Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Informatica per il management

RACCOLTA DI DATI SU LINKEDIN: ANALISI DELLE API E USO DI WEB SCRAPER

Relatore:
Prof.
Angelo Di Iorio

Presentata da:
Samuele Berni

Terza sessione
Anno Accademico 2021-2022

Abstract

L'idea della tesi nasce da un'esigenza evidenziata dal progetto RI-PLASMA.
RI-PLASMA è un progetto finanziato dall'Università di Bologna che ha l'obbiettivo di studiare gli spillover tecnologici dei lavori di ricerca, cioè analizzare la loro influenza e il loro effetto sulla tecnologia. Nel caso specifico verificare se la divulgazione di articoli su LinkedIn generi idee che si concretizzeranno nel deposito di brevetti.
Questa tesi ha lo scopo di eseguire uno studio di fattibilità sulle API e sui web scrapers utilizzabili per ottenere dati sulla biografia e sulla produzione scientifica di 6671 ricercatori i cui nominativi mi sono stati forniti dai responsabili del progetto RI-PLASMA. Questi nomi derivano dagli autori dei progetti finanziati dall'European Research Council (ERC) durante i programmi FP7 e H2020. In particolare cercherò, se presenti, gli articoli da loro pubblicati sulla piattaforma LinkedIn [1] con gli strumenti forniti dal social in questione o attraverso altri mezzi.
LinkedIn è uno dei social network più diffusi al mondo nell'ambito business, dove è possibile trovare offerte e opportunità di lavoro, pubblicare annunci e ricercare potenziali candidati. Inoltre permette anche di rimanere aggiornati sui vari trend dei settori di interesse oppure migliorare il proprio personal branding e la reputazione online. Gli utenti del social hanno la possibilità di creare profili visualizzabili da altre persone iscritte e non alla piattaforma.
Tenterò, senza successo, di utilizzare l'API ufficiale di LinkedIn per l'estrazione dei documenti generati dai ricercatori membri di LinkedIn.
L'API è un insieme di definizioni e protocolli da cui è possibile produrre e completare applicazioni software. Vengono utilizzate come punto d'accesso e possono quindi permettere una varietà di funzioni tra cui:

- Inserire dei contenuti nei servizi web.
- Inoltrare un comando al software e ricevere una risposta.
- Controllare l'accesso da parte di altri programmatori.

L'API di LinkedIn, per numerosi problemi di data breach causati da una sua vulnerabilità, non è utile allo scopo in quanto gli amministratori del social hanno deciso di limitarne notevolmente l'accesso facendo in modo che possa essere utilizzata solo da una

piccola cerchia di soggetti.
Per poterla sfruttare è necessario compilare un form inserendo dati aziendali che io come privato non possiedo e che ne impedisce di fatto l'utilizzo.
Non potendomi servire dell'API ufficiale di LinkedIn, sarà necessario ricercare gli oltre 6100 nomi di ricercatori e associarli ad un profilo LinkedIn. Questo mi permetterà di ottenere l'URL del profilo che sarà impiegabile da altre API per potere analizzare la pagina dei ricercatori. Occorrerà analizzare i risultati ottenuti e verificare se sarà possibile estrarre, una volta evidenziati i profili, gli articoli pubblicati dai vari ricercatori.
Per raggiungere la fase di associazione dei nomi all'effettivo account di LinkedIn, vista la grande quantità di nomi, è necessario utilizzare un web scraper.
Il web scraping è una tecnica informatica utilizzata per l'estrazione di dati da un sito web attraverso l'utilizzo di programmi software: essi sono in grado di riconoscere la struttura di una pagina web, senza l'intervento umano, ottenendo informazioni e dati significativi; in questo caso specifico dovrà estrarre l'URL relativo al ricercatore. Una volta ottenuto l'URL del profilo bisognerà estrapolare gli articoli pubblicati dal soggetto interessato; sarà quindi necessario utilizzare un API che con il link ottenuto in precedenza dallo scraper potrà inoltrare tutte le informazioni richieste.
L'attenzione verrà posta nello scoprire cosa si può ottenere, attraverso l'utilizzo di vari API e di vari scraper, dai profili di determinati utenti, i quali hanno postato in precedenza articoli di interesse scientifico.
Dopo aver eseguito un'analisi sui software, valutando quindi l'utilità e il prezzo, verranno studiati anche gli aspetti legali sull'utilizzo di queste piattaforme e quanto sia effettivamente possibile utilizzare questi programmi rispettando la legge o comunque non andando contro la policy di LinkedIn.

Indice

Capitolo 1

Descrizione del progetto

In questo capitolo vengono introdotte e descritte le fasi del progetto. Verrà trattato inizialmente il progetto RI-PLASMA, il progetto da cui è nata l'idea di questa tesi, a seguire il workflow dove verrà spiegato in dettaglio il perchè delle varie scelte e il percorso seguito per il raggiungimento dell'obbiettivo finale, ovvero ottenere gli articoli pubblicati dai ricercatori. Nelle ultime sezioni verrà fatta un'introduzione su alcuni temi tecnici che verranno visti in dettaglio nei successivi capitoli.

1.1 RI-PLASMA: Research for Innovation: Patents Linked to Altmetrics and Social Media in Academia

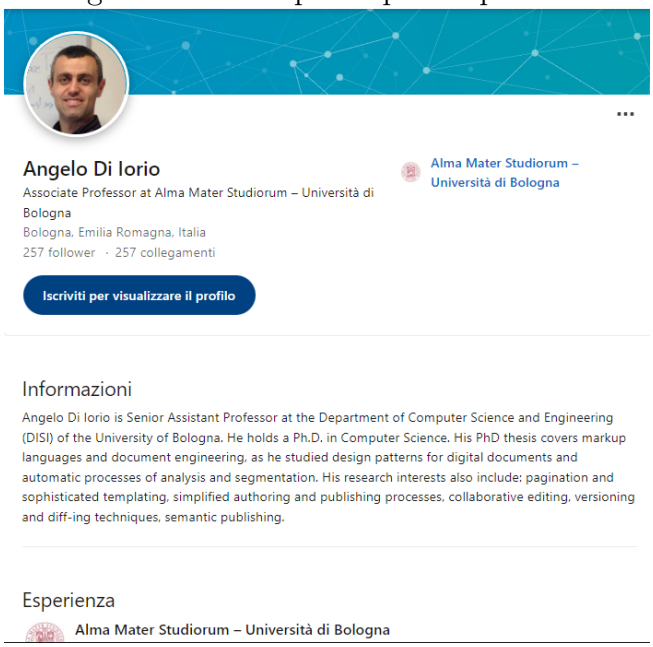
La seguente descrizione è stata direttamente estratta dal progetto RI-PLASMA.
Il progetto RI-PLASMA è un progetto finanziato dal bando AlmaIdea 2022 dell'Università di Bologna, e coordinato dalla Prof.ssa Laura Toschi, del Dipartimento di Scienze Aziendali; vede inoltre coinvolto come co-PI il Prof. Angelo Di Iorio del Dipartimento di Informatica Scienza e Ingegneria. Il progetto parte dall'idea, ormai condivisa, che le pubblicazioni scientifiche non sono solo l'output finale delle attività accademiche ma il risultato intermedio del processo di trasferimento della conoscenza. Questo ha portato diversi ricercatori ad interessarsi dei legami fra produzione di pubblicazioni e generazione di brevetti (Jaffe, 1989; Jaffe e Trajtenberg, 1998). Il progetto intende, in primo luogo, analizzare la capacità dei progetti di ricerca universitaria di generare pubblicazioni in grado di ispirare a loro volta brevetti successivi, generando spillover tecnologici. L'obiettivo è collegare progetti di ricerca, pubblicazioni e brevetti che citano tali pubblicazioni nella sezione 'Non Patent Literature' per rispondere alla domanda: qual è la portata dell'impatto tecnologico generato dalla ricerca universitaria? In particolare, si concentrerà su un campione di circa 6000 progetti di ricerca finanziati dall'European Research Coun-

cil (ERC) durante i programmi FP7 e H2020. Un secondo obiettivo è valutare se tali spillover tecnologici siano rafforzati dalle attività di divulgazione della ricerca scientifica attraverso canali non tradizionali, come ad esempio l'uso di piattaforme di social media in grado di dare visibilità della ricerca universitaria a un pubblico più ampio.
Una delle piattaforme che si prevede di studiare è LinkedIn, così come altre piattaforme ad esso collegate che permettono di ottenere dati sulla biografia e produzione scientifica e non dei ricercatori.
L'obiettivo della tesi è quindi quello di fare un primo studio di fattibilità su queste API e identificare metodi e processi da seguire per raccogliere dati, da riutilizzare anche per altre ricerche.

1.2 LinkedIn

LinkedIn è un social media professionale fondato nel 2002 da Reid Hoffman con lo scopo di connettere professionisti e aziende in tutto il mondo. E' stata acquisita nel 2016 da Microsoft per circa 26 miliardi di dollari. LinkedIn offre una serie di strumenti per le aziende, inclusi strumenti di recruiting, di marketing e di formazione. Gli utenti, all'interno del social media, possono creare un profilo includendo informazioni lavorative, permettendo loro di trovare lavoro o comunque cercare nuove opportunità di carriera. Esistono due tipi di profili LinkedIn: pubblici e privati.

