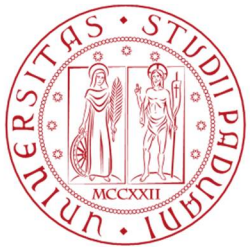

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Report Progetto Cyber Security

Social Networks – Scraping

Allegretti Tommaso

Matricola: 1201247

a.a. 2021-2022

Obbiettivi del progetto

Creare uno (o più) script che, dato nome e URL Instagram di una celebrità trovino:

- Profilo Facebook
- Profilo Twitter
- Sesso
- Data di nascita
- Etnia
- Interessi
- Orientamento sessuale
- Città di residenza attuale

Bisogna inoltre fornire le funzionalità di:

- Download foto profilo di Facebook
- Download foto profilo di Twitter
- Download foto profilo di Instagram
- Download degli ultimi 30 post di Instagram
- Generazione di un file di testo contenente l'ordine dei post scaricati

Introduzione

Per soddisfare i requisiti del progetto è stato adottato un approccio misto tra automatico e manuale.

Le informazioni vengono recuperate con processi di scraping automatico affiancate da un aiuto visivo per fare un check sulle informazioni trovate in automatico o nel caso fosse necessario l'inserimento manuale dei dati.

Sono stati utilizzati due siti web per effettuare lo scraping di informazioni: Wikipedia e Idolwiki.

Si è deciso di utilizzare due script separati per gestire l'inserimento nel database e il download delle immagini per evitare che l'utente debba sopportare inutili attese dopo ogni celebrità inserita in database.

Infatti, in questo modo, l'attesa del download è delegata completamente al secondo script che non richiede nessun input dell'utente.

Funzionamento degli script

Le funzionalità sono state suddivise in due script separati che utilizzano entrambi il modulo "**script.py**" il quale contiene le principali funzioni del programma.

create_database.py

Il primo script, "**create_database.py**", ha lo scopo di inserire nel file "dataset.csv" i dati che vengono ottenuti attraverso web scraping, senza eseguire il download.

L'esecuzione dello script segue questo sviluppo:

Viene stampata l'ultima voce della lista presente nel file "celebrities.json" e chiede se analizzarla o saltarla.

Se si decide di analizzarla viene chiesto il nome da inserire, si può dare input vuoto per confermare quello già presente.

Basandosi sul nome si effettua un'analisi sul sesso della celebrità presa in analisi attraverso il modulo **gender_guesser**, in caso il nome non sia presente nel database del modulo viene chiesta una conferma che la celebrità sia in effetti una persona vera (nel caso l'elemento analizzato sia un brand).

Viene chiesta conferma sul sesso, si può dare input vuoto per confermare quello trovato automaticamente.

Viene effettuato lo scraping di Idolwiki per trovare dati riguardanti data di nascita, etnia, orientamento sessuale e interessi.

Viene chiesta conferma sulla corrispondenza dei profili Facebook e Twitter trovati attraverso il modulo di **googlesearch**.

Viene effettuato lo scraping della pagina Wikipedia, si cerca di trovare le informazioni riguardanti il sesso del coniuge e la data di nascita:

- Il primo dato viene ricercato con **Selenium**, trovando il riquadro a destra della pagina dove sono presenti tutte le informazioni principali utilizzando l'xpath relativo ed all'interno di esso viene cercato il tag "tr" che contiene la stringa "Spouse(s)", viene recuperato l'intero contenuto del tag che viene quindi manipolato per ottenere solo i "first names" del/i coniuge/i che vengono poi usati per invocare la funzione di guess del gender che abbiamo visto prima. Attraverso questi risultati si può estrapolare se la celebrità abbia avuto solo coniugi maschi, solo femmine o entrambi, questo abbinato al sesso della celebrità stessa genera un risultato automatico che indica quale possa essere il suo orientamento sessuale
- La data di nascita viene trovata in un modo analogo a quello appena descritto ma in maniera semplificata, una volta trovato il riquadro usando l'xpath relativo basta cercare la voce che contiene la stringa "Born" o "Date of birth" per ricavare la data di nascita, questa viene poi manipolata per ricavarne giorno, mese e anno di nascita

Dopo aver effettuato lo scraping di Wikipedia viene confrontato il gender del/la coniuge con il gender della celebrità per ottenere l'orientamento sessuale.

Viene aperto su Google il risultato della ricerca "*nome celebrità* + age" per fornire un aiuto in caso la data di nascita trovata automaticamente non sia soddisfacente o sia mancante.

Viene mostrata la data di nascita suggerita in base ai dati trovati tramite lo scraping, la scelta consigliata viene stampata in diversi modi a seconda dei seguenti casi:

- È stata trovata una data di nascita sia su Idolwiki che su Wikipedia e sono uguali: viene suggerita la data trovata in entrambe le pagine
- È stata trovata una data di nascita sia su Idolwiki che si Wikipedia ma sono diverse l'una dall'altra: vengono suggerite due opzioni per confermare la prima o la seconda
- È stata trovata una data di nascita solo in una delle due pagine: viene suggerita la data trovata
- Se nessuna delle precedenti condizioni è soddisfatta allora viene chiesto di inserire manualmente la data di nascita

Viene stampato il suggerimento per l'etnia (se è stata trovata).

