Using Selenium to check Scopus journals are on WoS index or not?

In [ ]:
```
# using Selenium
# pip install selenium
```

In [ ]:
```
# import libraries
from selenium import webdriver

from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import Select  #thư viện cho Select box
from selenium.webdriver.common.keys import Keys  # để gán giá trị username, pass

from openpyxl import load_workbook   # để đọc file Excel

import json # thêm thư viện để đọc file JSON

from selenium.common.exceptions import NoSuchElementException
```

In [ ]:
```
# Đường dẫn đến Chrome WebDriver trên máy tính của bạn
chromedriver_path = r"C:\3.chromedriver-win32\chromedriver-win32\chromedriver.ex

# Khai báo đường dẫn cho Chrome WebDriver bằng cách thiết lập biến môi trường
webdriver.chrome.driver = chromedriver_path

# Khởi tạo trình duyệt Chrome
driver = webdriver.Chrome()

# Mở trang Web of Science
driver.get("https://mjl.clarivate.com/home")
```

In [ ]:
```
# Mở tạp chí mặc định
kwords = ""
inp = driver.find_element(By.ID, 'search-box')
inp.send_keys(kwords)
inp.submit() # xác nhận search
```

In [ ]:
```
# bắt đầu tìm
driver.find_element(By.ID, 'search-button').click() # click Search button
# Note: tắt pop-up Accept cookies thì sẽ KHÔNG bị lỗi
```

In [ ]:
```
# Tên file Excel của bạn
excel_file_path = "Scopus_index_Political Science and Relation.xlsx"
# Mở file Excel
workbook = load_workbook(excel_file_path)
sheet = workbook.active
# Lấy danh sách tên tạp chí từ cột tương ứng trong file Excel (thay thế 'B' bằng
journal_names = [cell.value for cell in sheet['B']]
```

In [ ]:
```
import time  # Import the time module

for row_index, journal_name in enumerate(journal_names, start=2):
    inp = driver.find_element(By.ID, 'search-box')

    # Clear the search box before entering a new journal name
```

```
    inp.clear()

    inp.send_keys(journal_name)
    inp.submit()  # xác nhận search
    driver.find_element(By.ID, 'search-button').click()  # click Search button



    # Tìm Index của tạp chí
    while True:
        try:
            title_elements = driver.find_elements(By.TAG_NAME, 'h2')  # tìm tất
            for title_element in title_elements:
                title = title_element.text
                if title == 'Exact Match Found':
                    # Duyệt qua các phần tử để tìm phần tử chứa văn bản "Web of
                    search_results_indexes = driver.find_elements(By.CLASS_NAME,
                    for index_title in search_results_indexes:
                        if "Web of Science Core Collection:" in index_title.text
                            core_collection_value = index_title.find_element(By.
                            indexc = core_collection_value.text
                            sheet[f'E{row_index}'] = indexc
                            break
                        else:
                            sheet[f'E{row_index}'] = 'No info'
                    break  # Exit the loop once "Exact Match Found" is encounter
        except StaleElementReferenceException:
            continue

        break  # Exit the while loop when the iteration is successful
    time.sleep(5)  # Add a delay of 5 seconds

# Note: This code uses a while loop to handle StaleElementReferenceException by
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
d:\Data science & Python  2022\2. Upwork jobs\job #5\Thach_SSCI_collecting_data_WoS_using_Selenium_.ipynb Cell 8 in 9
      <a href='vscode-notebook-cell:/d%3A/Data%20science%20%26%20Python%20%202022/2.%20Upwork%20jobs/job%20%235/Thach_SSCI_collecting_data_WoS_using_Selenium_.ipynb#X42sZmlsZQ%3D%3D?line=5'>6</a> # Clear the search box before entering a new journal name
      <a href='vscode-notebook-cell:/d%3A/Data%20science%20%26%20Python%20%202022/2.%20Upwork%20jobs/job%20%235/Thach_SSCI_collecting_data_WoS_using_Selenium_.ipynb#X42sZmlsZQ%3D%3D?line=6'>7</a> inp.clear()
----> <a href='vscode-notebook-cell:/d%3A/Data%20science%20%26%20Python%20%202022/2.%20Upwork%20jobs/job%20%235/Thach_SSCI_collecting_data_WoS_using_Selenium_.ipynb#X42sZmlsZQ%3D%3D?line=8'>9</a> inp.send_keys(journal_name)
     <a href='vscode-notebook-cell:/d%3A/Data%20science%20%26%20Python%20%202022/2.%20Upwork%20jobs/job%20%235/Thach_SSCI_collecting_data_WoS_using_Selenium_.ipynb#X42sZmlsZQ%3D%3D?line=9'>10</a> inp.submit()  # xác nhận search
     <a href='vscode-notebook-cell:/d%3A/Data%20science%20%26%20Python%20%202022/2.%20Upwork%20jobs/job%20%235/Thach_SSCI_collecting_data_WoS_using_Selenium_.ipynb#X42sZmlsZQ%3D%3D?line=10'>11</a> driver.find_element(By.ID, 'search-button').click()  # click Search button

File c:\Users\ADMIN\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\remote\webelement.py:232, in WebElement.send_keys(self, *value)
    228             remote_files.append(self._upload(file))
    229         value = "\n".join(remote_files)
    231 self._execute(
--> 232     Command.SEND_KEYS_TO_ELEMENT, {"text": "".join(keys_to_typing(value)), "value": keys_to_typing(value)}
    233 )

File c:\Users\ADMIN\AppData\Local\Programs\Python\Python310\lib\site-packages\selenium\webdriver\common\utils.py:138, in keys_to_typing(value)
    136         characters.extend(str(val))
    137     else:
--> 138         characters.extend(val)
    139 return characters

TypeError: 'NoneType' object is not iterable
```

