Crawl Data Company Glassdoor

1. Import các thư viện:

------------------------------------------------------------------------------------

from operator import ne

from selenium import webdriver

from selenium.webdriver.common.by import By

from bs4 import BeautifulSoup

import time

import pandas as pd

from selenium.webdriver.support.ui import WebDriverWait

from selenium.webdriver.support import expected\_conditions as EC

from selenium.webdriver.common.action\_chains import ActionChains

import os

------------------------------------------------------------------------------------

- from selenium import webdriver: Import module webdriver từ thư viện Selenium để tương tác với trình duyệt.

- from selenium.webdriver.common.by import By: Import class By để sử dụng trong việc xác định vị trí các phần tử.

- from bs4 import BeautifulSoup: Import class BeautifulSoup từ thư viện BeautifulSoup để parse HTML.

- import pandas as pd: Import thư viện pandas với alias là pd để làm việc với dữ liệu dưới dạng DataFrame.

- import os: Import module os để thao tác với hệ điều hành.

- from selenium.webdriver.support.ui import WebDriverWait: Import class WebDriverWait để thực hiện các chờ đợi thông minh trong Selenium.

- from selenium.webdriver.support import expected\_conditions as EC: Import module expected\_conditions với alias là EC để sử dụng các điều kiện chờ đợi trong Selenium.

- from selenium.webdriver.common.action\_chains import ActionChains: Import class ActionChains để thực hiện các hành động chuỗi trong Selenium.

- import time: Import module time để sử dụng các hàm liên quan đến thời gian.

------------------------------------------------------------------------------------

2. Hàm extract\_str:

------------------------------------------------------------------------------------

def extract\_str(a):

try:

if 'K' in a:

a = a.replace('K', '')

a = float(a) \* 1000

elif '+ Employees' in a:

a = a.replace('+ Employees', '')

a = float(a)

except:

a = None

return a

------------------------------------------------------------------------------------

- Hàm này được sử dụng để trích xuất giá trị số từ chuỗi. Nếu chuỗi chứa 'K', nó sẽ thay thế 'K' bằng rỗng, sau đó chuyển đổi sang kiểu số và nhân với 1000. Nếu chuỗi chứa '+ Employees', nó sẽ thay thế '+ Employees' bằng rỗng và chuyển đổi sang kiểu số. Nếu có lỗi, nó trả về None.

------------------------------------------------------------------------------------

3.Hàm crawl\_data:

------------------------------------------------------------------------------------

def crawl\_data(driver):

data = []

html\_doc = driver.page\_source

soup = BeautifulSoup(html\_doc, 'html.parser')

allCompanysContainer = soup.find("div", {"class": "col-md-12 col-lg-8"})

if allCompanysContainer:

companys = allCompanysContainer.findAll("div", {"class": "mt-0 mb-std p-std css-mdw3bo css-errlgf"})

if companys:

for company in companys:

obj = {}

try:

#name

obj['name'] = company.find('h2', attrs={'data-test': 'employer-short-name'}).text

#rate

span\_tag = company.find('span', class\_='pr-xsm', attrs={'data-test': 'rating'})

obj['rate'] = span\_tag.find('b').text

#image

obj['image'] = company.find('img', attrs={'class': 'employerInfo\_\_EmployerInfoStyles\_\_logoStyle',

'data-test': 'employer-logo', 'src': True})["src"]

company\_size = company.find('span', class\_='d-block mt-0 css-56kyx5',

attrs={'data-test': 'employer-size'}).text

#glb\_company\_size

obj['glb\_company\_size'] = extract\_str(company\_size)

#industry

obj['industry'] = company.find('span', class\_='d-block mt-0 css-56kyx5',

attrs={'data-test': 'employer-industry'}).text

#review

review = company.find('div', class\_='d-flex flex-column align-items-center',

attrs={'data-test': 'cell-Reviews'})

review = review.find('h3', attrs={'data-test': 'cell-Reviews-count'}).text

obj['review'] = extract\_str(review)

#salary

salary = company.find('div', class\_='d-flex flex-column align-items-center',

attrs={'data-test': 'cell-Salaries'})

salary = salary.find('h3', attrs={'data-test': 'cell-Salaries-count'}).text

obj['salary'] = extract\_str(salary)

#jobs

jobs = company.find('div', class\_='d-flex flex-column align-items-center', attrs={'data-test': 'cell-Jobs'})

jobs = jobs.find('h3', attrs={'data-test': 'cell-Jobs-count'}).text

obj['jobs'] = extract\_str(jobs)

#description

obj['descripton'] = company.find('p', class\_='css-1sj9xzx css-56kyx5').text

except:

obj['name'] = None

obj['rate'] = None

obj['image'] = None

obj['glb\_company\_size'] = None

obj['industry'] = None

obj['review'] = None

obj['salary'] = None

obj['jobs'] = None

obj['descripton'] = None

data.append(obj)

df = pd.DataFrame(data)

file\_name = 'test.csv'

if not os.path.isfile(file\_name):

df.to\_csv(file\_name, mode='w', header=True, index=False, encoding='utf-8')

else:

df.to\_csv(file\_name, mode='a', header=False, index=False, encoding='utf-8')

else:

print('Khong tim thay san pham')

else:

print('Khong tim thay san pham')

------------------------------------------------------------------------------------

- Hàm này thực hiện quá trình thu thập dữ liệu từ trang web Glassdoor.

- Parse mã nguồn HTML của trang web bằng BeautifulSoup.

- Tìm phần tử chứa tất cả các công ty (allCompanysContainer).

- Duyệt qua từng công ty và trích xuất thông tin cần thiết như tên, đánh giá, ảnh logo, kích thước công ty, ngành nghề, số lượng đánh giá, mức lương, số lượng - - công việc, mô tả công ty.

- Gọi hàm extract\_str để chuyển đổi các giá trị số từ chuỗi.

- Lưu thông tin của từng công ty vào danh sách data.

- Tạo DataFrame từ danh sách data.

- Ghi dữ liệu vào file CSV (test.csv).

------------------------------------------------------------------------------------

4. Hàm main:

------------------------------------------------------------------------------------

def main():

driver = webdriver.Safari()

driver.set\_window\_size(1920, 1080)

url = "https://www.glassdoor.com/Reviews/index.htm?overall\_rating\_low=3.5&page=1&filterType=RATING\_OVERALL"

driver.get(url)

print('Chuẩn bị crawl')

print('Vui lòng không tắt tool hoặc trình duyệt')

print('===============================')

visited\_urls = set()

for i in range(1, 3):

try:

print('chuan bi crawl trang', i)

url = f"https://www.glassdoor.com/Reviews/index.htm?overall\_rating\_low=3.5&page={i}&filterType=RATING\_OVERALL"

if url in visited\_urls:

continue

visited\_urls.add(url)

driver.get(url)

crawl\_data(driver)

except Exception as e:

print(f"Lỗi khi trích xuất dữ liệu từ trang: {e}")

print('Đã xong trang:', i)

print('=============================')

driver.quit()

if \_\_name\_\_ == "\_\_main\_\_":

main()

------------------------------------------------------------------------------------

- Hàm này là hàm chính của chương trình.

- Khởi tạo trình duyệt và mở trang web Glassdoor.

- Duyệt qua các trang (1 đến 2) và gọi hàm crawl\_data để thu thập thông tin từ mỗi trang.

- Lưu thông báo và thông tin đánh giá vào file CSV (test.csv).

- Đóng trình duyệt sau khi hoàn thành.