01_Extraccion_Datos

September 5, 2024

[ ]:
```
# 1. Importar librerías
from selenium import webdriver
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import pandas as pd
import numpy as np
from selenium.webdriver.common.by import By
import os
import time
import requests
```

[ ]:
```
# 2. Utilizar Sellenium para descargar el archivo
url = "https://www.datosabiertos.gob.pe/dataset/
 ↪postulantes-al-concurso-de-admisi%C3%B3n-de-la-universidad-nacional-de-ingenier%C3%ADa-uni"
driver  = webdriver.Chrome( )
driver.get(url)
driver.maximize_window()
wait = WebDriverWait(driver, 15)
download_button = wait.until(EC.element_to_be_clickable((By.XPATH, '/html/body/
 ↪div[4]/div/div/section/div/div/div/div/div[2]/div/div/div/article/div/div[3]/
 ↪div/div/ul/li[1]/div/span/a[2]')))
download_button.click()
```

[ ]:
```
# 3. Leer el archivo descargado
download_dir = os.path.expanduser('~/Downloads')
files = os.listdir(download_dir)
paths = [os.path.join(download_dir, basename) for basename in files if basename.
 ↪lower().endswith(('.csv'))]
latest_file = max(paths, key=os.path.getctime)
df = pd.read_csv(latest_file)
df.head()
```

[ ]:
```
                                                IDHASH  \
0  E4287C2FE19F63C5E6641955147E36684A5A2FF8064676…
1  2DC37F0B9727B6591EC72D7A942647797A200F45D47C9E…
2  3B8677B90781D7BB8F2F967C05FA2DBBE153BBB682DF05…
```

```
3  FA366704D9E9F6FB5E5F55C1FB0CEEE973C626A5616F55…
4  E31CF8F30F3AE60B3D8A14F6E1020E9AD26EE975F0823B…

                     COLEGIO COLEGIO_DEPA COLEGIO_PROV  \
0       LA DIVINA PROVIDENCIA         LIMA         LIMA
1           86019 LA LIBERTAD       ÁNCASH       HUARAZ
2  0113 DANIEL ALOMIAS ROBLES         LIMA         LIMA
3     SEBASTIAN SALAZAR BONDY         LIMA         LIMA
4           TRILCE LOS OLIVOS         LIMA         LIMA

             COLEGIO_DIST COLEGIO_PAIS  COLEGIO_ANIO_EGRESO  \
0               SURQUILLO         PERÚ                 2020
1                  HUARAZ         PERÚ                 2017
2  SAN JUAN DE LURIGANCHO         PERÚ                 2016
3              SANTA ANITA         PERÚ                 2018
4    SAN MARTÍN DE PORRES         PERÚ                 2020

                       ESPECIALIDAD  ANIO_POSTULA  CICLO_POSTULA  …  \
0            INGENIERÍA DE SISTEMAS          2021              1  …
1  INGENIERÍA DE TELECOMUNICACIONES          2021              1  …
2               INGENIERÍA MECÁNICA          2021              1  …
3            INGENIERÍA ELECTRÓNICA          2021              1  …
4                      ARQUITECTURA          2021              1  …

          DOMICILIO_DIST ANIO_NACIMIENTO NACIMIENTO_PAIS  NACIMIENTO_DEPA  \
0  SAN JUAN DE MIRAFLORES           2004            PERÚ             LIMA
1                  HUARAZ           2001            PERÚ           ÁNCASH
2  SAN JUAN DE LURIGANCHO           2000            PERÚ             LIMA
3              SANTA ANITA          2002            PERÚ             LIMA
4           INDEPENDENCIA           2004            PERÚ             LIMA

  NACIMIENTO_PROV    NACIMIENTO_DIST       SEXO CALIF_FINAL INGRESO  \
0            LIMA  VILLA EL SALVADOR  MASCULINO        3.11      NO
1          HUARAZ             HUARAZ   FEMENINO        3.58      NO
2            LIMA        JESÚS MARÍA  MASCULINO        8.04      NO
3            LIMA               LIMA  MASCULINO       10.82      NO
4            LIMA              COMAS   FEMENINO        6.21      NO

                                        MODALIDAD
0  EXTRAORDINARIO1 - DEPORTISTAS CALIFICADOS DE A…
1                                       ORDINARIO
2                                       ORDINARIO
3            EXTRAORDINARIO2 - INGRESO DIRECTO CEPRE
4                                       ORDINARIO

[5 rows x 22 columns]
```

[ ]:
```python
# 4. Exportamos los datos sin procesar
data_folder = 'data'
if not os.path.exists(data_folder):
    os.makedirs('data')

output_file = os.path.join(data_folder, 'data_raw.csv')
df.to_csv(output_file, index=False)
```