Studio e sviluppo di connettori di scraping di sorgenti social network per ingestione in sistemi di Big Data Analytics

Relatore:
Prof. Marco Anisetti

Tesi di Laurea di:
Francesco Avantaggiato

Università degli Studi di Milano
Sicurezza dei Sistemi e delle Reti Informatiche

15 Dicembre 2022

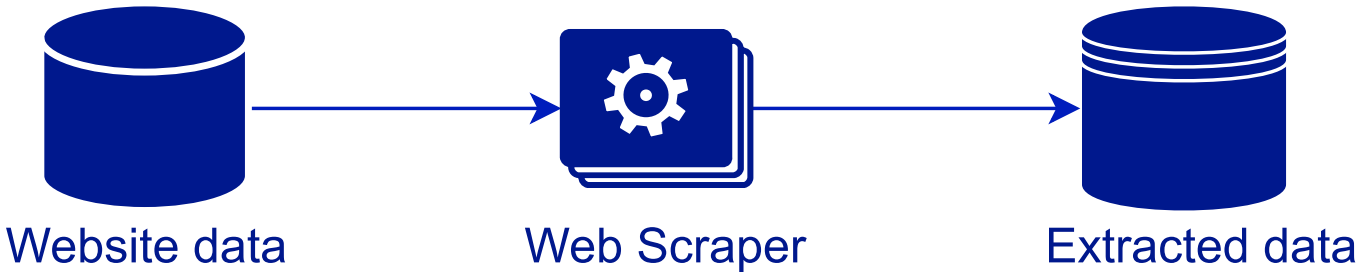
Attività svolta

1. Studio di soluzioni per web scraping da fonti social network (Facebook ed Instagram)
2. Identificazione del contrasto attuato dalle piattaforme
3. Sviluppo connettori di scraping
4. Sviluppo metodologie di elusione dei controlli

Web Scraping

Attività di raccolta automatica di dati da Internet attraverso:

- Application Programming Interface
- Richieste HTTP GET
- Tecniche dedicate



Aspetti legali e uso dei dati

L'attività rientra in una "zona grigia" della legalità in quanto non espressamente definita da alcuna norma.

- General Data Protection Regulation (GDPR - Europa)
- Computer Fraud and Abuse Act (CFAA - USA)
- Termini di servizio

I possibili casi d'uso dei dati estratti rientrano in settori investigativi e di Open Source Intelligence

