IT4930: Nhập môn Khoa học dữ liệu

BÁO CÁO BÀI THỰC HÀNH CRAWL DỮ LIỆU

Họ tên sinh viên: Trần Gia Định

MSSV: 20235036

Bài tập sử dụng thư viện selenium và beautifulsoup để crawl dữ liệu

BƯỚC 1. Import các thư viện cần dùng

In [1]:
```
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.chrome.service import Service
from bs4 import BeautifulSoup
import pandas as pd
import time
import chromedriver_autoinstaller
import random
```

BƯỚC 2. Khởi tạo driver

In [2]:
```
# Lấy đường dẫn driver phù hợp với phiên bản chrome hiện tại
driver_path = chromedriver_autoinstaller.install()
# Thiết lập các tùy chọn cho driver
option = Options()
# Khi lấy dữ liệu thì ẩn việc chrome được lấy dữ liệu sử dụng selenium
option.add_argument("--disable-blink-features=AutomationControlled")
# Ẩn cửa sổ Chrome khi chạy
option.add_argument("--headless")
# Giả mạo user-agent gửi tới server
option.add_argument("user-agent=Mozilla/5.0 (Windows NT 10.0; Win64; x64)"
                    "AppleWebKit/537.36 (KHTML, like Gecko)"
                    "Chrome/124.0.0.0 Safari/537.36")
service = Service(driver_path)
```

```
# Sử dụng driver với các tham số khởi tạo
driver = webdriver.Chrome(service=service, options=option)
# Thời gian đợi trang web phản hồi trước khi báo lỗi
driver.set_page_load_timeout(60)

# Url cần crawl
url = 'https://cafef.vn/tai-chinh-quoc-te.chn'
```

BƯỚC 3. Sử dụng beautifulsoup để phân tích mã HTML của trang web

In [3]:
```
driver.get(url)
# Tạo độ ngẫu nhiên giữa các lần truy cập (để không làm quá tải trang web cần crawl
time.sleep(random.uniform(1, 3))

# Lấy mã HTML của toàn bộ web
html = driver.page_source
# Khởi tạo đối tượng beautifulsoup để phân tích thông tin mã HTML
soup = BeautifulSoup(html, 'html.parser')
# Lấy ra các mục bài báo
contents = soup.find_all("div", class_="tlitem box-category-item")

# Tạo list để chứa dữ liệu của từng bài báo
data = []
# Đếm số bài báo được crawl (giới hạn 20 bài báo)
cnt = 0

for content in contents:
    if cnt == 20:
        break
    # Tiêu đề bài báo
    title = content.h3.a.get_text(strip=True)
    # Đường dẫn tới bài báo
    base_link = 'https://cafef.vn'
    link = content.h3.a["href"]
    link = base_link + link
    cnt += 1
    # Lưu dưới dạng dictionary
    data.append({
        "Index": cnt,
        "Title": title,
        "Link": link
    })
```

BƯỚC 4. Lưu dữ liệu crawl vào file excel sử dụng thư viện pandas

In [4]:
```
df = pd.DataFrame(data)
df.to_excel("data/crawl_data_practice.xlsx", index=False)
```

BƯỚC 5. Tắt chrome và tắt phiên làm việc với webdriver

In [5]:
```
driver.quit()
```

KẾT LUẬN

Sau khi sử dụng ChromeDrive cùng các thư viện khác như Selenium và BeautifulSoup để crawl dữ liệu (gửi yêu cầu đến trang web, trích xuất mã HTML), ta có được thông tin về tiêu đề (title) và links của 20 bài báo từ trang web 'https://cafef.vn/tai-chinh-quoc-te.chn'