Report Progetto Cyber Security

Web Scraping

Obbiettivo del progetto:

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

* 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

Raggiungimento dei requisiti

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

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:

Stampa la voce della lista presente nel file “celebrities.json” attuale e chiede se analizzarla o skipparla

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

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 google search

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 trovato 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 contenga 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 su 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” per ricavare la data 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 aperta 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:

* E’ stata trovata una data di nascita sia su Idolwiki che su Wikipedia e sono uguali: viene suggerita la data trovata in entrambe le pagine
* E’ 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
* E’ 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 stampata 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:

* E’ 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
* E’ 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
* E’ 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 aperta 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