In [ ]:

```
# 1. Tải và cài đặt Google Chrome (Phiên bản Stable)
!wget https://dl.google.com/linux/direct/google-chrome-stable_current_amd64.deb
!apt-get -y update
!apt-get install -y ./google-chrome-stable_current_amd64.deb

# 2. Cài đặt Selenium và Webdriver Manager (Tự động tìm driver phù hợp)
!pip install selenium webdriver-manager
```

```
--2025-12-17 15:44:20--  https://dl.google.com/linux/direct/google-chrome-stable_current_amd64.deb
Resolving dl.google.com (dl.google.com)... 142.251.10.136, 142.251.10.91, 142.251.10.93, ...
Connecting to dl.google.com (dl.google.com)|142.251.10.136|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 117816476 (112M) [application/x-debian-package]
Saving to: ‘google-chrome-stable_current_amd64.deb’

google-chrome-stabl 100%[===================>] 112.36M   158MB/s    in 0.7s

2025-12-17 15:44:21 (158 MB/s) - ‘google-chrome-stable_current_amd64.deb’ saved [117816476/117816476]

Hit:1 https://cli.github.com/packages stable InRelease
Hit:2 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:3 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Get:4 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease [3,632 B]
Get:5 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [128 kB]
Hit:6 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Get:7 https://r2u.stat.illinois.edu/ubuntu jammy InRelease [6,555 B]
Hit:8 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:9 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [127 kB]
Get:10 https://r2u.stat.illinois.edu/ubuntu jammy/main all Packages [9,550 kB]
Get:11 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1,598 kB]
Get:12 http://security.ubuntu.com/ubuntu jammy-security/restricted amd64 Packages [6,205 kB]
Get:13 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [3,965 kB]
Get:14 http://archive.ubuntu.com/ubuntu jammy-updates/restricted amd64 Packages [6,411 kB]
Get:15 http://archive.ubuntu.com/ubuntu jammy-updates/multiverse amd64 Packages [69.3 kB]
Get:16 http://archive.ubuntu.com/ubuntu jammy-backports/main amd64 Packages [114 kB]
Get:17 http://archive.ubuntu.com/ubuntu jammy-backports/universe amd64 Packages [40.3 kB]
Get:18 https://r2u.stat.illinois.edu/ubuntu jammy/main amd64 Packages [2,851 kB]
Get:19 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages [3,633 kB]
Fetched 34.8 MB in 8s (4,174 kB/s)
Reading package lists... Done
W: Skipping acquire of configured file 'main/source/Sources' as repository 'https://r2u.stat.illinois.e
du/ubuntu jammy InRelease' does not seem to provide it (sources.list entry misspelt?)
```

```
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
Note, selecting 'google-chrome-stable' instead of './google-chrome-stable_current_amd64.deb'
The following additional packages will be installed:
  at-spi2-core gsettings-desktop-schemas libatk-bridge2.0-0 libatk1.0-0
  libatk1.0-data libatspi2.0-0 libvulkan1 libxcomposite1 libxtst6
  mesa-vulkan-drivers session-migration
The following NEW packages will be installed:
  at-spi2-core google-chrome-stable gsettings-desktop-schemas
  libatk-bridge2.0-0 libatk1.0-0 libatk1.0-data libatspi2.0-0 libvulkan1
  libxcomposite1 libxtst6 mesa-vulkan-drivers session-migration
0 upgraded, 12 newly installed, 0 to remove and 3 not upgraded.
Need to get 11.2 MB/129 MB of archives.