Viene chiesto di inserire in input l'etnia della celebrità analizzata.

Viene stampato il suggerimento per gli interessi.

Viene chiesto di confermare il suggerimento o inserire manualmente la voce.

Viene stampato il suggerimento per la voce orientamento sessuale, i dati per ottenere una predizione automatica generano una soluzione simile a quella vista per la data di nascita:

- È stata trovato un orientamento sessuale sia su Idolwiki che tramite la comparazione del sesso della celebrità con il sesso del/i coniuge/i e sono uguali: viene suggerito l'orientamento sessuale trovato in entrambi i metodi
- È stato trovato l'orientamento sessuale sia su Idolwiki che tramite la comparazione del sesso della celebrità con il sesso del/i coniuge/i ma sono diversi: vengono suggerite due opzioni per confermare la prima o la seconda
- È stato trovato un orientamento sessuale solo in uno dei due metodi: viene suggerito l'orientamento sessuale trovato
- Se nessuna delle precedenti condizioni è soddisfatta allora viene chiesto di inserire manualmente l'orientamento sessuale

Viene aperto su Google il risultato della ricerca "where does + *nome celebrità* + live" per fornire un aiuto in caso la città di residenza trovata automaticamente non sia soddisfacente o sia mancante.

I dati raccolti vengono inseriti all'interno del file "dataset.csv" e viene chiesto se si desidera avanzare alla prossima iterazione o terminare il programma.

download_database.py

Questo script ha una funzione molto più semplice e facile da capire rispetto alla parte precedente del programma.

Lo script si limita a scorrere tutta le righe nel file “dataset.csv” e ad eseguire le seguenti operazioni su ognuna di esse:

- Download immagine profilo di Instagram
- Download ultimi 30 post in ordine cronologico di Instagram
- Creazione (o aggiornamento) di un file di testo contenente l’ordine cronologico dei post scaricati da Instagram
- Download immagine profilo da Facebook
- Download immagine profilo da Twitter

Requisiti per il funzionamento

Perché il programma funzioni correttamente è necessario soddisfare alcuni requisiti:

- Avere Google Chrome installato
- Avere la versione di Chromedriver corrispondente alla versione di Chrome e posizionarla nella stessa cartella dove sono presenti gli script
- Aver installato Python (personalmente ho usato la versione 3.10.2)
- Aver installato i moduli utilizzati dagli script (Selenium, Beautiful Soup 4, gender_guesser, pandas, googlesearch)
- Avere a disposizione una connessione ad internet
- Essere in possesso di credenziali di login Instagram valide

Analisi delle prestazioni

Inserimento

Il tempo necessario per l'aggiunta di un nuovo elemento è di circa 1 minuto, questo dato può variare a seconda della quantità di dati trovati in modo automatico.

Ad esempio, una celebrità molto conosciuta, di cui esiste una pagina Idolwiki, può richiedere solo 20 – 30 secondi per essere inserita in database.

Se esistono, i profili Facebook e Twitter hanno un'altissima probabilità di essere trovati (~99%).

Nei test effettuati non è mai avvenuto che le query non risalissero al profilo corrispondente della celebrità.

La probabilità che la celebrità analizzata abbia una pagina Idolwiki dedicata è di circa il 60% e, se questa viene trovata, c'è un'alta probabilità (~90%) che vengano trovate: data di nascita, etnia, interessi, orientamento sessuale.

Alta probabilità (~60%) che venga rilevato il gender corretto con gender_guesser, questa percentuale viene abbassata drasticamente dalle celebrità che usano pseudonimi/nomi d'arte.

Se si dovessero analizzare solo nomi propri la percentuale di riuscita sarebbe molto più alta poiché il modulo utilizzato è in grado di riconoscere nomi occidentali, est europei e indiani.

Un'idea per trovare la città corrente era stata quella di usare la pagina Wikipedia della celebrità e trovare la voce "Billed from" per risalire al domicilio, tuttavia questa soluzione si è rivelata molto inefficiente ed è risultato più semplice utilizzare un approccio più diretto per compilare il campo.

Download

Il tempo necessario per il download di tutti gli elementi richiesti (immagine profilo FB, TW, IG e ultimi 30 post IG) è di circa 2 minuti e 20 secondi per ogni celebrità.

Questo numero può scendere fino a poco più di 1 minuto nel caso sia già stato fatto il download degli elementi, infatti il modulo instaloader è in grado di riconoscere i post che sono già stati scaricati.

Trovati i profili Facebook e Twitter, il download delle immagini profilo ha sempre avuto successo.

Ci sono state solo 3 occasioni nelle quali non si è riuscito ad accedere alla pagina Instagram (la mia opinione è che il link inserito nel file json fosse sbagliato o obsoleto).

Il modulo instaloader presenta dei problemi nell'effettuare molte operazioni ravvicinate nel tempo e dopo circa 10 profili scaricati si ferma per circa 5 minuti.

Questo non costituisce un problema grave poiché allo scadere del timeout riprenderà il programma automaticamente.