In [ ]:

```
# # Lặp qua từng tên tạp chí để kiểm tra và cập nhật thông tin
# for row_index, journal_name in enumerate(journal_names, start=2):

#     inp = driver.find_element(By.ID, 'search-box')
#     inp.send_keys(journal_name)
#     inp.submit() # xác nhận search
#     driver.find_element(By.ID, 'search-button').click() # click Search button

#     # Tìm Index của tạp chí
#     title_elements = driver.find_elements(By.TAG_NAME, 'h2')  # tìm tất cả các
#     for title_element in title_elements:
#         title = title_element.text
#         if title == 'Exact Match Found':
#             # Duyệt qua các phần tử để tìm phần tử chứa văn bản "Web of Scienc
#             search_results_indexes = driver.find_elements(By.CLASS_NAME, "sear
#             for index_title in search_results_indexes:
#                 if "Web of Science Core Collection:" in index_title.text:
#                     core_collection_value = index_title.find_element(By.XPATH,
#                     indexc = core_collection_value.text
#                     #print(indexc)
```

```
#                     break
#                 sheet[f'E{row_index}'] = indexc
#                 else
#                 sheet[f'E{row_index}'] = 'No info'
```

```
  Cell In [13], line 23
    else
    ^
SyntaxError: invalid syntax
```

