SPAREPART CRAWLER

HOW TO INSTALL APPLICATION

Linux

- In Linux environment execute the script file `start.sh`

```
source ./start.sh
```

- Logs & Data results from crawling can be accessed "/logs" & "/data"

Windows

- In Windows, we will need to use Docker. First install docker engine on the computer. Assuming the docker service is up and running, build docker image

```
docker build -t="crawler"
```

- after successful building docker image, then run it with mounting volume on windows e.g "c:\Users\shared"

```
docker run -v "/c/Users/shared/logs:/app/logs" -v "/c/Users/shared/data:/app/data"
--rm -d -p 6800:6800 -p 9000:9000 crawling
```

- Logs & Data results from crawling can be accessed at mounted volume e.g "c:\Users\shared"

HOW TO USE APPLICATION

1. SCHEDULE CRAWLING

Description

Schedule crawling / scraping a website for a given car merk & model.

Endpoint

GET `/crawl`

Parameter

- Request Header
    - Basic Authentication

```
Authorization: Basic dXNlcm5hbWU6cGFzc3dvcmQ=
```

- Query String:

    - required:
    - `spider, merk, model`

    - **available spider:**

        - parts.com
        - megazip
        - daihatsu
        - suzuki
        - isuzu

    - optional:

    - `year, varian, mesin`
    - **for example:**

```
?spider=megazip&merk=toyota&model=vios&year=2009&varian=...&mesin=…&region=...&resume_job_id=abc123
```

Example Request

```
curl -H "Content-Type:application/json" -H "Authorization: Basic
dXNlcm5hbWU6cGFzc3dvcmQ="
-X GET "http://192.168.105.102/crawl?website=isuzu&merk=isuzu&model=panther"
```

2. CRAWL JOBS

Description

See pending, running, and finished crawl jobs.

Endpoint

GET `/jobs`

Example Request

```
curl -X GET "http://192.168.105.102:9000/jobs"
```

3. CANCEL JOBS

Description

Cancel Job on Crawl

Endpoint

GET /cancel_job/{job_id}

Request Header

- Basic Authentication
    - Authorization: Basic dXNlcm5hbWU6cGFzc3dvcmQ=

Example Request

```
curl -X GET
"http://192.168.105.102:9000/cancel_job/73ef9f78a52211e8ac210242ac110004"
```

4. CANCEL QUEUE

Description

For Cancel Queue on Crawler.

Endpoint

GET /cancel_queue/{spider}/{queue_id}

Request Header

- Basic Authentication
    - Authorization: Basic dXNlcm5hbWU6cGFzc3dvcmQ=

Example Request

```
curl -X GET "http://192.168.105.102:9000/cancel_queue/megazip/2"
```