After this operation, 444 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu jammy/main amd64 libatk1.0-data all 2.36.0-3build1 [2,824 B]
Get:2 http://archive.ubuntu.com/ubuntu jammy/main amd64 libatk1.0-0 amd64 2.36.0-3build1 [51.9 kB]
Get:3 http://archive.ubuntu.com/ubuntu jammy/main amd64 libatspi2.0-0 amd64 2.44.0-3 [80.9 kB]
Get:4 http://archive.ubuntu.com/ubuntu jammy/main amd64 libatk-bridge2.0-0 amd64 2.38.0-3 [66.6 kB]
Get:5 http://archive.ubuntu.com/ubuntu jammy/main amd64 libvulkan1 amd64 1.3.204.1-2 [128 kB]
Get:6 http://archive.ubuntu.com/ubuntu jammy/main amd64 libxcomposite1 amd64 1:0.4.5-1build2 [7,192 B]
Get:7 http://archive.ubuntu.com/ubuntu jammy/main amd64 libxtst6 amd64 2:1.2.3-1build4 [13.4 kB]
Get:8 http://archive.ubuntu.com/ubuntu jammy/main amd64 session-migration amd64 0.3.6 [9,774 B]
Get:9 http://archive.ubuntu.com/ubuntu jammy/main amd64 gsettings-desktop-schemas all 42.0-1ubuntu1 [3
1.1 kB]
Get:10 http://archive.ubuntu.com/ubuntu jammy/main amd64 at-spi2-core amd64 2.44.0-3 [54.4 kB]
Get:11 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 mesa-vulkan-drivers amd64 23.2.1-1ubun
tu3.1~22.04.3 [10.7 MB]
Get:12 /content/google-chrome-stable_current_amd64.deb google-chrome-stable amd64 143.0.7499.146-1 [118
MB]
Fetched 11.2 MB in 2s (5,186 kB/s)
Selecting previously unselected package libatk1.0-data.
(Reading database ... 117528 files and directories currently installed.)
Preparing to unpack .../00-libatk1.0-data_2.36.0-3build1_all.deb ...
```

```
Unpacking libatk1.0-data (2.36.0-3build1) ...
Selecting previously unselected package libatk1.0-0:amd64.
Preparing to unpack .../01-libatk1.0-0_2.36.0-3build1_amd64.deb ...
Unpacking libatk1.0-0:amd64 (2.36.0-3build1) ...
Selecting previously unselected package libatspi2.0-0:amd64.
Preparing to unpack .../02-libatspi2.0-0_2.44.0-3_amd64.deb ...
Unpacking libatspi2.0-0:amd64 (2.44.0-3) ...
Selecting previously unselected package libatk-bridge2.0-0:amd64.
Preparing to unpack .../03-libatk-bridge2.0-0_2.38.0-3_amd64.deb ...
Unpacking libatk-bridge2.0-0:amd64 (2.38.0-3) ...
Selecting previously unselected package libvulkan1:amd64.
Preparing to unpack .../04-libvulkan1_1.3.204.1-2_amd64.deb ...
Unpacking libvulkan1:amd64 (1.3.204.1-2) ...
Selecting previously unselected package libxcomposite1:amd64.
Preparing to unpack .../05-libxcomposite1_1%3a0.4.5-1build2_amd64.deb ...
Unpacking libxcomposite1:amd64 (1:0.4.5-1build2) ...
Selecting previously unselected package google-chrome-stable.
Preparing to unpack .../06-google-chrome-stable_current_amd64.deb ...
Unpacking google-chrome-stable (143.0.7499.146-1) ...
Selecting previously unselected package libxtst6:amd64.
Preparing to unpack .../07-libxtst6_2%3a1.2.3-1build4_amd64.deb ...
Unpacking libxtst6:amd64 (2:1.2.3-1build4) ...
Selecting previously unselected package session-migration.
Preparing to unpack .../08-session-migration_0.3.6_amd64.deb ...
Unpacking session-migration (0.3.6) ...
Selecting previously unselected package gsettings-desktop-schemas.
Preparing to unpack .../09-gsettings-desktop-schemas_42.0-1ubuntu1_all.deb ...
Unpacking gsettings-desktop-schemas (42.0-1ubuntu1) ...
Selecting previously unselected package at-spi2-core.
Preparing to unpack .../10-at-spi2-core_2.44.0-3_amd64.deb ...
Unpacking at-spi2-core (2.44.0-3) ...
Selecting previously unselected package mesa-vulkan-drivers:amd64.
Preparing to unpack .../11-mesa-vulkan-drivers_23.2.1-1ubuntu3.1~22.04.3_amd64.deb ...
Unpacking mesa-vulkan-drivers:amd64 (23.2.1-1ubuntu3.1~22.04.3) ...
```

```
Setting up session-migration (0.3.6) ...
Created symlink /etc/systemd/user/graphical-session-pre.target.wants/session-migration.service → /usr/l
ib/systemd/user/session-migration.service.
Setting up libxtst6:amd64 (2:1.2.3-1build4) ...
Setting up libatspi2.0-0:amd64 (2.44.0-3) ...
Setting up libvulkan1:amd64 (1.3.204.1-2) ...
Setting up libatk1.0-data (2.36.0-3build1) ...
Setting up libatk1.0-0:amd64 (2.36.0-3build1) ...
