Gebruikershandleiding

IPFIT6

Yassir Laaouissi | INF3A | 18-10-2020

Inhoud

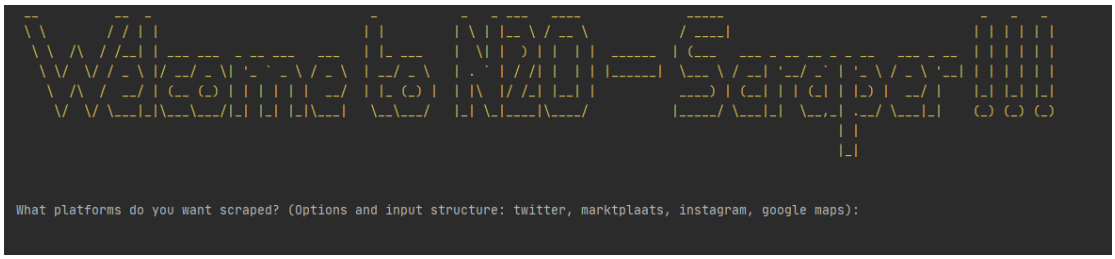
Over de tool

Deze tool functioneert als scraper voor de websites Marktplaats, Google Maps en Instagram op basis van een aantal zoektermen die vooraf worden aangegeven. Middels deze tool kan een gedeelte van een OSINT onderzoek geautomatiseerd worden.

INSTALATION AND BASIC USAGE

Python

- Zorg dat je de volgende packages hebt geinstalleerd via apt:
    - git
    - openjdk-11-jdk
    - chromium
    - chromium-driver
    - gcc
    - g++
    - python3
    - python3-pip
    - python3-dev
    - python3-wheel
    - ffmpeg
- Zorg daarna dat je de volgende packages met pip3 installeert:
    - Alle packages in het bestand requirements.txt
    - En als je twitter wilt gebruiken ook het commando: pip3 install --upgrade git+https://github.com/yunusemrecatalcam/twint.git@twitter_legacy2
- Start een terminal in de map van de scraper en voer het commando python3 main.py uit om te starten. In het volgende venster kan je kiezen welke platformen gescraped moeten worden, de output wordt in de map “Output” opgeslagen:



Docker

- Download de map genaamd scraper_linux.zip en pak het uit op een zelfgekozen locatie. En start een terminal in de uitgepakte map.
- Voer het volgende commando uit om een docker container te maken van de scraper: *docker-compose up –build*
- Maak nu een cronjob voor het commando "*docker-compose up*" in de map waar de scraper staat.
- De output wordt automatisch opgeslagen in de map "Scraper/Output".

Output

De output van de tool is verdeeld in vier mappen. Iedere map heeft de resultaten van een specifieke website. Oftewel vier mappen met resultaten. Hieronder is toegelicht hoe de resultaten voor iedere map eruitzien:

- **Marktplaats:** Voor dit platform zijn CSV en XLSX ingericht en gecategoriseerd per zoekterm en tijd waarop wordt gescraped.
- **Google Maps:** Voor dit platform is een XLSX-bestand ingericht met daarin alle Google Maps resultaten voor alle zoektermen. De naam van dit bestand is gebaseerd op de tijd waarop wordt gescraped.
- **Twitter:** Voor dit platform is een JSON-bestand ingericht met alle tweets die gevonden zijn aan de hand van de zoektermen. De naam van dit bestand is gebaseerd op de tijd waarop wordt gescraped.
- **Instagram:** Voor platform worden steeds mappen aangemaakt per tijd waarop gescraped met daarin dan weer mappen per zoekterm. In deze mappen worden de foto's, video's, beschrijvingen en comments van Instagram posts gedownload.