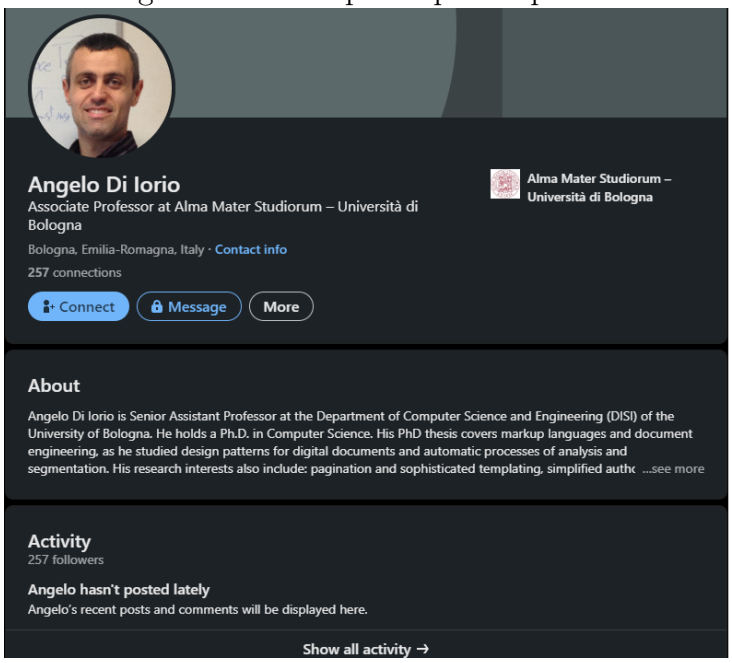
Figura 1.1: Esempio di profilo pubblico



Nei profili pubblici si ha la possibilità di visualizzare il contenuto senza la necessità di

essere iscritto alla piattaforma. Essi contengono meno informazioni, ad esempio non è descritto quali siano le sue skills e non c'è la possibilità di inviargli un messaggio; inoltre possono venire nascoste notizie quali educazione ed esperienze. Quindi i profili pubblici possono essere visualizzati da chiunque sul sito, anche da utenti non connessi o non membri del network. La pagina pubblica può risultare utile per ricerche di lavoro, la creazione di una rete di contatti professionali e la ricerca delle informazioni disponibili su persone e aziende.

Figura 1.2: Esempio di profilo privato

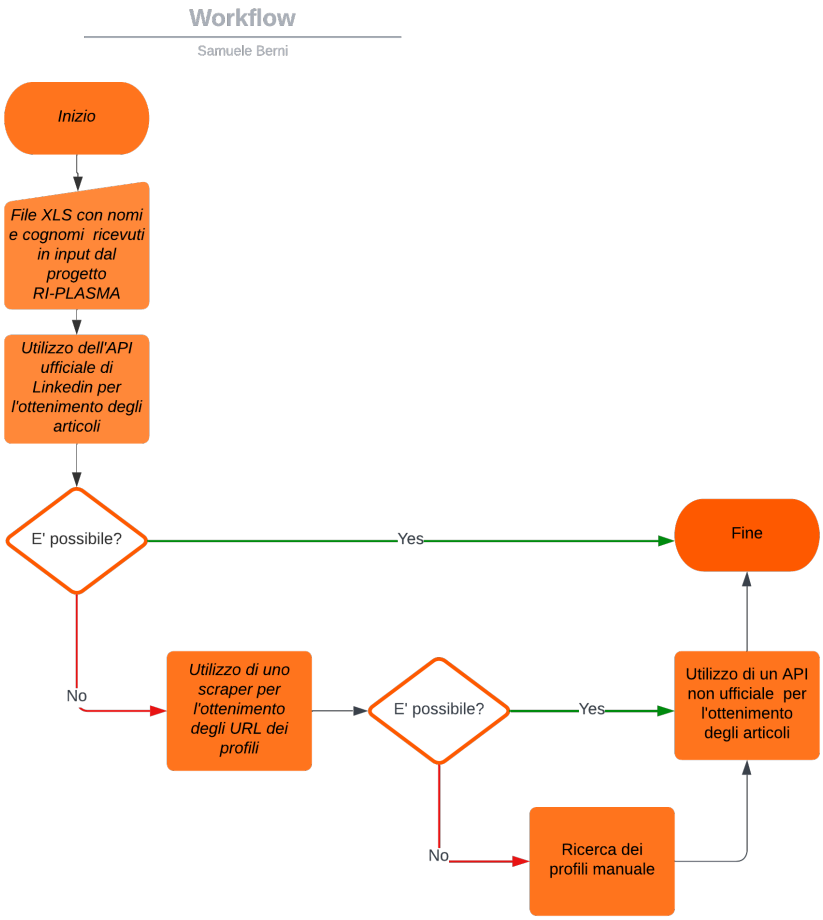


I profili privati non sono visualizzabili senza aver eseguito l'accesso e contengono al loro interno una maggiore quantità di informazioni. Non sono quindi visibili da tutti gli utenti del sito, ma è necessario essere connessi e registrati alla piattaforma. Un profilo privato può aiutare a proteggere la privacy e quindi limitarne la visibilità dei propri dati personali e/o professionali. Nello stesso tempo limitandone l'accesso sarà più difficile la costruzione di una rete di contatti o la ricerca di un opportunità di lavoro. I profili privati hanno la possibilità di utilizzare la private mode: permette di essere completamente invisibile su LinkedIn anche dagli altri utenti loggati e si ha la possibilità di visualizzare tutti i profili senza notificare il proprietario; allo stesso tempo però il tuo profilo non sarà ricercabile in alcun modo. La differenza fondamentale, per quel che mi riguarda ai fini della tesi, è il fatto che sui profili pubblici può avvenire lo scraping, mentre su quelli privati non è possibile. Attualmente LinkedIn continua il suo tentativo di bloccare ogni tentativo di scraping (verrà trattato nel dettaglio in una sezione successiva) sui profili privati, con l'intento di proteggere i propri utenti. Il 1° Febbrario 2022 LinkedIn ha rilasciato una

comunicazione : ha citato in causa Mantheos [2], un azienda con sede a Singapore che, utilizzando numerosi profili fake e riuscendo ad ottenere l'iscrizione a LinkedIn Sales Navigator, ha potuto ad eseguire lo scraping di milioni di profili privati. Le informazioni di questi profili erano quindi visibili solamente in caso di accesso. Queste azioni hanno violato l'accordo per gli utenti e l'informativa sulla privacy di LinkedIn e la legge. Questo caso verrà trattato in maniera più approfondita all'interno del capitolo di LinkedIn.

1.3 Workflow

Figura 1.3: Workflow



Il flusso di lavoro della tesi è una sequenza di azioni che definiscono il processo di lavoro per poter raggiungere un determinato obbiettivo che, nel mio caso, è quello di poter

ottenere dai nomi dei ricercatori tutti gli articoli pubblicati nel loro profilo personale di LinkedIn. Ci sono diverse strade possibili da seguire e da verificare se effettivamente sono praticabili. La via più conveniente, che permette di ottenere gli articoli in maniera semplice ed efficace, è sicuramente quella in cui viene utilizzata l'API ufficiale di LinkedIn; ma non sarà percorribile per diversi motivi che verranno trattati nel dettaglio nel prossimo capitolo. Sfrutterò le potenzialità degli scraper per prelevare l'URL appartenente al profilo di un ricercatore: solo a questo punto sarà possibile passare l'URL all'API.
L'API esaminata al suo interno eseguirà uno scraping di un determinato profilo fornendo gli articoli pubblicati appartenenti al ricercatore considerato. Nel caso fosse possibile, bisogna verificare quanti articoli di risposta è possibile ricavare, e nel caso ci fosse un limite capire quale sia.

1. In input verranno trattati oltre 6100 nomi di ricercatori provenienti dal progetto RI-PLASMA. Essi saranno contenuti all'interno di un file XLS (che poi verrà trasformato in XLSX).

2. Bisognerà chiedersi se è possibile servirsi dell'API ufficiale di LinkedIn per richiedere tutti gli articoli dei ricercatori profilati. Nel caso fosse possibile non saranno necessari altri passaggi.

3. In caso contrario, dati i nomi sarà necessario utilizzare uno scraper per poter ottenere l'URL associato. Si otterrà quindi il link relativo al profilo del ricercatore. Verranno presi in esame solamente i profili pubblici, in quanto come spiegato in precedenza, saranno gli unici che potranno essere tenuti in considerazione. Nel caso non fosse possibile utilizzare uno scraper bisognerà effettuare una ricerca dei profili manualmente. Dopo aver fatto ciò, sarà necessario caricare gli URL nell'editor, che verranno utilizzati per fare le richieste all'API.

4. Bisognerà ricercare un API non ufficiale che sia in grado di eseguire la stessa operazione e verificarne eventuali limiti legali.

1.4 API

In questa sezione verrà spiegata la nascita e il funzionamento delle API. Verrà quindi descritta la sua evoluzione, alcuni vincoli strutturali e come permetterne l'accesso in sicurezza.

1.4.1 L'evoluzione delle API

Le API (Application Programming Interface) sono state sviluppate negli anni '60 come parte delle prime architetture di sistemi operativi. La loro funzione principale era quella

di fornire un modo standardizzato per i programmi di accedere ai servizi di sistema, come l'input/output, la gestione della memoria e altre funzioni di basso livello. L'utilizzo delle API nel web è iniziato attorno agli anni 2000. Da allora l'interesse è costantemente aumentato fino a portarci nella moderna economia delle API. Questo è dovuto dal fatto che siamo interconnessi come mai prima. Anche il settore delle applicazioni mobile ne fa un grande utilizzo. Ci sono varie definizioni di API, fatte da diverse prospettive: alcuni autori decidono di fornire una concreta definizione, mentre altri preferiscono descrivere gli attributi e le caratteristiche dell'API. La caratteristica più importante rimane il fatto che sono un elemento di intermediazione tra gli utenti o i clienti e le risorse o servizi web: inoltre è il mezzo con il quale un'organizzazione può condividere risorse e informazioni.

Definition 1.1. APIs as the collection of codes, packaged with interfaces that aid other developers to use it [3]

La documentazione di un'API (Application Programming Interface) è una descrizione dettagliata delle funzionalità e dei comportamenti offerti dall'API: è cruciale per gli sviluppatori che vogliono utilizzare l'API per costruire applicazioni.

Essa può includere informazioni sulle richieste e le risposte accettate dall'API, sulla struttura dei dati e sulle eventuali limitazioni o restrizioni. Può anche includere esempi di codice e descrizioni dettagliate dei singoli endpoint dell'API.

Questo manuale viene spesso generato automaticamente a partire dal codice sorgente o dalla specifica di un'API, utilizzando strumenti come Swagger o Postman. Tuttavia, è importante che la documentazione sia verificata e mantenuta manualmente, poiché i cambiamenti nel codice potrebbero non essere riflessi automaticamente nella documentazione.