In [ ]:
```
# Lưu file Excel sau khi cập nhật
workbook.save(excel_file_path)
```

Phần sau bỏ qua

In [ ]:
```
# Đóng trình duyệt sau khi hoàn thành
driver.quit()
```

```
---------------------------------------------------------------------------
InvalidArgumentException                  Traceback (most recent call last)
d:\Data science & Python  2022\2. Upwork jobs\job #5\collecting_data_WoS_using_Se
lenium.ipynb Cell 5 in 1
     <a href='vscode-notebook-cell:/d%3A/Data%20science%20%26%20Python%20%202022/
2.%20Upwork%20jobs/job%20%235/collecting_data_WoS_using_Selenium.ipynb#W4sZmlsZQ%
3D%3D?line=15'>16</a> driver.implicitly_wait(10)  # Chờ 10 giây
     <a href='vscode-notebook-cell:/d%3A/Data%20science%20%26%20Python%20%202022/
2.%20Upwork%20jobs/job%20%235/collecting_data_WoS_using_Selenium.ipynb#W4sZmlsZQ%
3D%3D?line=17'>18</a> # Đọc thông tin từ trang web
---> <a href='vscode-notebook-cell:/d%3A/Data%20science%20%26%20Python%20%202022/
2.%20Upwork%20jobs/job%20%235/collecting_data_WoS_using_Selenium.ipynb#W4sZmlsZQ%
3D%3D?line=18'>19</a> journal_info = driver.find_element("class_name", "journal-i
nfo")
     <a href='vscode-notebook-cell:/d%3A/Data%20science%20%26%20Python%20%202022/
2.%20Upwork%20jobs/job%20%235/collecting_data_WoS_using_Selenium.ipynb#W4sZmlsZQ%
3D%3D?line=19'>20</a> info_text = journal_info.text
     <a href='vscode-notebook-cell:/d%3A/Data%20science%20%26%20Python%20%202022/
2.%20Upwork%20jobs/job%20%235/collecting_data_WoS_using_Selenium.ipynb#W4sZmlsZQ%
3D%3D?line=20'>21</a> print(info_text)

File c:\Users\ADMIN\AppData\Local\Programs\Python\Python310\lib\site-packages\sel
enium\webdriver\remote\webdriver.py:739, in WebDriver.find_element(self, by, valu
e)
    736     by = By.CSS_SELECTOR
    737     value = f'[name="{value}"]'
--> 739 return self.execute(Command.FIND_ELEMENT, {"using": by, "value": value})
["value"]

File c:\Users\ADMIN\AppData\Local\Programs\Python\Python310\lib\site-packages\sel
enium\webdriver\remote\webdriver.py:345, in WebDriver.execute(self, driver_comman
d, params)
    343 response = self.command_executor.execute(driver_command, params)
    344 if response:
--> 345     self.error_handler.check_response(response)
    346     response["value"] = self._unwrap_value(response.get("value", None))
    347     return response

File c:\Users\ADMIN\AppData\Local\Programs\Python\Python310\lib\site-packages\sel
enium\webdriver\remote\errorhandler.py:229, in ErrorHandler.check_response(self,
response)
    227         alert_text = value["alert"].get("text")
    228     raise exception_class(message, screen, stacktrace, alert_text)  # typ
e: ignore[call-arg]  # mypy is not smart enough here
--> 229 raise exception_class(message, screen, stacktrace)

InvalidArgumentException: Message: invalid argument: invalid locator
  (Session info: chrome=116.0.5845.111)
Stacktrace:
        GetHandleVerifier [0x00007FF6311152A2+57122]
        (No symbol) [0x00007FF63108EA92]
        (No symbol) [0x00007FF630F5E3AB]
        (No symbol) [0x00007FF630F97C02]
        (No symbol) [0x00007FF630F97E2C]
        (No symbol) [0x00007FF630FD0B67]
        (No symbol) [0x00007FF630FB701F]
        (No symbol) [0x00007FF630FCEB82]
        (No symbol) [0x00007FF630FB6DB3]
        (No symbol) [0x00007FF630F8D2B1]
        (No symbol) [0x00007FF630F8E494]
```

```
        GetHandleVerifier [0x00007FF6313BEF82+2849794]
        GetHandleVerifier [0x00007FF631411D24+3189156]
        GetHandleVerifier [0x00007FF63140ACAF+3160367]
        GetHandleVerifier [0x00007FF6311A6D06+653702]
        (No symbol) [0x00007FF63109A208]
        (No symbol) [0x00007FF6310962C4]
        (No symbol) [0x00007FF6310963F6]
        (No symbol) [0x00007FF6310867A3]
        BaseThreadInitThunk [0x00007FFEFAC97BD4+20]
        RtlUserThreadStart [0x00007FFEFAFCCED1+33]
```