Setting up libxcomposite1:amd64 (1:0.4.5-1build2) ...
Setting up gsettings-desktop-schemas (42.0-1ubuntu1) ...
Setting up mesa-vulkan-drivers:amd64 (23.2.1-1ubuntu3.1~22.04.3) ...
Setting up libatk-bridge2.0-0:amd64 (2.38.0-3) ...
Setting up google-chrome-stable (143.0.7499.146-1) ...
update-alternatives: using /usr/bin/google-chrome-stable to provide /usr/bin/x-www-browser (x-www-brows
er) in auto mode
update-alternatives: using /usr/bin/google-chrome-stable to provide /usr/bin/gnome-www-browser (gnome-w
ww-browser) in auto mode
update-alternatives: using /usr/bin/google-chrome-stable to provide /usr/bin/google-chrome (google-chro
me) in auto mode
Processing triggers for mailcap (3.70+nmu1ubuntu1) ...
Processing triggers for libglib2.0-0:amd64 (2.72.4-0ubuntu2.6) ...
Processing triggers for libc-bin (2.35-0ubuntu3.11) ...
/sbin/ldconfig.real: /usr/local/lib/libur_loader.so.0 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libur_adapter_level_zero.so.0 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtcm_debug.so.1 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_0.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libur_adapter_opencl.so.0 is not a symbolic link
```

```
/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_5.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc.so.2 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc_proxy.so.2 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libur_adapter_level_zero_v2.so.0 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libumf.so.1 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbb.so.12 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtcm.so.1 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libhwloc.so.15 is not a symbolic link

Processing triggers for man-db (2.10.2-1) ...
Setting up at-spi2-core (2.44.0-3) ...
Collecting selenium
  Downloading selenium-4.39.0-py3-none-any.whl.metadata (7.5 kB)
Collecting webdriver-manager
  Downloading webdriver_manager-4.0.2-py2.py3-none-any.whl.metadata (12 kB)
Requirement already satisfied: urllib3<3.0,>=2.5.0 in /usr/local/lib/python3.12/dist-packages (from url
lib3[socks]<3.0,>=2.5.0->selenium) (2.5.0)
Collecting trio<1.0,>=0.31.0 (from selenium)
  Downloading trio-0.32.0-py3-none-any.whl.metadata (8.5 kB)
Collecting trio-websocket<1.0,>=0.12.2 (from selenium)
  Downloading trio_websocket-0.12.2-py3-none-any.whl.metadata (5.1 kB)
Requirement already satisfied: certifi>=2025.10.5 in /usr/local/lib/python3.12/dist-packages (from sele
nium) (2025.11.12)
Requirement already satisfied: typing_extensions<5.0,>=4.15.0 in /usr/local/lib/python3.12/dist-package
s (from selenium) (4.15.0)
Requirement already satisfied: websocket-client<2.0,>=1.8.0 in /usr/local/lib/python3.12/dist-packages
(from selenium) (1.9.0)
```

```
Requirement already satisfied: requests in /usr/local/lib/python3.12/dist-packages (from webdriver-mana
ger) (2.32.4)
Requirement already satisfied: python-dotenv in /usr/local/lib/python3.12/dist-packages (from webdriver
-manager) (1.2.1)
Requirement already satisfied: packaging in /usr/local/lib/python3.12/dist-packages (from webdriver-man
ager) (25.0)
Requirement already satisfied: attrs>=23.2.0 in /usr/local/lib/python3.12/dist-packages (from trio<1.0,
>=0.31.0->selenium) (25.4.0)
Collecting sortedcontainers (from trio<1.0,>=0.31.0->selenium)
  Downloading sortedcontainers-2.4.0-py2.py3-none-any.whl.metadata (10 kB)
Requirement already satisfied: idna in /usr/local/lib/python3.12/dist-packages (from trio<1.0,>=0.31.0-
>selenium) (3.11)
Collecting outcome (from trio<1.0,>=0.31.0->selenium)
  Downloading outcome-1.3.0.post0-py2.py3-none-any.whl.metadata (2.6 kB)
Requirement already satisfied: sniffio>=1.3.0 in /usr/local/lib/python3.12/dist-packages (from trio<1.