1.4.2 Quando si può parlare di API RESTful

REST non è nè un protocollo nè uno standard, bensì un insieme di vincoli architetturali. Viene introdotto per la prima volta nel 2000 all'interno di una tesi per il dottorato di Roy Fielding [4]. Esso utilizza un sistema di trasmissione dei dati che adopera le funzioni HTTP necessarie per identificare precise risorse. Il REST descrive un insieme di vincoli e principi, che devono essere rispettati, stabiliti dallo stesso Roy Fielding.

- Stateless: il principio si basa sull'assenza di stato durante la comunicazione tra client e server. E' fondamentale che durante la richiesta, il client invii tutte le informazioni necessarie affinchè questa venga evasa correttamente. Ogni sessione client-server è unica e non deve essere quindi connesssa in alcun modo alle precedenti o future

- Client-server: Si basa sull'architettura client-server, separando i compiti tra loro, evitando quindi un qualsiasi tipo di sovrapposizione di competenze

- Cacheable: i dati vengono memorizzati nella cache per ottimizzare le interazioni tra client-server. Le risposte date del server devono però essere esplicitamente etichettate come memorizzabili nella cache per evitare che venga salvata ogni risposta. In questo modo si riesce ad evitare che il client utilizzi stati vecchi o errati.

- Sistema a più livelli: è basato su livelli composti che garantiscono politiche di sicurezza.

- Interfaccia uniforme: l'interfaccia deve essere uniforme per i componenti.

Andiamo a vedere quali sono i vantaggi principali che ci fornisce la struttura REST.
Le API REST sono indipendenti da qualsiasi linguaggio o piattaforma utilizzata, garantendo massima libertà. Si ha quindi la possibilità di usarle su un server con qualsiasi linguaggio. Un vincolo fondamentale, client-server, rappresenta anche un vero e proprio vantaggio. I componenti infatti hanno la possibilità di evolversi indipendentemente, quindi è possibile modificare una sola parte progettuale senza dover modificare l'altra. Ciò si traduce anche in una migliore scalabilità.

1.4.3 Sicurezza di un API

Innanzitutto è necessario capire perchè è così fondamentale la sicurezza di un API.
API esposte sono la causa di gravi violazioni di dati che permettono quindi di rendere pubbliche delle informazioni sensibili. L'esposizione di dati sensibili non è però l'unica minaccia presente, infatti è possibile che avvenga un iniezione di codice malevolo. La tipologia più diffusa sono le SQL injection che prende di mira tutte quelle applicazioni che si poggiano ai database relazionali che quindi usano il linguaggio SQL. La sicurezza di un sistema coinvolge i seguenti aspetti:

- L'autenticazione: è necessario riuscire ad identificare le parti che vengono coinvolte in una comunicazione.

- L'autorizzazione: bisogna concedere la possibilità di accesso ad una determinata risorsa.

- Integrità: i dati non possono essere modificati da terzi.

- Riservatezza: la capacità di mantenere i dati privati durante lo scambio di informazioni.

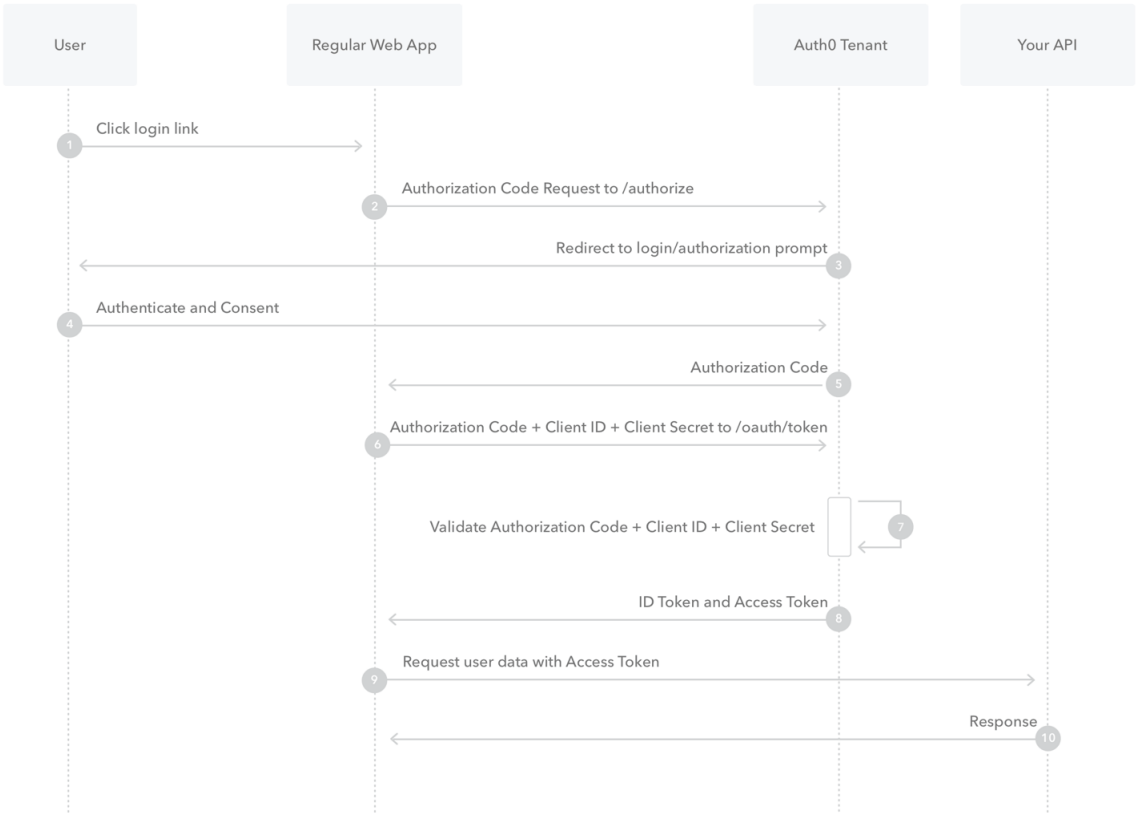
OAuth è lo standard open che regola la delega all'accesso, consentendo agli utenti di concedere a terzi l'accesso alle risorse web senza dover fornire nè username nè password. L'API di LinkedIn utilizza OAuth2, versione rilasciata nel 2012 come revisione completa della precedente versione e attualmente lo ha in gran parte rimpiazzato. Come OAuth2

gestisce le autenticazioni? Esistono due flussi principali per garantire l'accesso ad un determinato client: Client Credentials e l'Authorization Code. Verrà trattato solo il secondo dato che è ampiamente utilizzato nelle API e in particolare in quella di LinkedIn.

Authorization Code flow

Denominata "three-legged OAuth". Viene scambiato un Authorization Code per un token. Durante lo scambio viene inoltre passato il Client Secret.

Figura 1.4: Authorization Code Flow



1. L'utente accede regolarmente all'applicazione web
2. L'SDK reindirizza l'utente nel Auth0 Authorization Server
3. L'Authorization Server a sua volta, reindirizza l'utente nella pagina di login richiedendo l'autorizzazione
4. L'utente esegue l'autenticazione

5. L'Authorization Server reinderizza nuovamente l'utente all'applicazione con un authorization code, che potrà essere utilizzato una sola volta

6. L'SDK spedisce l'authorization code all'Authorization Server insieme al Client ID e al Client Secret.

7. L'Authorization Server verifica le credenziali

8. L'Authorization Server risponde con un ID Token e un Access Token

9. L'utente potrà utilizzare l'Access Token per eseguire le richieste alla specifica API.

1.5 Cos'è lo scraping

Lo scraping verrà impiegato per riuscire a mettere in relazione i nomi dei vari ricercatori con il loro profilo di LinkedIn, ottenendo quindi l'URL di quest'ultimo: sarà necessario per ricavare dei dati da pagine html. In particolare lo web scraping è uno strumento utile per la raccolta di dati online. Nella sua forma più generale si intende una tecnica in cui un programma informatico estrae alcuni dati dall'output generato da un altro programma: funziona inviando richieste a un sito web o a un'altra fonte online, quindi estraendo e raccogliendo i dati dalle risposte ottenute.. Lo scraping può essere utilizzato per molte finalità diverse, come la raccolta di informazioni per la ricerca di mercato, la raccolta di dati per l'analisi, la creazione di database di contatti, l'automazione di attività ripetitive e molto altro ancora. In realtà il processo del web scraping è abbastanza semplice e si suddivide in tre fasi:

1. Lo scraper, ovvero il codice che è in grado di estrarre le informazioni, invia una richesta GET ad un determinato sito web

2. Ricevuta la risposta, lo scraper analizza il documento HTML ottenuto ricercando una sequenza specifica di dati.

3. Lo scraper estrae i dati strutturandoli in un formato specifico a scelta.

Un utilizzo comune di scraper avviene quando Google o un qualsiasi altro motore di ricerca analizza un sito, immagazzina informazioni per poterlo poi indicizzare. Non commette alcun tipo di reato, ma è comunque prevista un'area riservata del sito con dati sensibili degli utenti, i quali sono protetti dagli accordi sulla privacy e che quindi non possono essere letti o copiati in alcun modo. Per esempio, la maggior parte dei servizi di comparazione dei prezzi usa i web scraper per leggere le informazioni sui prezzi di diversi negozi online.
I programmi di scraping utilizzano spesso linguaggi di programmazione come Python o

R per inviare richieste e estrarre i dati. Possono anche utilizzare librerie o framework specifici per lo scraping, come BeautifulSoup o Scrapy, ma che non andremo a trattare. Ho eseguito quindi una ricerca di uno scraper che dato il nome completo contenuto in una colonna di un file XLSX mi fornisse il profilo LinkedIn del soggetto designato.

