimu...
2	[\nTitle: PUBG: BATTLEGROUNDS\nGenre: Action, ...	[\nAbout This Game\nPUBG: BATTLEGROUNDS is a b...	[\n\n\nMinimum:Requires a 64-bit processor and...

```
#  fonction pour l'Elimination des caractères indésirables
def replace_char(characters):
  return characters.replace('\n', '').replace('\t', '').replace('[','').replace(']
```

4	[\nTitle: New World\nGenre:	[\nNew World - Deluxe Edition	[\n\n\nMinimum:Requires a

```
# Elimination des caractères indésirables du jeu de données
for column in jeuDeDonnees.columns:
  jeuDeDonnees[column] = jeuDeDonnees[column].apply(replace_char)


# test d'affichage
jeuDeDonnees.head()
```

	Détails	Description	Configuration minimale requise
0	Title: Counter-Strike: Global OffensiveGenre: ...	About This GameCounter-Strike: Global Offensiv...	Windows, Minimum:OS: Windows® 7/Vista/XPProces...
1	Title: Dota 2Genre: Action, Free to Play, Stra...	Reviews“A modern multiplayer masterpiece.”9.5/...	Windows, Minimum:OS: Windows 7 or newerProcess...
2	Title: PUBG: BATTLEGROUNDSGenre: Action, Adven...	About This GamePUBG: BATTLEGROUNDS is a battle...	Minimum:Requires a 64-bit processor and operat...
3	Title: Apex Legends™Genre: Action, Adventure, ...	Reviews“The champion of Battle Royales.”9/10 –...	Minimum:Requires a 64-bit processor and operat...
4	Title: New WorldGenre: Action, Adventure, Mass...	New World - Deluxe Edition Enter the land of ...	Minimum:Requires a 64-bit processor and operat...

```
# transformation du jeu de données en dictionnaire
dictionnaire_data = jeuDeDonnees.to_dict()


#Création d'un fichier CSV pour le jeu de données

jeuDeDonnees.to_csv('Fichier.csv', index = False)


pd.read_csv('Fichier.csv')
```

	Détails	Description	Configuration minimale requise
0	Title: Counter-Strike: Global OffensiveGenre: ...	About This GameCounter-Strike: Global Offensiv...	Windows, Minimum:OS: Windows® 7/Vista/XPProces...
1	Title: Dota 2Genre: Action, Free to Play, Stra...	Reviews“A modern multiplayer masterpiece.”9.5/...	Windows, Minimum:OS: Windows 7 or newerProcess...
2	Title: PUBG: BATTLEGROUNDSGenre: Action, Adven...	About This GamePUBG: BATTLEGROUNDS is a battle...	Minimum:Requires a 64-bit processor and operat...
3	Title: Apex Legends™Genre: Action, Adventure, ...	Reviews“The champion of Battle Royales.”9/10 –...	Minimum:Requires a 64-bit processor and operat...
4	Title: New WorldGenre: Action, Adventure, Mass...	New World - Deluxe Edition Enter the land of ...	Minimum:Requires a 64-bit processor and operat...
...	...	...	...
95	Title: The Sims™ 4Genre: Casual, SimulationDev...	Digital Deluxe Edition Deluxe Edition includes...	Minimum:Requires a 64-bit processor and operat...
96	Title: Slay the SpireGenre: Indie, StrategyDev...	Reviews“The devs have taken the best of the de...	Windows, Minimum:OS: Windows XP, Vista, 7, 8/8...
97	Title: Oxygen Not IncludedGenre: Indie, Simula...	Language Packs , About This GameIn the space-c...	Windows, Minimum:Requires a 64-bit processor a...
98	Title: Fall Guys: Ultimate KnockoutGenre: Acti...	Collector's Edition Dress your Fall Guys in al...	Minimum:Requires a 64-bit processor and operat...
99	Title: NBA 2K21Genre: Simulation, SportsDevelo...	NBA 2K21 Mamba Forever Edition This edition in...	Minimum:OS: Windows 7 64-bit, Windows 8.1 64-b...

100 rows × 3 columns

✓ 0 s terminée à 13:47