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

Francesco Avantaggiato matr: 942790

Negli ultimi anni la crescente diffusione dei social network ha portato ad un incremento dei dati prodotti e condivisi su queste piattaforme. Sono quindi necessarie soluzioni Big Data che permettano di produrre valore dai dati prodotti, sia attraverso la collezione delle informazioni condivise dagli utenti che tramite l'analisi delle stesse. In tal senso risulta fondamentale il web scraping ovvero l'attività di acquisizione automatica di dati dal web ed è stata scelta nello sviluppo del progetto di tesi. Il lavoro verte sull'estrazione dati da fonti social, in particolare da Facebook ed Instagram, i quali applicano metodologie di contrasto tecnico allo scraping per tutelare e garantire la privacy degli utenti. Si è proposta e sviluppata una soluzione software consistente in due connettori in grado di gestire lo scraping, l'elusione dei controlli anti-scraping e l'ingestione dei dati su piattaforme di Big Data Analytics. In particolare l'operatività continua di questi connettori è garantita grazie alle soluzioni in grado di aggirare i controlli anti-estrazione.
Oggetto di attenzione nel progetto è anche l'aspetto legale di questa attività e l'utilizzo dei dati estratti, valutando soluzioni in ambito investigativo e di Open Source Intelligence.
Il lavoro svolto si articola come riportato di seguito.

1. **Studio dello stato dell'arte del web scraping da fonti social network**: individuazione della base di partenza e dello stato di fatto dei sistemi di scraping e del loro funzionamento. Sono stati studiati i punti di forza, le peculiarità e le relative problematiche dell'estrazione dati dai social network. L'obiettivo dichiarato dalle piattaforme, inerente alla tutela ed alla protezione dei dati dell'utente, si traduce nell'implementazione di metodologie di contrasto tecnico.

2. **Analisi concettuale di connettori di scraping da fonti social, requisiti e tecnologie**: presentazione di un connettore dati ideale, in grado di garantire la continuità operativa anche a fronte di problematiche dovute al contrasto dello scraping. Le tecnologie da impiegare prevedono la creazione di Application Programming Interface e di software container per determinare la portabilità del software. Il connettore deve inoltre prevedere, per la successiva ingestione in sistemi di Big Data Analytics, uno standard per la gestione dei dati in output, uniformando ed aggregando i dati eterogenei estratti. Il sistema ideale include accortezze tecniche e metodologie in grado di eludere le maggiori tipologie di contrasto, sempre più specifiche e funzionali.

3. **Sviluppo dei connettori ed integrazione dei tool**: presentazione del lavoro di sviluppo con il funzionamento dei tool impiegati e le peculiarità delle informazioni da estrarre dai social network. Sono stati individuati i dati di maggiore interesse operativo tra quelli messi a disposizione dal normale utilizzo di un utente del social. Sono stati sviluppati metodi di elusione quali: "cookie-rotation", "account-rotation", ritardo tra le operazioni e limitazione del numero di richieste. Il metodo innovativo e funzionale inerente la rotazione dei cookies consente il mantenimento della stabilità dell'applicazione, fornendo un'estrazione dati continua. Quando il processo viene interrotto a causa di provvedimenti contro l'account in uso, si avvia la rotazione randomica degli utenti e si riavvia lo scraping.

4. **Test e operatività**: le funzioni ed i connettori sviluppati sono stati valutati in termini di performance e velocità.

Il lavoro di tesi, ad integrazione di quanto sviluppato e proposto, si presta a sviluppi futuri tra i quali:

- l'ampliamento delle modalità di elusione, ad esempio integrando una rotazione degli indirizzi ip tramite proxy, oppure sviluppando i connettori in sistemi distribuiti.
- l'aumento della resilienza dei connettori, intesa come mantenimento della capacità operativa degli stessi a fronte di contrasto e problematiche tecniche.