1.6 Software utilizzati per la gestione delle richieste

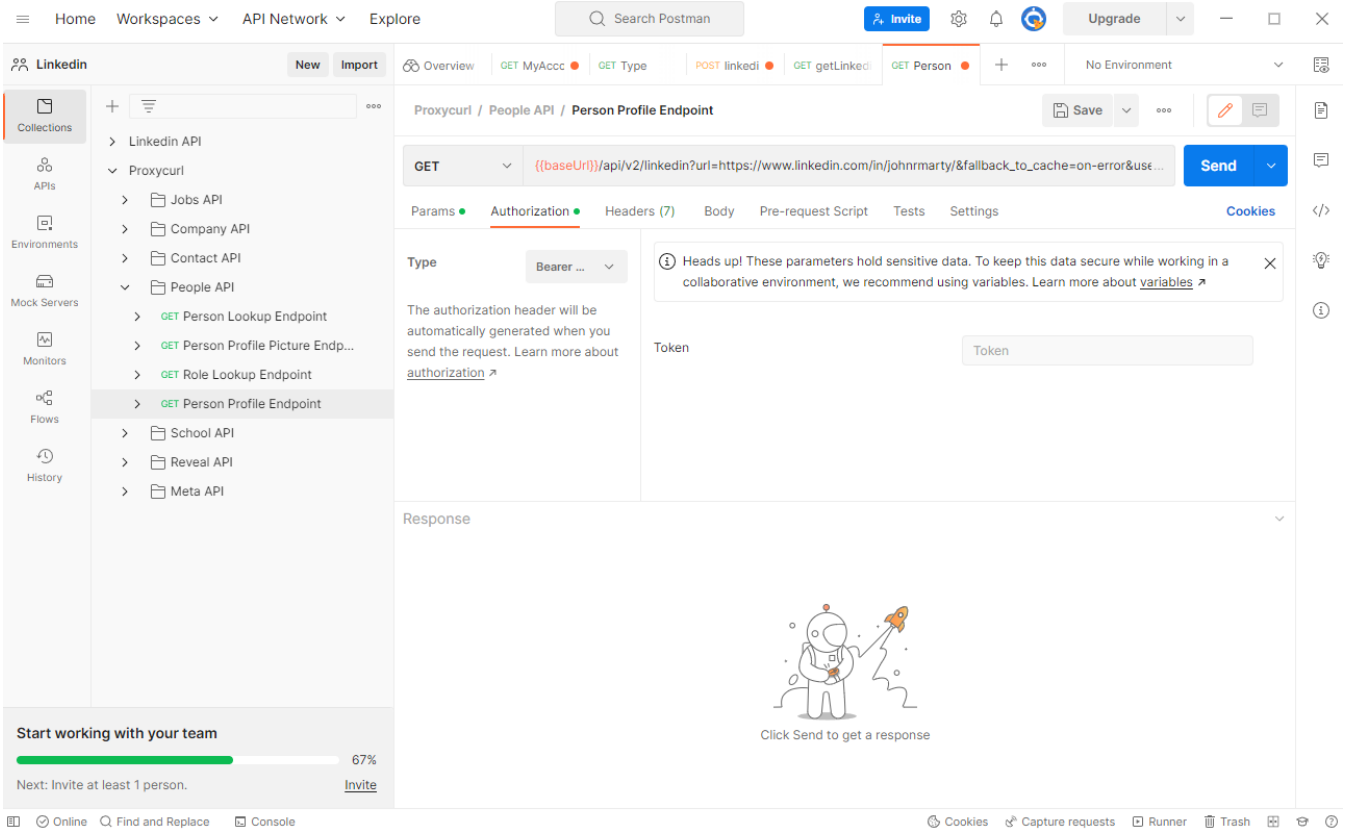
1.6.1 Visual Studio Code

E' l'editor di code sorgente sviluppato da Microsoft. E' l'editor gratuito più popolare tra gli sviluppatori Javascript. È gratuito, open source e disponibile per Windows, macOS e Linux. Visual Studio Code offre funzionalità come l'evidenziazione della sintassi, la completazione automatica del codice, la visualizzazione del debug e l'integrazione con Git. Supporta molte lingue di programmazione e può essere esteso tramite una vasta gamma di plugin sviluppati dalla community. Viene utilizzato da sviluppatori web, desktop e cloud per scrivere, testare e debugguare il codice.
Visual Studio Code verrà utilizzato come editor principale all'interno della ricerca e sarà necessario per interrogare le varie API.

1.6.2 Postman

Postman [5] è una piattaforma API utilizzata dagli sviluppatori per poter creare, progettare e testare le loro API. Con Postman, è possibile creare e inviare richieste HTTP e visualizzare la risposta del server. È possibile creare raccolte di richieste API, testare endpoint e automatizzare test API. Postman offre anche funzionalità come la documentazione integrata delle API, la condivisione e la collaborazione con il team, la gestione dell'autenticazione e la generazione di codice.
Le sue funzionalità consentono quindi di semplificare ogni passaggio per la gestione e monitoraggio delle API, inoltre permette anche di valorizzare la collaborazione tra gli sviluppatori velocizzando il processo di sviluppo.

Figura 1.5: Esempio di collections con Postman



Capitolo 2

API di Linkedin

In questo capitolo verrà trattata l'API di Linkedin ed eseguita una analisi generale della piattaforma. Inoltre verranno anche trattati casi legali passati.
E' necessario capire se questa API possa risultare utile per la ricerca di qualsiasi genere di produzione scientifica o sarà necessario eseguire un'ulteriore ricerca nel tentativo di trovare API non ufficiali che permettano di estrapolare informazioni dall'URL di un profilo pubblico.

2.1 La documentazione

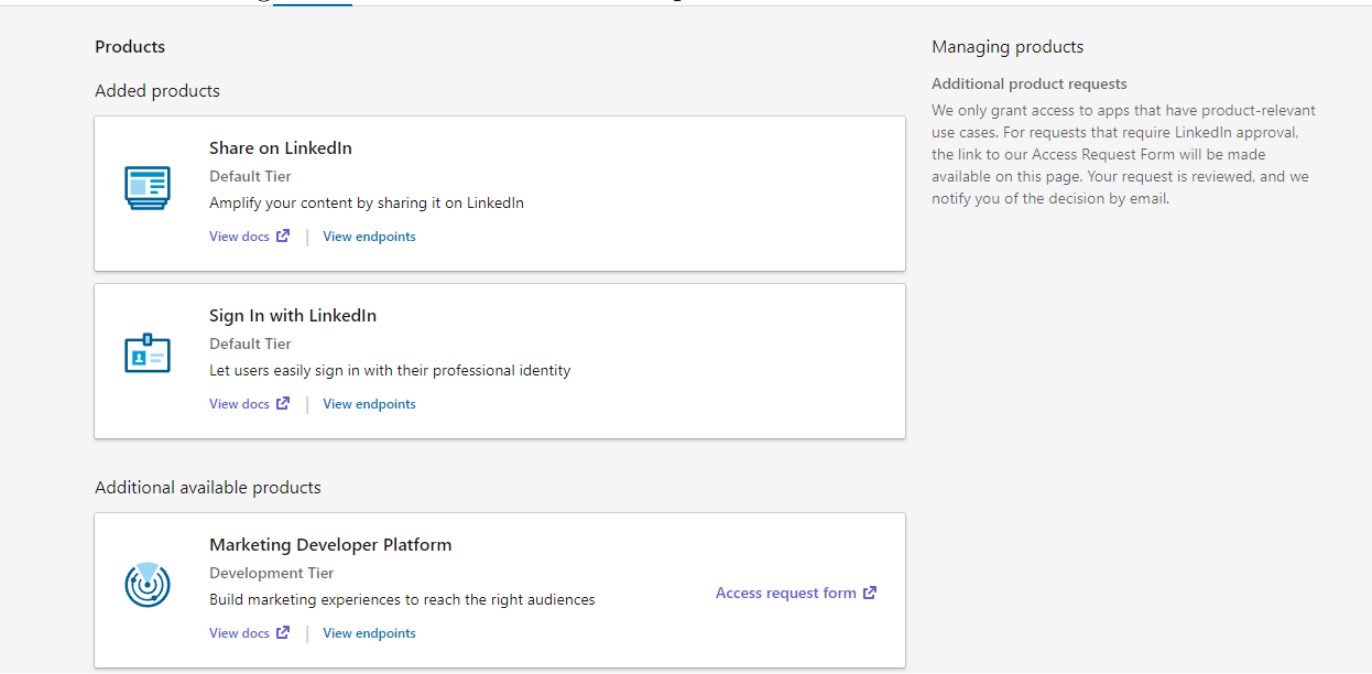
La documentazione dell'API di LinkedIn è una risorsa online che fornisce informazioni dettagliate su come utilizzare le API di LinkedIn per sviluppare integrazioni con la piattaforma. La documentazione include informazioni su come ottenere l'accesso alle API, come autenticare le richieste, come utilizzare le diverse chiamate API e come gestire gli errori. La documentazione descrive i prodotti offerti dall'API:

- Profili dei membri: offre l'accesso a informazioni di profilo dei membri, come il nome, il lavoro attuale, la formazione e le esperienze professionali.
- Post di LinkedIn: offre la possibilità di pubblicare, leggere e commentare i post su LinkedIn.
- Messaggi di LinkedIn: offre l'opportunità di inviare e ricevere messaggi privati su LinkedIn.
- Aziende: offre informazioni sulla presenza online delle società su LinkedIn, come il numero di follower e le statistiche di engagement.
- Gruppi: offre anche un metodo per gestire e partecipare ai gruppi su LinkedIn.

2.2 Autorizzazione

Per poter ottenere l'autorizzazione per l'utilizzo dell'API di LinkedIn è necessario creare un'applicazione all'interno della sezione developers. A questo punto si ha la possibilità di richiedere vari permessi tra cui: condivisione su LinkedIn, accedere su LinkedIn e la sezione di marketing (Marketing Developer Platform) che è quella che effettivamente ci interessa. I primi due hanno un accesso libero, per il terzo invece è necessario compilare un form dove vengono richieste informazioni riguardanti l'azienda che necessita dell'API. Il Marketing Developer Platform (MDP) consente agli utenti di creare e integrare app, widget e servizi di terze parti. Le autorizzazioni offerte possono variare : l'accesso a informazioni di profilo dei membri dell'azienda, l'invio di messaggi ai membri, la lettura e la scrittura di post. Inoltre fornisce la possibilità di ottenere dati biografici, ma rispettando politiche molto severe e limitanti che non permettono quindi la possibilità del suo utilizzo per la raccolta di informazioni di ricercatori e non. Sarebbe stato necessario compilare un form che necessita obbligatoriamente dell'utilizzo di dati aziendali, cosa che sono impossibilitato fornire, in aggiunta è necessario il permesso dell'utente specifico per la richiesta dei suoi dati tramite l'API. Ho deciso per questi motivi di scartare la possibilità di utilizzare l'API ufficiale di LinkedIn.