0,>=0.31.0->selenium) (1.3.1)
Collecting wsproto>=0.14 (from trio-websocket<1.0,>=0.12.2->selenium)
  Downloading wsproto-1.3.2-py3-none-any.whl.metadata (5.2 kB)
Requirement already satisfied: pysocks!=1.5.7,<2.0,>=1.5.6 in /usr/local/lib/python3.12/dist-packages
(from urllib3[socks]<3.0,>=2.5.0->selenium) (1.7.1)
Requirement already satisfied: charset_normalizer<4,>=2 in /usr/local/lib/python3.12/dist-packages (fro
m requests->webdriver-manager) (3.4.4)
Requirement already satisfied: h11<1,>=0.16.0 in /usr/local/lib/python3.12/dist-packages (from wsproto>
=0.14->trio-websocket<1.0,>=0.12.2->selenium) (0.16.0)
Downloading selenium-4.39.0-py3-none-any.whl (9.7 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 9.7/9.7 MB 94.6 MB/s eta 0:00:00
Downloading webdriver_manager-4.0.2-py2.py3-none-any.whl (27 kB)
Downloading trio-0.32.0-py3-none-any.whl (512 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 512.0/512.0 kB 31.1 MB/s eta 0:00:00
Downloading trio_websocket-0.12.2-py3-none-any.whl (21 kB)
Downloading outcome-1.3.0.post0-py2.py3-none-any.whl (10 kB)
Downloading wsproto-1.3.2-py3-none-any.whl (24 kB)
Downloading sortedcontainers-2.4.0-py2.py3-none-any.whl (29 kB)
Installing collected packages: sortedcontainers, wsproto, outcome, webdriver-manager, trio, trio-websoc
```

```
ket, selenium
Successfully installed outcome-1.3.0.post0 selenium-4.39.0 sortedcontainers-2.4.0 trio-0.32.0 trio-webs
ocket-0.12.2 webdriver-manager-4.0.2 wsproto-1.3.2
```

In [ ]:

```python
import time
import pandas as pd
from selenium import webdriver
from selenium.webdriver.chrome.service import Service as ChromeService
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.common.by import By

# --- 1. CẤU HÌNH TRÌNH DUYỆT ---
options = webdriver.ChromeOptions()
options.add_argument('--headless')
options.add_argument('--no-sandbox')
options.add_argument('--disable-dev-shm-usage')

print("🚀 Đang khởi động crawler cho dự án cuối kỳ...")
try:
    service = ChromeService(ChromeDriverManager().install())
    driver = webdriver.Chrome(service=service, options=options)
    print("✅ Trình duyệt sẵn sàng!")
except Exception as e:
    print(f"❌ Lỗi driver: {e}")

# --- 2. THIẾT LẬP THAM SỐ (PHẦN QUAN TRỌNG) ---
# Link chuyên mục: CNTT - Phần mềm (c1)
# Cấu trúc link: https://careerviet.vn/viec-lam/cntt-phan-mem-c1-trang-{}-vi.html
base_url = "https://careerviet.vn/viec-lam/cntt-phan-mem-c1-trang-{}-vi.html"

# SỐ TRANG MUỐN LẤY: Để làm đồ án, bạn nên lấy khoảng 30-50 trang
so_trang_muon_lay = 50

data_list = []
```

```python
print(f"🎯 Mục tiêu: Crawl {so_trang_muon_lay} trang thuộc nhóm ngành CNTT...")

# --- 3. BẮT ĐẦU CRAWL ---
for page in range(1, so_trang_muon_lay + 1):
    # Tạo link cho từng trang
    if page == 1:
        url = "https://careerviet.vn/viec-lam/cntt-phan-mem-c1-vi.html"
    else:
        url = base_url.format(page)

    print(f"   ⏳ Đang xử lý trang {page}/{so_trang_muon_lay}...", end="\r") # In đè dòng để gọn

    try:
        driver.get(url)
        time.sleep(2) # Nghỉ 2s mỗi trang để tránh bị chặn

        jobs = driver.find_elements(By.CSS_SELECTOR, ".job-item")

        for job in jobs:
            try:
                # Lấy Title
                try: title = job.find_element(By.CSS_SELECTOR, ".title a").text
                except: title = ""

                # Lấy Company
                try: company = job.find_element(By.CSS_SELECTOR, ".company-name").text
                except: company = ""

                # Lấy Salary
                try: salary = job.find_element(By.CSS_SELECTOR, ".salary p").text
                except: salary = "Thỏa thuận"

                # Lấy Location
                try: location = job.find_element(By.CSS_SELECTOR, ".location").text
```

```
                except: location = "Vietnam"

                if title:
                    data_list.append({
                        "Job Title": title,
                        "Company": company,
                        "Salary": salary,
                        "Location": location
                    })
            except:
                continue

    except Exception as e:
        print(f"\n❌ Lỗi tại trang {page}: {e}")

driver.quit()

# --- 4. LƯU KẾT QUẢ ---
print(f"\n\n✅ HOÀN THÀNH! Tổng số job thu thập được: {len(data_list)}")

if len(data_list) > 0:
    df = pd.DataFrame(data_list)

    # Loại bỏ các dòng trùng lặp (nếu có)
    df.drop_duplicates(inplace=True)
    print(f"📉 Sau khi loại bỏ trùng lặp còn: {len(df)} jobs")

    # Lưu file
    file_name = "jobs_it_final.csv"
    df.to_csv(file_name, index=False, encoding="utf-8-sig")
    print(f"📁 Đã lưu file: {file_name}")

    # Hiển thị thống kê nhanh
    print("\n--- Thống kê sơ bộ ---")
```

```
    print(df['Location'].value_counts().head())
else:
    print("❌ Không lấy được dữ liệu.")
