Web Scraping para obtención de texto

Vamos a captar información de la web del Congreso de los Diputados para obtener las transcripciones de las intervenciones en la XIV legislatura (2019-Actualidad) de los siguientes representantes políticos:

- Pedro Sánchez Pérez-Castejón (GS)
- Santiago Abascal Conde (GV)
- Pablo Casado Blanco (GP)

Procedimiento

La página de las intervenciones del congreso permite filtrar en funcion de distintos campos, nosotros usaremos: "Legislatura" y "Orador". Además selecionaremos la opción "Por orden cronológico" para obtener las intervenciones de más actuales a más antigüas.

Una vez tengamos cargada la información requerida, bastará con ir haciendo click en las salidas resultantes para llegar al texto de la intervención, el cual descargaremos en un .txt que luego procesaremos para obtener solo la información relevante.

```
# Requisitos
# pip install selenium
# pip install webdriver-manager

# Librerias necesarias que utilizaremos
import os
import time
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.action_chains import ActionChains

# Genero una variable donde voy a almacenar los plenos ya capturados
con intención
# de no duplicar información que ya esta recolectada
plenosCapturados = []

def obtenerIntervenciones(orador, dirPath, numPags):

    # Configuramos las opciones de chrome para que descargue los pdfs
a los que accedamos
    # directamente sin abrirlos y en la carpeta que deseamos
    options = webdriver.ChromeOptions()
    options.add_experimental_option('prefs', {
    "download.default_directory": dirPath, # Cambiamos el directorio
```

```
de las descargas
    "download.prompt_for_download": False,
    "download.directory_upgrade": True,
    "plugins.always_open_pdf_externally": True
    })

    # Para mantener silenciado el webdriver_manager
    os.environ['WDM_LOG'] = '0'
    # Fácil instalación y despliegue del driver de
    # Selenium gracias a https://pypi.org/project/webdriver-manager/
    # Generamos la instancia del driver de Chrome de Selenium para
navegar
    driver =
webdriver.Chrome(service=Service(ChromeDriverManager().install()),opti
ons=options)
    driver.implicitly_wait(10)
    driver.maximize_window()
    # Generamos una "espera", nos servirá para decirle al driver
    # que espere hasta una condición antes de actuar u obtener
    # información de la página
    wait = WebDriverWait(driver,10)

    ### Comienzo del proceso de scrapeo

    # Vamos a la web del congreso
    driver.get('https://www.congreso.es/busqueda-de-intervenciones')
    # Aceptamos las dichosas cookies (Hasta en la web del congreso
venden tus datos)
    # El driver esperara a poder clickar el elemento seleccionado
(Aceptar cookies)
    wait.until(EC.element_to_be_clickable((By.XPATH,
'//*[@id="banda_cookies"]/a[1]'))).click()

    ## Rellenamos los campos del filtro
    # La legislatura actual está seleccionada por defecto, lo cual nos
viene perfecto.
    # Solo tenemos que introducir el orador, darle al botón buscar.
Esperar que cargue
    # y presionar el botón 'Por orden cronologico'
    # Despues vamos clicando en las distintas intervenciones, que sean
de distintos plenos, para descargar sus pdfs

    # Ubicamos y rellenamos el input del orador
    inputOrador = wait.until(EC.element_to_be_clickable((By.XPATH,
'//*[@id="_intervenciones_orador"]')))
    inputOrador.clear()
    inputOrador.send_keys(orador)
    # Click en 'Buscar'
    wait.until(EC.element_to_be_clickable((By.XPATH,
'/html/body/div[3]/div[2]/div[1]/div[2]/section/div/div[2]/div[2]/
```

```
div/div/section/div/div[2]/div/div[1]/div/div/div/div/div/div[2]/div/
div[3]/button[1]'))).click()
    # Esperamos 2 segundos a que cargue la información correctamente
    time.sleep(2)
    # Click en 'Por orden cronologico'
    wait.until(EC.element_to_be_clickable((By.XPATH,
'//*[@id="num"]'))).click()

    # Ya tenemos la lista de intervenciones cargada. Hacemos scroll
    # hasta el pie de pagina para asegurarnos que el driver carga todo
el html (Solo la primera vez)
    time.sleep(1)
    driver.execute_script("window.scrollTo(0,
document.body.scrollHeight);")
    time.sleep(1)
    driver.execute_script("window.scrollTo(0, 100);")

    while numPags > 0:
        # Ahora localizamos el elemento HTML que contiene
        # todos los elementos que no interesan
        contenedor = driver.find_element(By.XPATH,
'//*[@id="_intervenciones_contentPaginationIntervenciones"]' )

        # Todas las intervenciones tienen 'Pleno' en el que se dieron
y posteriormente
        # una tabla con más información sobre las intervenciónes y el
botón
        # para abrir el pdf del pleno.
        plenos = contenedor.find_elements(By.CLASS_NAME, 'pleno')
        tables = contenedor.find_elements(By.TAG_NAME, 'tbody')
        cajasIntervenciones = []

        # Se comprueba si el pleno en cuestión ha sido descargado
        # Si no esta descargado se captura y se añade a la lista
        # plenosCapturados
        for i,t in enumerate(tables):
            if plenos[i].text not in plenosCapturados:
                plenosCapturados.append(plenos[i].text)

cajasIntervenciones.append(t.find_elements(By.TAG_NAME, 'tr')[0])

        # Bucle que recorre los elementos donde están las
        # intervenciones y los clicka para descargar sus respectivos
pdfs
        for i in cajasIntervenciones:
            if (i.text != ''):

driver.execute_script("arguments[0].scrollIntoView();", i)
                time.sleep(2)
                elems = i.find_elements(By.TAG_NAME, 'i')
```

```
                elems[3].click()
            else:
                continue

        # Pasamos de pagina
        driver.execute_script("window.scrollTo(0,
document.body.scrollHeight);")
        driver.find_element(By.XPATH,
'//*[@id="_intervenciones_paginationLinksFooterIntervenciones"]/li[8]/
a').click()
        time.sleep(2)
        numPags-=1


    # Matamos el proceso de ejecución del driver (Cerramos el Chrome)
    driver.quit()

obtenerIntervenciones('Sánchez Pérez-Castejón, Pedro', 'C:\\Users\\
gonef\\Desktop\\Plenos', 5 )
obtenerIntervenciones('Abascal Conde, Santiago', 'C:\\Users\\gonef\\
Desktop\\Plenos', 5 )
obtenerIntervenciones('Casado Blanco, Pablo', 'C:\\Users\\gonef\\
Desktop\\Plenos', 5 )
len(plenosCapturados)
```

```
62
```

```
# Obtenemos un total de 62 pdfs de distintos plenos en la carpeta
'C:\\Users\\gonef\\Desktop\\Plenos'
```