Figura 2.1: Permessi richiedibili per l'API di LinkedIn



2.3 Termini d'uso

I termini d'uso di LinkedIn [6] stabiliscono le regole e le condizioni che devono essere rispettate dagli utenti della piattaforma. Accettando di sviluppare un'applicazione utilizzando la piattaforma LinkedIn è necessario anche l'approvazione dei suoi termini d'uso. Ci sono numerosi limiti descritti all'interno dei termini. Ad esempio non è possibile eseguire oltre 250000 chiamate giornaliere, non si possono implementare features che possano danneggiare la figura di un membro della piattaforma e in particolar modo è severamente vietato salvare ogni tipo di contenuto. E' possibile memorizzare i dati solamente se c'è un modo per identificarli, separarli e elminarli selettivamente e in ogni caso non possono essere in alcuna maniera ceduti a terzi. In particolare:

- Uso dei servizi: gli utenti sono tenuti a utilizzare LinkedIn in modo legale e rispettoso degli altri utenti. I servizi non sono utilizzabili dagli utenti con meno di 16 anni.

- Contenuti degli utenti: gli utenti sono responsabili del contenuto che pubblicano su LinkedIn. Nel momento in cui un utente posta informazioni, foto, commenti o qualsiasi altro tipo di contenuto su LinkedIn, accetta di rispettare le regole che includono la responsabilità per ciò che si ha pubblicato; devono garantire che non violi i diritti di terze parti o le leggi applicabili. Inoltre devono essere consapevoli delle loro azioni online e dell'impatto che possono avere sul loro profilo o sulla loro reputazione professionale.

- Proprietà intellettuale: LinkedIn protegge i propri diritti di proprietà intellettuale e chiede agli utenti di rispettare i diritti di proprietà intellettuale di terze parti. LinkedIn afferma di rispettare la proprietà intelettuale degli altri richiedendo che anche gli utenti della piattaforma facciano lo stesso. Se un utente pubblica contenuti che violano i diritti d'autore o altri diritti di proprietà intellettuale, il social può prendere misure come la rimozione del contenuto e la sospensione o cancellazione del profilo.

- Privacy: LinkedIn stabilisce le regole per la raccolta, l'utilizzo e la protezione delle informazioni personali degli utenti. La prossima sezione sarà dedicata alla Privacy Policy.

- Modifiche ai termini d'uso: LinkedIn si riserva il diritto di modificare i termini d'uso in qualsiasi momento e senza preavviso. Queste modifiche possono riguardare le regole per la pubblicazione di contenuti, la responsabilità dei contenuti degli utenti, la privacy o la proprietà intellettuale. Gli utenti sono tenuti a rispettare queste modifiche.

- Limitazione di responsabilità: La limitazione di responsabilità descrive tutte le circostanze in cui la piattaforma non è effettivamente responsabile per eventuali danni o perdite subite da parte degli utenti. LinkedIn limita la propria responsabilità per danni diretti e indiretti, inoltre non si ritiene responsabile per le azioni o omissioni degli utenti o di terze parti. Tuttavia, gli utenti dovrebbero essere consapevoli che questa clausola può limitare la loro capacità di ottenere un rimborso o un risarcimento in caso di danni subiti sulla piattaforma.

- Risoluzione delle controversie: eventuali controversie tra LinkedIn e gli utenti saranno risolte attraverso la mediazione o la giurisdizione competente.

2.4 Privacy Policy

La Privacy Policy [7] è un documento, che viene redatto all'interno del sito, che ha lo scopo di informare gli utenti circa il trattamento dei loro dati personali. Se si possiede un sito web bisogna dotarlo di una Privacy Policy, la quale è obbligatoria per la legge anche in caso di tracciamento delle visite per mezzo di strumenti di web analyitics.

- Raccolta di informazioni: LinkedIn raccoglie informazioni personali tra cui nome, indirizzo email, numero di telefono, data di nascita, sesso, informazioni sul posto di lavoro, informazioni sulle connessioni, informazioni sulle preferenze, informazioni tecniche, quali posizione geografica e altri dati che gli utenti forniscono volontariamente. Inoltre, LinkedIn raccoglie informazioni sulle attività degli utenti sulla piattaforma, come le connessioni e le interazioni con i contenuti.

- Uso delle informazioni: LinkedIn utilizza le informazioni personali degli utenti per fornire e migliorare i propri servizi, come la personalizzazione della piattaforma e la fornitura di suggerimenti di contatti e opportunità di lavoro. LinkedIn può anche utilizzare le informazioni per scopi pubblicitari e per condividerle con partner commerciali. I dati possono essere utilizzati inoltre per garantire la sicurezza della piattaforma e proteggere gli utenti da attività dannose, ma anche per condurre ricerche e sviluppare nuovi prodotti o servizi futuri.

- Protezione dei dati: LinkedIn adotta numerose misure tecniche e organizzative per proteggere le informazioni personali degli utenti da perdite, accessi non autorizzati, usi impropri o alterazioni. Tuttavia, LinkedIn non garantisce che i dati degli utenti siano al sicuro da eventuali violazioni della sicurezza. Le tecniche utilizzate possono essere la crittografia, per proteggere i dati trasmessi tra gli utenti, accesso limitato o monitoraggio e rilevamento delle violazioni.

- Condivisione delle informazioni: LinkedIn ha la possibilità di condividere le informazioni personali degli utenti con aziende affiliate, partner commerciali e altri

terzi per scopi pubblicitari o di analisi dei dati. Inoltre, LinkedIn le può condividere informazioni con le autorità per conformarsi alle leggi e alle richieste giudiziarie.

- Diritti degli utenti: gli utenti possono richiedere l'accesso alle loro informazioni personali e la correzione o la cancellazione di esse. In particolare gli utenti hanno i seguenti diritti:
  - Diritto di accesso
  - Diritto di rettifica
  - Diritto alla cancellazione
  - Diritto alla portabilità dei dati
  - Diritto alla limitazione del trattamento
  - Diritto di opposizione

2.5 Data breach

Un data breach, o violazione dei dati, è una situazione in cui i dati sensibili o confidenziali di un'azienda o di un individuo vengono esposti o rubati da parte di hacker o di altri individui non autorizzati. In caso di data breach, le informazioni possono essere utilizzate in modo non autorizzato o venduti a terze parti, creando rischi per la privacy e la sicurezza degli individui interessati. Durante l'estate del 2021 attraverso l'utilizzo delle API è avvenuto un attacco che ha permesso di ottenere oltre il 90% dei dati degli utenti presenti su LinkedIn. Questa enorme quantitità di dati contententi informazioni personali ha quindi permesso ad altri malintenzionati di lanciare attacchi di social engineering a utenti mirati. I dati online sono fondamentali in un mondo sempre più digitale. Possono essere utilizzati per varie ragioni tra cui la personalizzazione dell'esperienza online, miglioramento dei prodotti e dei servizi, ricerca e sviluppo, ma essi possono anche essere adoperati in modo improprio o per scopi dannosi come ad esempio per furti d'identità o per truffe di phishing. L'impatto del data breach subito ha obbligato LinkedIn a migliorare la sicurezza delle proprio API e ridurre effettivamente le informazioni ottenibili da esse.

2.6 Caso HiQ Labs

HiQ Labs [8] è una piccola società che raccoglie e analizza dati pubblici provenienti da diverse fonti per poi fornire informazioni e analisi ai propri clienti. Il caso "HiQ Labs v. LinkedIn" è un caso legale importante che si è svolto negli Stati Uniti riguardante la questione sulla legalità dell'estrazione di dati da un sito web tramite una terza parte. HiQ Labs dopo aver estratto dati pubblici dalla piattaforma di LinkedIn è stata citata

in causa da quest'ultima sostenendo che violasse i termini d'uso e la privacy dei suoi utenti. In primo luogo la corte ha deciso che i dati pubblici estratti attraverso tecniche di scraping non erano effettivamente protetti dalla legge sulla protezione della privacy e che quindi HiQ Labs potesse raccogliere e utilizzare questi dati per le proprie attività commerciali. LinkedIn ha sostenuto che l'estrazione delle informazioni pubbliche potesse violare il diritto d'autore di LinkedIn sui suoi dati, poichè questi dati erano protetti come "opere d'ingegno di carattere creativo", tuttavia anche questa volta la corte ha deciso che l'atto compiuto dalla piccola società non violasse alcun diritto d'autore dato che i dati ottenuti non erano nè originali nè creativi. La sentenza emessa dalla Corte di appello del nono circuito non ha convinto LinkedIn che ha deciso di procedere con il giudizio della Corte Suprema. La più alta corte della magistratura federale degli Stati Uniti d'America ha emesso una decisione il 14 giugno 2021 confermato la sentenza della Corte di appello del nono circuito descrivendo come HiQ Labs avesse la possibilità di continuare con lo scraping sui dati pubblici di LinkedIn; inoltre LinkedIn non poteva impedire ad HiQ Labs di estrarre i dati pubblici attraverso un avviso di blocco automatico. Grazie a questa sentenza chiunque possegga e gestisca dati online è stato obbligato a cambiare le modalità di protezione di cui si era servino fino a quel momento, vista l'impossibilità di impedire l'estrazione di dati pubblici dalle proprie piattaforme. Ha anche influenzato la procedura per la difesa sul diritto alla privacy e d'autore, portando ad una maggiore attenzione alle tecniche per la protezione dei dati personali in favore dei propri utenti.
LinkedIn ha deciso di accettare la decisone della Corte Suprema e conseguentemente ha cercato di rendere i suoi termini di servizio più chiari e stringenti per quanto riguarda l'utilizzo dei dati provenienti dalla piattaforma. Ha deciso inoltre di monitorare maggiormente l'accesso ai dati pubblici sul suo sito web.