```

```
🚀 Đang khởi động crawler cho dự án cuối kỳ...
✅ Trình duyệt sẵn sàng!
🎯 Mục tiêu: Crawl 50 trang thuộc nhóm ngành CNTT...


✅ HOÀN THÀNH! Tổng số job thu thập được: 1314
📉 Sau khi loại bỏ trùng lặp còn: 994 jobs
📁 Đã lưu file: jobs_it_final.csv

--- Thống kê sơ bộ ---
Location
Hà Nội                  522
Hồ Chí Minh             359
Bình Dương               27
Hà Nội\nHồ Chí Minh      12
Đồng Nai                 10
Name: count, dtype: int64
```

In [ ]:

```
import time
import pandas as pd
from selenium import webdriver
from selenium.webdriver.chrome.service import Service as ChromeService
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.common.by import By

# --- 1. CẤU HÌNH TRÌNH DUYỆT ---
options = webdriver.ChromeOptions()
options.add_argument('--headless')
options.add_argument('--no-sandbox')
options.add_argument('--disable-dev-shm-usage')
```

```
print("🚀 Đang khởi động crawler cho dự án cuối kỳ...")
try:
    service = ChromeService(ChromeDriverManager().install())
    driver = webdriver.Chrome(service=service, options=options)
    print("✅ Trình duyệt sẵn sàng!")
except Exception as e:
    print(f"❌ Lỗi driver: {e}")

# --- 2. THIẾT LẬP THAM SỐ (PHẦN QUAN TRỌNG) ---
# Link chuyên mục: CNTT - Phần mềm (c1)
# Cấu trúc link: https://careerviet.vn/viec-lam/cntt-phan-mem-c1-trang-{}-vi.html
base_url = "https://careerviet.vn/viec-lam/cntt-phan-mem-c1-trang-{}-vi.html"

# SỐ TRANG MUỐN LẤY: Để làm đồ án, bạn nên lấy khoảng 30-50 trang
so_trang_muon_lay = 200

data_list = []
print(f"🎯 Mục tiêu: Crawl {so_trang_muon_lay} trang thuộc nhóm ngành CNTT...")

