TÀI LIỆU SỬ DỤNG CRAWL TOOLS

1. Tạo image Docker

Đảm bảo thiết bị đã cài đặt Docker nhập lệnh `docker -version` để kiểm tra Docker đã được cài đặt chưa

```
C:\Users\AN>docker --version
Docker version 28.3.2, build 578ccf6
```

Di chuyển vào thư mục chứa file Dockerfile và nhập lệnh sau để tạo docker image

```
docker build . --tag selenium_crawl:extend
```

```
C:\Users\AN\Collection\Work\Giang\GiangJob>docker build . --tag selenium_crawl:extend
[+] Building 10.6s (9/9) FINISHED
 => [internal] load build definition from Dockerfile
 => => transferring dockerfile: 233B
 => [internal] load metadata for docker.io/selenium/standalone-chrome:latest
 => [internal] load .dockerignore
 => => transferring context: 2B
 => [internal] load build context
 => => transferring context: 70B
 => [1/4] FROM docker.io/selenium/standalone-chrome:latest@sha256:bcb1d054f3c88ef618cd92a8124d894038fb5f670bf7d427f5e5bbb60882b6f6
 => => resolve docker.io/selenium/standalone-chrome:latest@sha256:bcb1d054f3c88ef618cd92a8124d894038fb5f670bf7d427f5e5bbb60882b6f6
 => [2/4] COPY requirements.txt /opt/project/requirements.txt
 => [3/4] RUN pip install --upgrade pip
 => [4/4] RUN pip install --no-cache-dir  -r /opt/project/requirements.txt
 => exporting to image
 => => exporting layers
 => => exporting manifest sha256:a63d470e724b77a4e9bb90c0d99d3215cb88753608ebdf36c4844a417459ab75
 => => exporting config sha256:6db6a960d0f62980462a0b55dbf5beed69aaa2f8f7b9d7d2bda1b4425d50e6a4
 => => exporting attestation manifest sha256:fbb74712a7a284922f16cef875bf2676523730aa5e20b32e8b3a53203ba0bbf3
 => => exporting manifest list sha256:bb87528a69234a578938a05e87d3e1aea507e9c982304af783d64496e3301dbc
 => => naming to docker.io/library/selenium_crawl:extend
 => => unpacking to docker.io/library/selenium_crawl:extend

View build details: docker-desktop://dashboard/build/desktop-linux/desktop-linux/2pdye0memzxltxsdwvfmv2bt0
```

Nhập lệnh sau để kiểm tra xem image đã cài đặt thành công hay chưa

```
docker image ls
```

```
C:\Users\AN>docker image ls
REPOSITORY                   TAG       IMAGE ID       CREATED          SIZE
selenium_crawl               extend    bb87528a6923   10 minutes ago   3.09GB
selenium/standalone-chrome   latest    bcb1d054f3c8   5 days ago       3.04GB
```

Trong cùng thư mục, tiếp tục nhập lệnh sau để chạy container

```
docker-compose up
```

2. Chạy tools crawl

Mở một cmd hoặc terminal và nhập lệnh sau,

```
docker exec -it crawl-selenium-1 --country "Tên quốc gia" --keyword
"Keyword muốn tìm kiếm"
```

Kết quả tìm kiếm sau đó sẽ được lưu vào file crawl_resul.jsonl

```
{"name": "ENVIS School - Luyện thi IELTS", "address": "26 Đ. Láng, Thịnh Quang, Đống Đa, Hà Nội", "phone":
"0972 952 083", "web_link": "http://envis.edu.vn/", "email": "hello@envis.edu.vn"}
{"name": "ILA - Tây Sơn", "address": "Tòa nhà Phương Đông, 324 P. Tây Sơn, Ngã Tư Sở, Đống Đa, Hà Nội 100000",
"phone": "024 7307 1168", "web_link": "https://ila.edu.vn/", "email": null}
{"name": "Clever Academy", "address": "Viet Tower, 01 P. Thái Hà, Trung Liệt, Đống Đa, Hà Nội 10000", "phone":
"0975 861 994", "web_link": "https://cleveracademy.vn/", "email": "info@cleveracademy.vn"}
{"name": "PPSVietnam", "address": "5 Ngh. 95/8 P. Chùa Bộc, Trung Liệt, Đống Đa, Hà Nội 100000", "phone":
"0966 861 650", "web_link": null, "email": null}
{"name": "Trung Tâm Anh Ngữ Washington Language Center", "address": "66 P. Võ Thị Sáu, Thanh Nhàn, Hai Bà
Trưng, Hà Nội", "phone": "024 3625 0952 ext. 13", "web_link": null, "email": null}
{"name": "Pps Vietnam English center", "address": "ngõ 183 P. Trần Đại Nghĩa, Bách Khoa, Hai Bà Trưng, Hà
Nội", "phone": null, "web_link": null, "email": null}
{"name": "Scots English Nguyễn Xiển", "address": "BI-BT2, Khu đô thị mới, Thanh Trì, Hà Nội", "phone": "1900
252575", "web_link": "https://scotsenglish.edu.vn/", "email": null}
{"name": "Scots English Nguyễn Tuân", "address": "Lô 6 - Liền kề 2, số 90, 90 Đ. Nguyễn Tuân, Thanh Xuân
Trung, Thanh Xuân, Hà Nội", "phone": "1900 252575", "web_link": "https://scotsenglish.edu.vn/", "email": null}
{"name": "GEMS EDU- Hệ Thống anh ngữ quốc tế", "address": "2 P. Vương Thừa Vũ, Khương Trung, Thanh Xuân, Hà
Nội 120701", "phone": "0785 758 659", "web_link": null, "email": null}
{"name": "Amslink Láng Hạ 1", "address": "ngõ 59 P. Láng Hạ, Chợ Dừa, Ba Đình, Hà Nội", "phone": "024 7305
0384", "web_link": "http://amslink.edu.vn/", "email": "nguyensyminh@gmail.com, Hethonganhngu@amslink.edu.vn"}
```