2.7 Caso Mantheos

Il primo Febbraio del 2022 LinkedIn ha accusato la società Mantheos Ptd, fondata a Singapore, di eseguire uno scraping non autorizzato di milioni di profili LinkedIn. La tecnica utilizzata dalla società accusata è stata quella di creare centinaia di fake accounts con lo scopo di fare scraping su tutte quelle informazioni che sono solamente accessibili attraverso un account loggato e che quindi non sarebbero ottenibili. In particolare Mantheos dopo aver creato un fake accounts era in grado di ottenere il trials alla Linkedin Sales Navigator, utilizzata dalle aziende che vogliono costruire relazioni con i clienti e per cercare nuovi lead in modo più efficiente, arricchendo nello stesso momento il volume di affari. A questo punto la società di Singapore ha utilizzato l'API ufficiale di LinkedIn per riuscire a mettere le mani sopra le informazioni dei vari profili, cercando di ottenere più dati possibili fino al ban dell'account. Per continuare ad ottenere altre notizie personali degli utenti di LinkedIn basta creare un nuovo account con un nuovo indirizzo IP per evitare di essere immediatamente riconosciuto. Queste azioni hanno violato l'User

Agreement e la Privacy Policy di LinkedIn. In data 6 Maggio 2022, LinkedIn dopo aver vinto la causa ha cercato nuove vie per ridurre la creazione dei fake accounts e reso più complesso l'ottenimento della LinkedIn Sales Navigator. Come parte della risoluzione, Mantheos ha accettato di cancellare per sempre tutti i dati ottenuti attraverso lo scraping illecito, distruggendo inoltre ogni software che permettesse di fare ciò. Durante il processo Mantheos ha definito LinkedIn come una "Gold Mine" di informazioni personali; dopo aver pagato il risarcimento e il costo della causa, ha dovuto inoltre notificare di aver perso la causa a tutti i propri clienti che hanno acquisito o ottenuto l'accesso a questi dati e invitandoli a cancellarli.

2.8 Analisi dei casi

Analizzando i due casi appena descritti sono giunto alla conclusione di come sia attualmente possibile fare scraping di dati pubblici. Il caso HiQ Labs ha permesso di capire che se viene rispettata la Privacy Policy e il diritto d'autore non c'è alcun tipo di problema nell'estrapolazione dei dati pubblici; al contrario, la società Mantheos con il suo utilizzo sconsiderato di fake accounts per poter fare scraping di informazioni che non sono raggiungibili se non dagli utenti iscritti sulla piattaforma, ha messo in evidenza il fatto che non sia in nessun caso possibile ottenere dati privati e protetti dalla Privacy Policy di LinkedIn. Questi due casi devono comunque essere presi con le dovute cautele visto che sono estremamente recenti. Il riproprosi di queste situazioni potrebbe dare origini a sentenze diverse e modificando quindi le decisioni prese in precedenza.

Capitolo 3

Ricerca profili

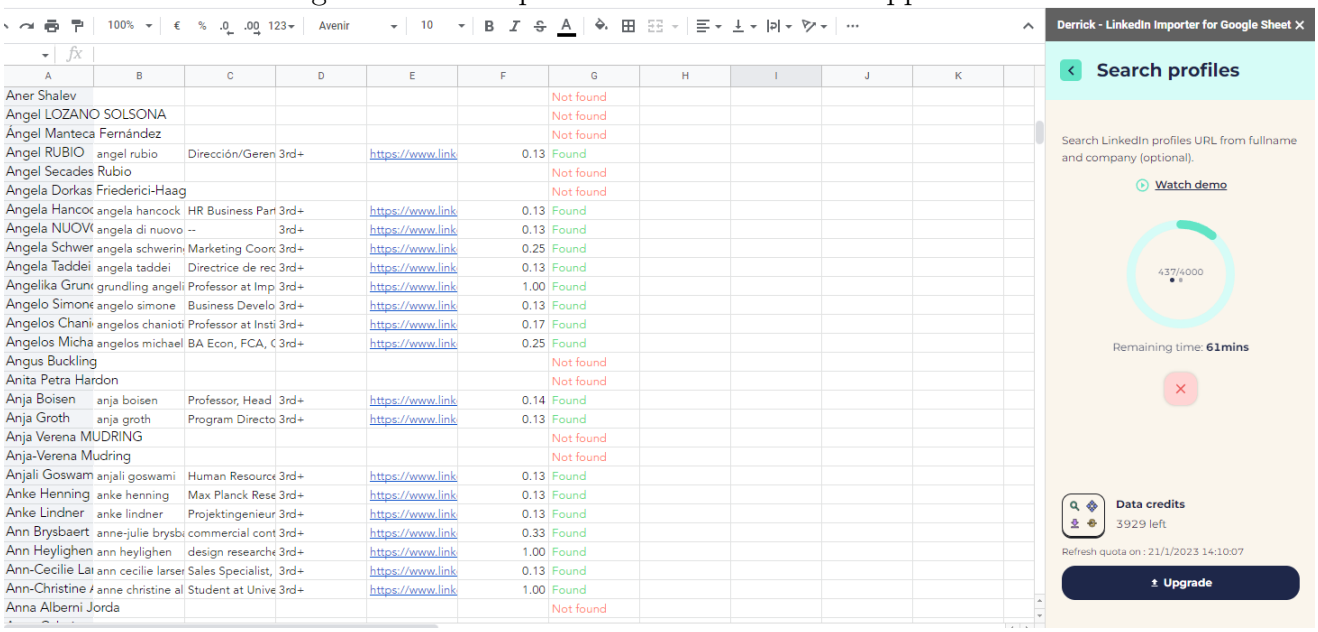
Dai 6671 nomi di ricercatori forniti in input dal progetto RI-PLASMA è necessario ottenere i loro profili LinkedIn. Essendo un lavoro troppo lungo da eseguire manualmente ho deciso di utilizzare uno scraper che, dato un nome e cognome, fosse in grado di fornirmi, se esistente, il profilo relativo. In questo modo ho la possibilità di automatizzare il processo di ricerca risparmiando tempo e permettendo in futuro, nel caso venissero aggiunti ulteriori nomi, di ricercarli in modo automatico.
Verranno analizzati in particolare due scraper: Derrick App, estensione installabile direttamente sul browser, e PhantomBuster, tool che attraverso l'utilizzo di bot estrapola le informazioni richieste. Il loro compito è quello di ottenere l'URL relativo al profilo della persona inserita in input. Con Derrick App non è stato possibile generare il link corretto che mi portasse al profilo pubblico e quindi ne ho ho scartato l'utilizzo. PhantomBuster ha invece raggiunto lo scopo ed è lo scraper che sarà utilizzato per questo step del workflow.

3.1 Derrick App

Derrick App [9] è un'estensione, installabile sul browser Google Chrome, che è in grado di importare all'interno di un Google Sheets tutte le informazioni di cui abbiamo effettivamente bisogno utilizzando solamente una colonna contenente nome e cognome. Derrick App necessita di un cookie di LinkedIn, ovvero di frammenti di dati memorizzati sul computer e utilizzati per migliorare la navigazione. I cookie vengono creati dal server e inviati sul browser. Lo scambio di informazioni consente ai siti di riconoscere il tuo computer e inviargli informazioni personalizzate in base alle tue sessioni. In questo caso Derrick App utilizza il LinkedIn Cookie Session per potere eseguire richieste usando il tuo account di LinkedIn: senza di esso Derrick App non può eseguire alcun tipo di azione. Per aiutare a recuperare più facilemente questo determinato cookie è possibile

utilizzare un ulteriore estensione chiamata Derrick extension che è in grado di prelevarlo in maniera automatica.

Figura 3.1: Esempio di utilizzo di Derrick App

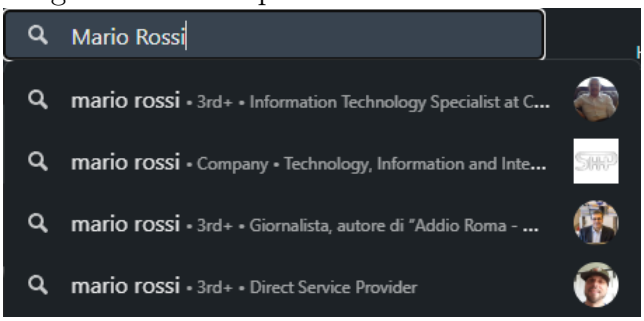


A	B	C	D	E	F	G
Aner Shalev						Not found
Angel LOZANO SOLSONA						Not found
Ángel Manteca Fernández						Not found
Angel RUBIO	angel rubio	Dirección/Geren	3rd+	https://www.link	0.13	Found
Angel Secades Rubio						Not found
Angela Dorkas Friederici-Haag						Not found
Angela Hancoc	angela hancock	HR Business Part	3rd+	https://www.link	0.13	Found
Angela NUOVO	angela di nuovo	--	3rd+	https://www.link	0.13	Found
Angela Schwer	angela schwerin	Marketing Coor	3rd+	https://www.link	0.25	Found
Angela Taddei	angela taddei	Directrice de rec	3rd+	https://www.link	0.13	Found
Angelika Grun	grundling angeli	Professor at Imp	3rd+	https://www.link	1.00	Found
Angelo Simone	angelo simone	Business Develo	3rd+	https://www.link	0.13	Found
Angelos Chani	angelos chanioti	Professor at Insti	3rd+	https://www.link	0.17	Found
Angelos Micha	angelos michael	BA Econ, FCA, C	3rd+	https://www.link	0.25	Found
Angus Buckling						Not found
Anita Petra Hardon						Not found
Anja Boisen	anja boisen	Professor, Head	3rd+	https://www.link	0.14	Found
Anja Groth	anja groth	Program Directo	3rd+	https://www.link	0.13	Found
Anja Verena MUDRING						Not found
Anja-Verena Mudring						Not found
Anjali Goswam	anjali goswami	Human Resource	3rd+	https://www.link	0.13	Found
Anke Henning	anke henning	Max Planck Rese	3rd+	https://www.link	0.13	Found
Anke Lindner	anke lindner	Projektingenieur	3rd+	https://www.link	0.13	Found
Ann Brysbaert	anne-julie brysb	commercial cont	3rd+	https://www.link	0.33	Found
Ann Heylighen	ann heylighen	design researche	3rd+	https://www.link	1.00	Found
Ann-Cecilie La	ann cecilie larse	Sales Specialist,	3rd+	https://www.link	0.13	Found
Ann-Christine A	anne christine al	Student at Unive	3rd+	https://www.link	1.00	Found
Anna Alberni Jorda						Not found