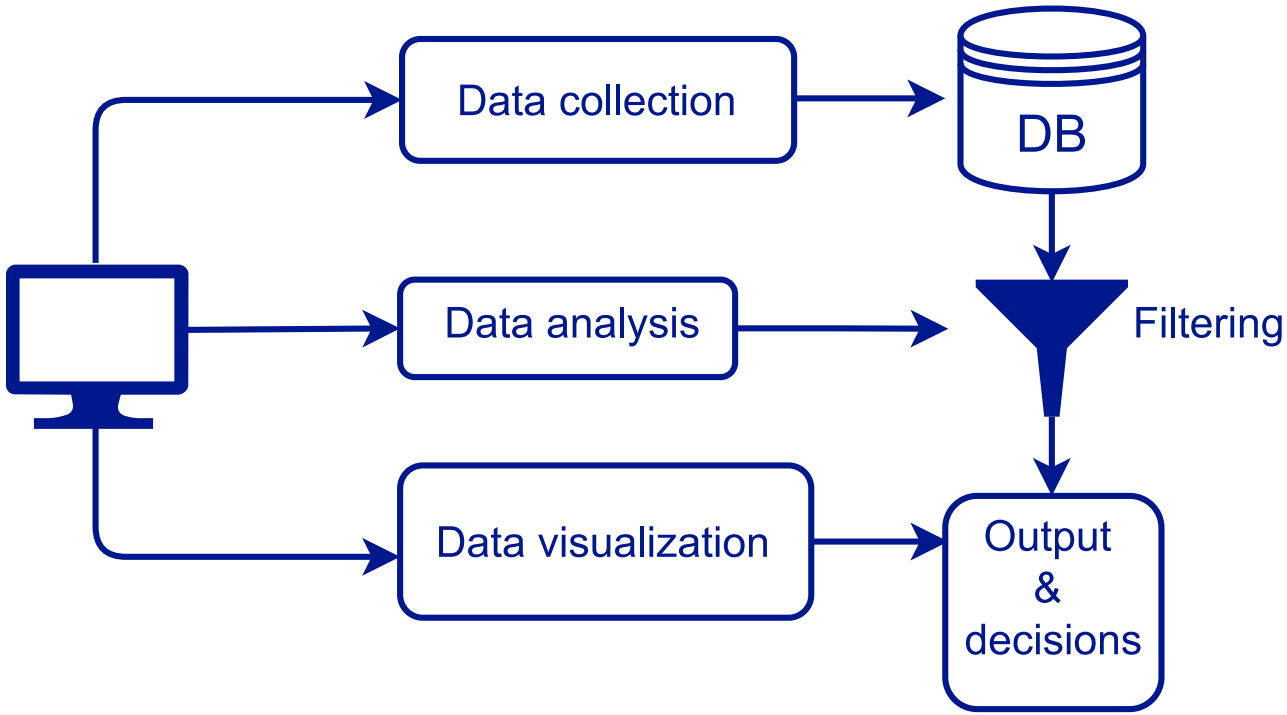
Connettori sviluppati

Facebook

- Tool open source “facebook-scraper”
- No API ufficiali

Instagram

- Tool open source “instaloader”
- No API ufficiali



Estrazione dei dati

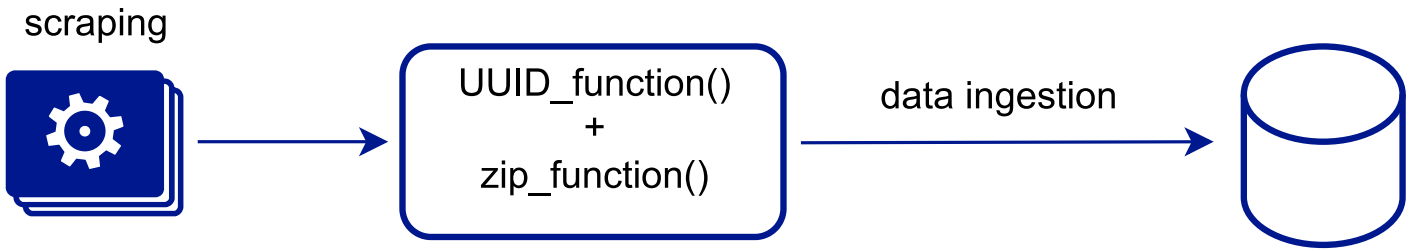
Facebook

- Informazioni generali da profili
- Post
- Gruppi pubblici e privati
- Pagine

Instagram

- Informazioni generali da profili
- Post
- Storie
- Follower e seguiti

La gestione dei dati in output avviene nella stessa modalità per entrambi i connettori.



Controlli anti-scraping

Le soluzioni di contrasto allo scraping si basano su:

- Autenticazione
- Ban
- Controllo richieste ed indirizzo IP
- Fingerprinting
- Aggiornamenti

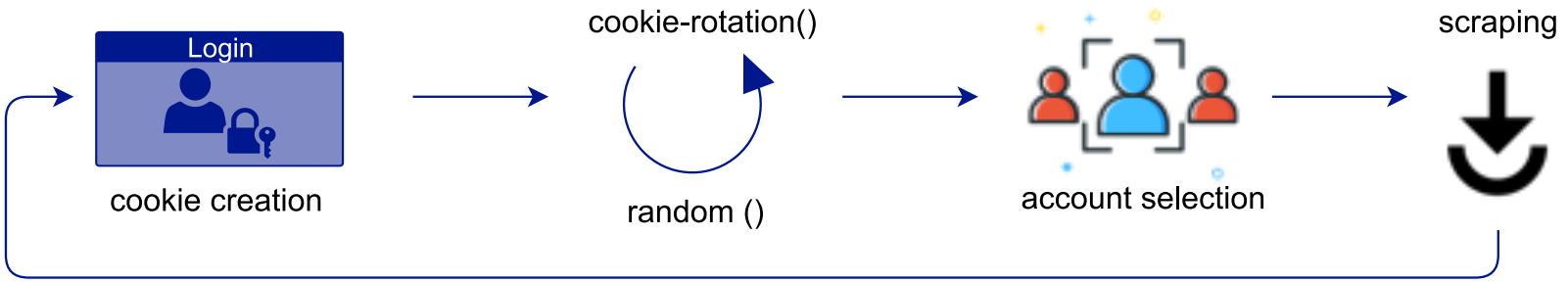
Metodi di elusione dei controlli

I metodi di elusione sviluppati sono:

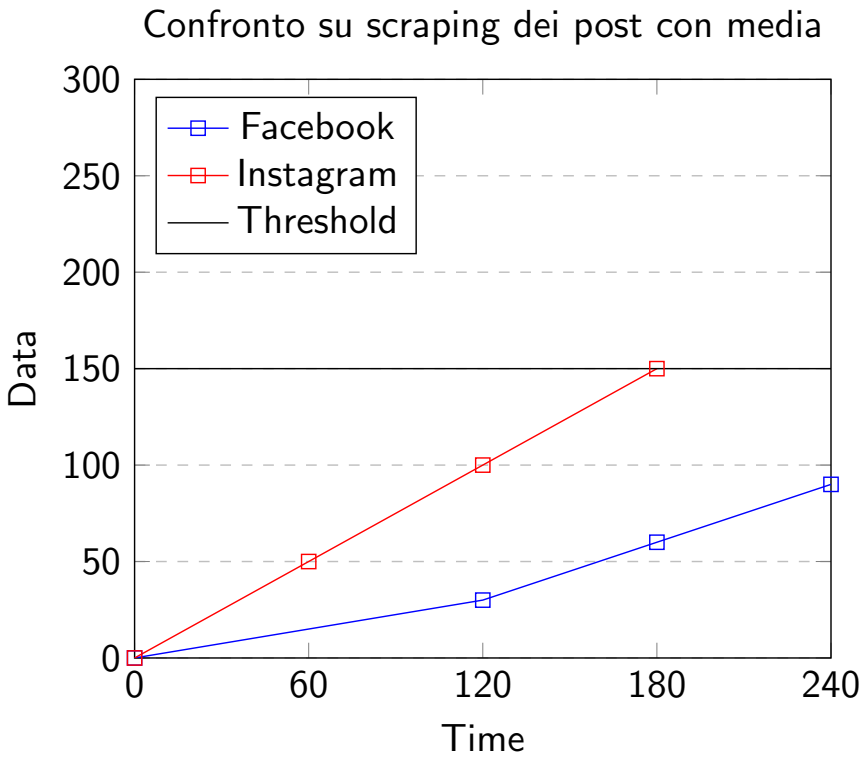
- Gestione di account multipli
- Rotazione di cookies ed account
- Ritardo tra le operazioni

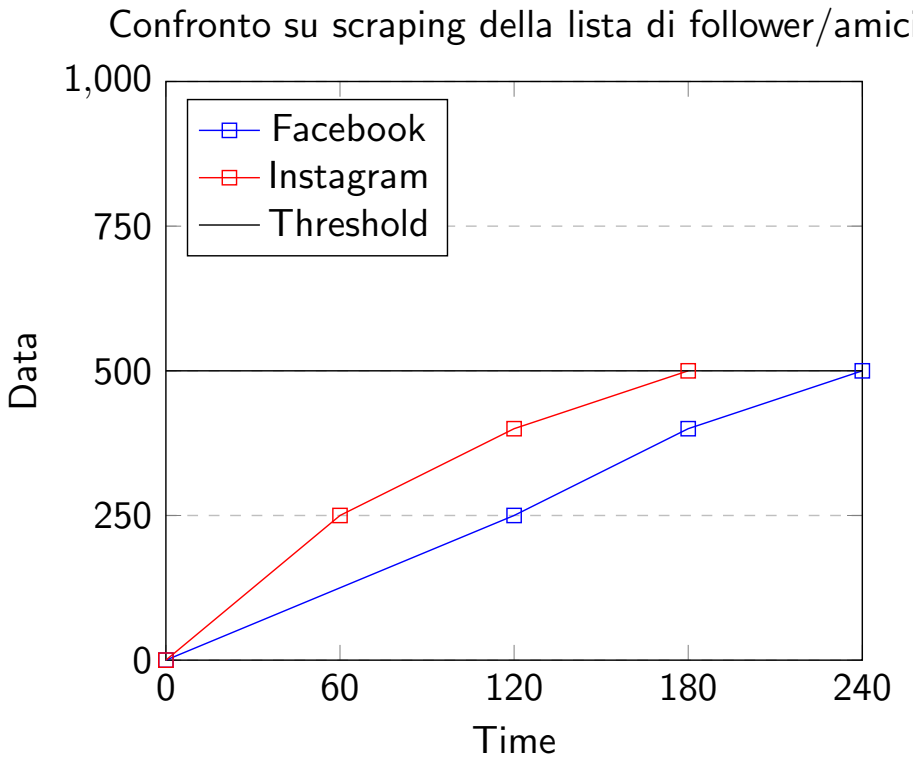
Cookie-rotation e automazione

- Utilizzo di Selenium per automazione browser
- Estrazione dei cookies
- Rotazione random e selezione account



Test e prestazioni





Conclusioni e sviluppi futuri

Grazie ai connettori sviluppati si è garantita:

- Continuità operativa dello scraping
- Buona elusione dei controlli
- Portabilità del software

Possibili sviluppi futuri possono basarsi su:

- Nuovi metodi di elusione (es. "ip-rotation")
- Implementazione dei connettori in sistemi distribuiti

Grazie per l'attenzione