# --- 3. BẮT ĐẦU CRAWL ---
for page in range(1, so_trang_muon_lay + 1):
    # Tạo link cho từng trang
    if page == 1:
        url = "https://careerviet.vn/viec-lam/cntt-phan-mem-c1-vi.html"
    else:
        url = base_url.format(page)

    print(f"   ⏳ Đang xử lý trang {page}/{so_trang_muon_lay}...", end="\r") # In đè dòng để gọn

    try:
        driver.get(url)
        time.sleep(2) # Nghỉ 2s mỗi trang để tránh bị chặn
```

```
        jobs = driver.find_elements(By.CSS_SELECTOR, ".job-item")

        for job in jobs:
            try:
                # Lấy Title
                try: title = job.find_element(By.CSS_SELECTOR, ".title a").text
                except: title = ""

                # Lấy Company
                try: company = job.find_element(By.CSS_SELECTOR, ".company-name").text
                except: company = ""

                # Lấy Salary
                try: salary = job.find_element(By.CSS_SELECTOR, ".salary p").text
                except: salary = "Thỏa thuận"

                # Lấy Location
                try: location = job.find_element(By.CSS_SELECTOR, ".location").text
                except: location = "Vietnam"

                if title:
                    data_list.append({
                        "Job Title": title,
                        "Company": company,
                        "Salary": salary,
                        "Location": location
                    })
            except:
                continue

    except Exception as e:
        print(f"\n❌ Lỗi tại trang {page}: {e}")

driver.quit()
```

```python
# --- 4. LƯU KẾT QUẢ ---
print(f"\n\n✅ HOÀN THÀNH! Tổng số job thu thập được: {len(data_list)}")

if len(data_list) > 0:
    df = pd.DataFrame(data_list)

    # Loại bỏ các dòng trùng lặp (nếu có)
    df.drop_duplicates(inplace=True)
    print(f"📉 Sau khi loại bỏ trùng lặp còn: {len(df)} jobs")

    # Lưu file
    file_name = "jobs_it_final.csv"
    df.to_csv(file_name, index=False, encoding="utf-8-sig")
    print(f"📁 Đã lưu file: {file_name}")

    # Hiển thị thống kê nhanh
    print("\n--- Thống kê sơ bộ ---")
    print(df['Location'].value_counts().head())
else:
    print("❌ Không lấy được dữ liệu.")
```

```
🚀 Đang khởi động crawler cho dự án cuối kỳ...
✅ Trình duyệt sẵn sàng!
🎯 Mục tiêu: Crawl 200 trang thuộc nhóm ngành CNTT...


✅ HOÀN THÀNH! Tổng số job thu thập được: 1730
📉 Sau khi loại bỏ trùng lặp còn: 10 jobs
📁 Đã lưu file: jobs_it_final.csv

--- Thống kê sơ bộ ---
Location
Hà Nội                  6
Hà Tĩnh                 2
Quảng Nam | Đà Nẵng     1
Long An                 1
Name: count, dtype: int64
```

crawl lại

In [ ]:

```python
import time
import pandas as pd
from selenium import webdriver
from selenium.webdriver.chrome.service import Service as ChromeService
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.common.by import By

# --- 1. CẤU HÌNH TRÌNH DUYỆT ---
options = webdriver.ChromeOptions()
options.add_argument('--headless')
options.add_argument('--no-sandbox')
options.add_argument('--disable-dev-shm-usage')
```

```
print("🚀 Đang khởi động crawler cho danh mục CNTT (Phần cứng & Phần mềm)...")
try:
    service = ChromeService(ChromeDriverManager().install())
    driver = webdriver.Chrome(service=service, options=options)
    print("✅ Trình duyệt sẵn sàng!")
except Exception as e:
    print(f"❌ Lỗi driver: {e}")
    exit()

# --- 2. THIẾT LẬP THAM SỐ ---
base_url_page_1 = "https://careerviet.vn/viec-lam/cntt-phan-cung-mang-cntt-phan-mem-c63,1-vi.html"
base_url_other = "https://careerviet.vn/viec-lam/cntt-phan-cung-mang-cntt-phan-mem-c63,1-trang-{}-vi.h

so_trang_muon_lay = 23
data_list = []
last_successful_page = 0

print(f"🎯 Mục tiêu: Crawl tối đa {so_trang_muon_lay} trang...")

# --- 3. BẮT ĐẦU CRAWL ---
for page in range(1, so_trang_muon_lay + 1):
    if page == 1:
        url = base_url_page_1
    else:
        url = base_url_other.format(page)

    print(f"📄 Đang crawl trang {page}...")

    try:
        driver.get(url)
        time.sleep(2)

        jobs = driver.find_elements(By.CSS_SELECTOR, ".job-item")
```

```
        if not jobs:
            print(f"    ⚠️ Trang {page}: Không tìm thấy việc làm. DỪNG CRAWL tại đây.")
            break

        job_count = 0
        for job in jobs:
            try:
                title_elem = job.find_elements(By.CSS_SELECTOR, ".title a")
                title = title_elem[0].text if title_elem else ""
                if not title.strip():
                    continue

                company_elem = job.find_elements(By.CSS_SELECTOR, ".company-name")
                company = company_elem[0].text if company_elem else ""

                salary_elem = job.find_elements(By.CSS_SELECTOR, ".salary p")
                salary = salary_elem[0].text if salary_elem else "Thỏa thuận"

                location_elem = job.find_elements(By.CSS_SELECTOR, ".location")
                location = location_elem[0].text if location_elem else "Vietnam"

                data_list.append({
                    "Job Title": title,
                    "Company": company,
                    "Salary": salary,
                    "Location": location
                })
                job_count += 1

            except Exception:
                continue

        print(f"    ✅ Trang {page}: Lấy được {job_count} job(s).")