3.1.1 Analisi risultati

E' necessaria eseguire una piccola analisi sui dati ottenuti per verificare quanto sia stato preciso Derrick App, infatti non è possibile accertarsi della completa correttezza dei risulati del software. Innanzitutto sui 6671 nomi sono stati trovati 3507 profili, quindi solamente il 57%. Questo può essere dovuto principalmente da due motivi: il soggetto in questione non ha effettivamente un profilo LinkedIn oppure la nostra applicazione non è stato in grado di trovarlo. Ho deciso quindi di eseguire, utilizzando un campione di 100 nomi che non sono stati trovati, una ricerca manuale per verificare quanti effettivamente non esistessero. Mi sono reso conto che solamente 3 nomi effettivamente avessero un profilo che non è stato trovato, ma tra quelli non trovati ho notato un pattern comune, ovvero che il nome completo contenesse anche il secondo nome di battesimo e una volta rimosso, LinkedIn riusciva ad indirizzarmi al profilo del soggetto. Un altro punto sicuramente interessante è capire come Derrick App fosse in grado di scegliere un profilo dato un nome che ne indirizza a più di uno; ho deciso di eseguire un test come quello precedente tenendo quindi sempre in considerazione 100 campioni (in questo caso nomi e cognomi comuni che nel caso venissero ricercarti porterebbero a più profili). Mi sono reso conto e accertandomene completando il test che l'applicazione ha sempre preso il

Figura 3.2: Esempio di ricerca con omonimi



primo profilo trali omonimi che appaiono nella ricerca. C'è quindi una possibilità che non venga preso il profilo che ci interessa, ma un suo omonimo, e non c'è alcun modo per filtrare la scelta del profilo all'interno delle funzionalità del software. Derrick App mi ha inoltre estratto anche la descrizione dei profili. Il costo di ogni singola richiesta può variare in base al tempo che il software spende per la ricerca, in media si aggira attorno ai 30 crediti per richiesta. Con un pagamento di 9 euro mensili si ottengono 4000 crediti (i quali vengono riaggiornati ogni mese). Il problema principale è dovuto dal fatto che gli URL, creati da Derrick App, fossero effettivamente diversi rispetto a quelli del profilo. L'URL generato mi reinderizzava effettivamente all'utente che ci interessa, ma per visualizzarlo è necessario effettuare il login, come se tutti i ricercatori avessero il profilo privato. A causa di questo motivo è necessario cercare un altro tool più adatto, che sia in grado di trovarmi il link del profilo pubblico. Un esempio di URL creato attraverso Derrick App:

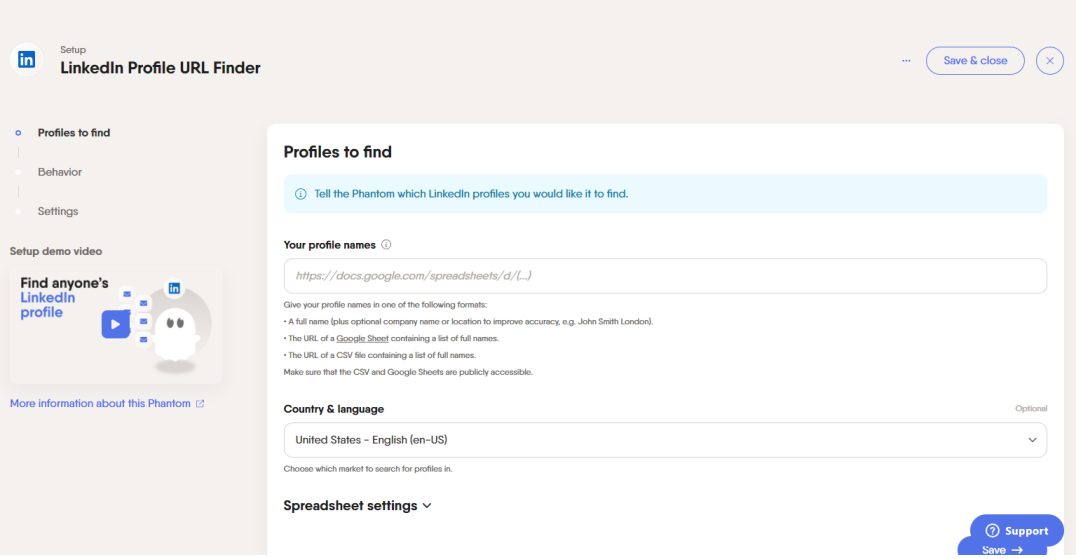
Example 1. https://www.linkedin.com/in/ACoAAAwrHksB2WBzm8pnl0wCCdVa6dEYn0pL51g

3.2 PhantomBuster

PhantomBuster [10] è un tool in grado di svolgere azioni automatizzate di interazione sui social media e di estrazione di dati dal web. Il software si presenta all'utente come un valido sostituto per diversi compiti professionali. Esso contiene oltre 100 "phantom", bot che sono in grado di estrarre dati e/o svolgere altre task. Con PhantomBuster gli utenti possono automatizzare attività come l'estrazione di dati da profili social, l'invio di messaggi automatici e la raccolta di informazioni di contatto. In particolare:

- Estrarre i follower di un determinato account.
- Estrarre i membri di un gruppo social.
- Esportare le conversazioni effettuate in rete.

Figura 3.3: Phantom utilizzato



- Scaricare numerose informazioni come intere liste di risultati di ricerca relativi a profili, aziende, posti di lavoro, ecc..
- Ottenere i dati di base del profilo per ogni utente (nome utente, nome e URL del profilo).
- Estrarre, per ogni URL, i dati di un profilo o di un'azienda, come nome, lavoro, istruzione, dati di contatto (compresi gli indirizzi email), il tutto salvato in un foglio di calcolo pronto per il download.

Verrà utilizzato per l'estrazione di URL fornendo al tool un nome completo . Dopo aver eseguito la registrazione sul sito, riceveremo la possibilità di eseguire un qualsiasi phantom per una durata massima di 2 ore. Utilizzando quindi il phantom LinkedIn Profile URL Finder e passandogli nel parametro richiesto un file XLSX contentente 100 fullnames, è stato in grado di trovare 100 profili, estraendo l'url, la descrizione e il titolo dell'utente. A questo punto PhantomBuster mi ha dato la possibilità di scaricare il file in formato csv:, di seguito è stato convertito in xml ed è pronto per l'utilizzo. PhantomBuster offre un trial plan, concedendo la possibilità di usare un qualsivoglia dei suoi phantom per un tempo massimo di 2 ore. I pacchetti che offre necessitano di un pagamento mensile, il primo parte da 48 dollari, con 20 ore di scraping eseguibili, fino ad un massimo di 320 dollari con circa 300 ore di scraping. Tutti i benefici e le funzionalità ottenute tramite l'acquisto del pacchetto vengono riaggiorante ogni mese. Il phantom LinkedIn Profile URL Finder ha impiegato circa 1 minuto per la ricerca di un singolo URL. In seguito è stato eseguito lo stesso procedimento avvenuto in precedenza

con Derrick App per il caricamento dei dati. In questo caso PhantomBuster mi ha fornito degli URL utilizzabili e che hanno identificato il profilo pubblico degli utenti (ovviamente nel caso lo avessero). In questo caso, a differenza di Derrick App, l'URL ottenuto è corretto ed inoltre è in grado di reindirizzarmi al profilo pubblico

Example 2. https://www.linkedin.com/in/adrian-perrig

Capitolo 4

API alternative a LinkedIn

La ricerca di un API alternativa è necessaria data l'impraticabilità di utilizzare l'API ufficiale a causa dei precedenti motivi elencati. Verranno inizialmente trattati solamente gli aspetti tecnici delle API e a seguire verrà esaminata la questione legale. L'API avrà lo scopo di eseguire l'estrazione dei dati dato l'URL ottenuto attraverso le tecniche di scraping avvenute in precedenza. L'ultima fase, dopo aver correttamente caricato tutti i dati, è quella di eseguire le richieste all'endpoint di un API che permettesse di estrapolare le informazioni dei ricercatori.

4.1 Scrapingbot API

Scrapingbot [11] fornisce un'API che permette l'estrazione di contenuto HTML. E' in grado di localizzare l'informazione all'interno del linguaggio di Markup, estrarre e di conseguenza strutturare i dati. Questa API fornisce dei metodi per recuperare informazioni da diverse fonti sul web e utilizzarle per una vasta gamma di scopi, come analisi dei dati o creazione di report. L'API utilizza il web scraping per estrarre i dati, fornendo un'interfaccia API per accedere ad essi.

4.1.1 Documentazione

Scrapingbot fornisce una Social Media Api da cui è possibile ottenere tutte le informazioni di un determinato profilo di LinkedIn, personale o di un'azienda, attraverso l'utilizzo di un URL passato in richiesta. In particolare nella documentazione è descritto come è necessario eseguire due passaggi per poter ottenere un risultato. Bisogna eseguire una richiesta POST con all'interno i parametri scelti, nel nostro caso l'URL, ottenendo come risposta un codice univoco utilizzabile per la seconda richiesta. Essa sarà una richiesta in GET dove sarà visualizzto il risultato.

4.1.2 Analisi

Dopo aver eseguito vari test utilizzando l'API di Scrapingbot sono giunto alla conclusione che attualmente non funzioni. Mi sono limitato ad eseguire solamente 10 test, a causa della necessità di utilizzare dei crediti per ogni richiesta, e nessuna di esse ha funzionato tranne una. L'unica richiesta funzionante era effettivamente quella usata come esempio nella loro documentazione. Ho deciso quindi di scartare la possibilità di utilizzare questa specifica API. Non sono riuscito a risalire ai motivi del non funzionamento e non c'è alcuna comunicazione, all'interno del sito di Scrapingbot, a riguardo.