        last_successful_page = page  # Ghi lại trang cuối thành công
```

```
    except Exception as e:
        print(f"   ❌ Lỗi khi crawl trang {page}: {e}")
        break

# --- 4. KẾT THÚC ---
driver.quit()

print(f"\n✅ HOÀN THÀNH!")
print(f"🔗 Đã crawl thành công đến trang: {last_successful_page} / {so_trang_muon_lay}")
print(f"📥 Tổng số job thu thập được: {len(data_list)}")

if data_list:
    df = pd.DataFrame(data_list)
    df.drop_duplicates(subset=["Job Title", "Company", "Location"], inplace=True)
    print(f"🧹 Sau khi loại bỏ trùng lặp: {len(df)} job(s)")

    file_name = "jobs_it_hardware_software.csv"
    df.to_csv(file_name, index=False, encoding="utf-8-sig")
    print(f"📁 Đã lưu file: {file_name}")

    print("\n--- Top 5 địa điểm ---")
    print(df['Location'].value_counts().head())
else:
    print("❌ Không có dữ liệu.")
```

```
🚀 Đang khởi động crawler cho danh mục CNTT (Phần cứng & Phần mềm)...
✅ Trình duyệt sẵn sàng!
🎯 Mục tiêu: Crawl tối đa 23 trang...
📄 Đang crawl trang 1...
   ✅ Trang 1: Lấy được 50 job(s).
📄 Đang crawl trang 2...
   ✅ Trang 2: Lấy được 50 job(s).
📄 Đang crawl trang 3...
   ✅ Trang 3: Lấy được 50 job(s).
📄 Đang crawl trang 4...
   ✅ Trang 4: Lấy được 50 job(s).
📄 Đang crawl trang 5...
   ✅ Trang 5: Lấy được 50 job(s).
📄 Đang crawl trang 6...
   ✅ Trang 6: Lấy được 50 job(s).
📄 Đang crawl trang 7...
   ✅ Trang 7: Lấy được 50 job(s).
📄 Đang crawl trang 8...
   ✅ Trang 8: Lấy được 50 job(s).
📄 Đang crawl trang 9...
   ✅ Trang 9: Lấy được 50 job(s).
📄 Đang crawl trang 10...
   ✅ Trang 10: Lấy được 50 job(s).
📄 Đang crawl trang 11...
   ✅ Trang 11: Lấy được 50 job(s).
📄 Đang crawl trang 12...
   ✅ Trang 12: Lấy được 50 job(s).
📄 Đang crawl trang 13...
   ✅ Trang 13: Lấy được 50 job(s).
📄 Đang crawl trang 14...
   ✅ Trang 14: Lấy được 50 job(s).
📄 Đang crawl trang 15...
   ✅ Trang 15: Lấy được 20 job(s).
📄 Đang crawl trang 16...
```

```
   ✅ Trang 16: Lấy được 50 job(s).
📄 Đang crawl trang 17...
   ✅ Trang 17: Lấy được 50 job(s).
📄 Đang crawl trang 18...
   ✅ Trang 18: Lấy được 50 job(s).
📄 Đang crawl trang 19...
   ✅ Trang 19: Lấy được 50 job(s).
📄 Đang crawl trang 20...
   ✅ Trang 20: Lấy được 50 job(s).
📄 Đang crawl trang 21...
   ✅ Trang 21: Lấy được 50 job(s).
📄 Đang crawl trang 22...
   ✅ Trang 22: Lấy được 50 job(s).
📄 Đang crawl trang 23...
   ✅ Trang 23: Lấy được 35 job(s).

✅ HOÀN THÀNH!
🔗 Đã crawl thành công đến trang: 23 / 23
📥 Tổng số job thu thập được: 1105
🧹 Sau khi loại bỏ trùng lặp: 1045 job(s)
📁 Đã lưu file: jobs_it_hardware_software.csv

--- Top 5 địa điểm ---
Location
Hà Nội                547
Hồ Chí Minh           371
Bình Dương             27
Hà Nội\nHồ Chí Minh    13
Đồng Nai                9
Name: count, dtype: int64
```