4.2 Proxycurl API

L'API di Proxycurl [12] è un set di strumenti progettati per permettere di ottenere dati elaborati. Il compito principale che esegue è di interporsi tra l'applicazione e i dati grezzi in modo che non ci sia necessità di occuparsi dello scraping e/o elaborazione dei dati in scala. L'API è divisa in sottocategorie, analizzerò le principali.

4.2.1 Documentazione

Con L'API di ProxyCurl è possibile:

- Cercare utenti.
- Cercare aziende
- Estraporare informazioni dai profili.
- Estrapolare informazioni dalle aziende.
- Cercare contatti di persone o aziende.

E' possibile eseguire 300 chieste all'API ogni minuto. Viene data in risposta l'error code 429 quando il limite viene superato. Per potere eseguire una richiesta è necessario usare un credito, in caso di successo viene restutuito 200, in caso di errore 404. La velocità di risposta dell'API, secondo la documentazione, si aggira attorno ai 2 secondi.

Jobs API

Contiene all'interno due endpoint.
Job Profile Endpoint permette di ottenere notizie riguardanti il lavoro del profilo passato all'interno della richiesta. Nella risposta è contenuta la descrizione del lavoro, il titolo di impiego, la posizione e il nome della azienda in cui lavora.
Jobs Listing Endpoint ti permette di ottenere la lista delle occupazioni postati da parte di una determinata azienda.

Company API

Company API contiene principalmente tutti i dati e informazioni utili presenti all'interno del profilo dell'azienda interessata. Grazie a questa API è possibile ottenere varie notizie di un'azienda come il nome, la descrizione, il website, la grandezza, il tipo (es: NonProfit, Public Company..), l'anno di fondazione, la posizione, il numero e la lista dei vari dipendenti e anche, in caso, imprese commerciali simili.

Contact API

Contact API include varie informazioni di contatto a partire da un profilo. Vengono strutturati dati quali email e numero di telefoni lavorati.

People API

People API viene utilizzata per visualizzare l'intero profilo. E' possibile ottenere dati personali dell'utente, come la sua occupazione e ruolo, la sua immagine di profilo e tutti i vari social media a cui è collegato. In particolare possiede il Person Profile Endpoint che passando come paramentro l'URL di un profilo è in grado di rispondermi con tutte le informazioni di esso, tra cui tutti gli articoli pubblicati.

4.2.2 Funzionamento e analisi

Proxycurl necessita dei crediti per funzionare, ogni credito permette di eseguire una singola richiesta. Per l'autenticazione è necessario utilizzare il Bearer Token del profilo. Esso è un particolare tipo di Access Token, usato per ottenere l'autorizzazione ad accedere ad una risorsa protetta da un Authorization Server conforme con lo standard OAuth2. Il solo fatto di conoscere e quindi possedere il Bearer Token fornisce la possibilità di accedere ad una determinata risorsa. Ho eseguito delle richieste al Person Profile Endpoint passandogli come parametro l'URL di alcuni profili che avevano effettivamente delle pubblicazioni.

Endpoint 1. GET /proxycurl/api/v2/linkedin

L'unico parametro obbligatorio richiesto è l'URL del profilo che ProxyCurl analizzerà. Sono presenti numerosi parametri non obbligatori:

- Fallback to cache: In caso di errore l'API è in grado di fornirti una risposta in base alle richieste avvenute in precedenza utilizzando quell'URL.
- Use cache: utilizza in default la cache per fornirti una risposta.
- Skills: include nella risposta anche le skills se sono presenti nel profilo ricercato.

- Personal email: include nella risposta anche l'email personale se presente nel profilo ricercato.
- Personal contact number: include nella risposta anche il numero se presente nel profilo ricercato.
- Twitter Id: include nella risposta anche l'id di twitter se presente nel profilo ricercato.
- Extra: include nella risposta anche campi extra specificati dal richiedente.

Nella risposta verranno forniti numerosi campi quali: Nome e cognome, numero di followers, occupazione, paese e città di residenza, esperienze passate, lingua, connessioni, certificati e gli articoli.
Gli articoli sono la parte del profilo interessante allo scopo all'interno di questa tesi. Dopo aver eseguito varie richieste a profili pubblici ho ottenuto sempre risposte quasi immediate. Nella risposta sono contenuti tutti i campi appena elencati ed erano presenti un massimo di 3 articoli in risposta (anche se il profilo ne contenesse effettivamente di più). Questo perchè il massimo di articoli visualizzabili in un profilo pubblico è tre e precisamenti gli ultimi tre inseriti dall'utente in questione. Il risultato è quindi considerato soddisfacente e l'API ProxyCurl può essere benissimo utilizzata per la fase finale del workflow.

4.2.3 Questioni legali

Proxycurl si è impegnato a rispettare gli User Agreement e la Privacy Policy di LinkedIn; in particolare viene descritto all'interno del suo blog, che dopo aver osservato i due casi descritti in precedenza e le relative sentenze, ha valutato che fosse possibile eseguire uno scraping dei soli profili pubblici.
Dal 25 maggio 2018 l'Unione Europea ha iniziato ad applicare il GDPR (Regolamento generale sulla protezione dei dati) nel tentativo di rafforzare la sicurezza e la protezione dei dati personali dei residenti all'interno dell'UE. Vengono previsti diversi requisiti a seconda di come l'azienda interagisce con i dati personali degli utenti. Per dati personali si intendono dati che si riferiscono ad una persona fisica vivente che può essere identificata o attraverso tali dati o da tali dati e da altre informazioni che possono o potranno entrare in possesso. I responsabili del trattamento dei dati sono aziende che forniscono beni o servizi residenti nell'UE, quindi nel caso vengono trattati i dati di residenti dell'UE si viene considerati responsabili del trattamento dei dati e di conseguenza obbligati dai vincoli descritti all'interno del GDPR: uno di questi è quello di lavorare sui dati solo con elaboratori conformi. Proxycurl è considerato un elaboratore di dati e all'interno del suo sito è possibile trovare tutti gli accordi sul trattamento dei dati all'interno del suo DPA (Accordo sul trattamento dei dati). Il DPA è un contratto legale, in questo caso tra

Proxycurl e i suoi incaricati del trattamento. Lo scopo del DPA è quello di stabilire ruoli e obblighi chiari per gli incaricati del trattamento quando trattano dati personali per conto di Proxycurl. Anche se il responsabile del trattamento ha degli obblighi, in ultima analisi è il titolare del trattamento ad essere il responsabile dei dati personali. Proxycurl può solamente garantire in modo sufficiente che il trattamento soddisfi i requisiti del GDPR. Gli elementi da ricercare all'interno del DPA sono:

- Trasferimento e archiviazione dei dati al di fuori dell'UE. Nelle clausole contenute all'interno del DPA di Proxycurl viene descritto come esso si impegna a proteggere qualsiasi dato proveniente dal SEE(spazio economico europeo) in linea con gli standard europei di protezione dei dati

- Misure di sicurezza tecniche e organizzative: Proxycurl adotta un approccio alla sicurezza olistico e basato sul rischio, ovvero la piattaforma limita e protegge l'accesso ai dati e fornisce un monitoraggio degli incidenti possibili.

- Trattamento secondo le istruzioni del responsabile del trattamento: i dati vengono trattati secondo le istruzioni dei clienti

- Notifiche delle violazioni: Proxycurl si impegna di informare tempestivamente qualsiasi trasgressione.

Il problema principale dell'utilizzo di Proxycurl deriva dal fatto che la sua difesa legale non sia effettivamente così solida, ma i due casi visti in precedenza sono degli ottimi indicatori di come sia probabilmente possibile poter estrapolare informazioni pubbliche.

Capitolo 5

Conclusione

Ripercorrendo il workflow il primo passo era l'utilizzo dell'API di LinkedIn. Ciò non è stato possibile a causa delle politiche molto severe e limitanti e in ogni caso l'API di LinkedIn è nata per scopi di marketing e di recruiting. Inoltre sono stato bloccato dalla impossibilità nel reperire i dati necessari a compilare un form che mi desse l'autorizzazione per l'accesso alle varie API fornite dal social media.
Scartata questa opzione ho dovuto proseguire attraverso l'utilizzo dei vari web scrapers. Per primo è stato utilizzato Derrick App il quale, creando in modo erroneo gli URL dei profili, non è stato utile allo scopo perchè gli URL generati non mi permettavano di raggiungere il profilo pubblico del ricercatore in questione.
Ho deciso per questo motivo di provare PhantomBuster. PhantomBuster è riuscito a dare origine URL funzionanti che mi hanno permesso l'accesso ai profili. Risolto questo step ho proseguito con la ricerca di un API che dato l'URL appena ottenuto fosse in grado di estrapolarmi gli articoli postati dal ricercatore evidenziato. Il primo test è avvenuto con l'utilizzo di Scrapingbot, ma che si è da subito rivelato un fallimento. L'API non era in grado di inviarmi le informazioni richieste ed è stato scartato immediatamente.
Ho scelto quindi di utilizzare Proxycurl, il quale ha svolto il suo compito in maniera egregia. Dato il link del profilo è riuscito senza alcun tipo di problema a rimandarmi tutte le ricerche del soggetto in esame.
A causa dei vari costi d'utilizzo delle API ho eseguito l'intero procedimento per un campione di 100 ricercatori, i risultati mi hanno fatto giungere alla conclusione che sia effettivamente possibile estrapolare gli articoli dai loro profili se pubblici e presenti.
Rispetto alle varie questioni legali, all'interno della tesi è stato posto in evidenza il fatto che fosse possibile solamente estrapolare informazioni definite pubbliche, ovvero accessibili anche senza l'iscrizione al social media. Nei due casi principali viene ribadito il fatto che lo sfruttamente di ogni dato privato e di conseguenza protetto dalla Privacy Policy di LinkedIn sia attualmente non utilizzabile. Proxycurl citando in particolare il caso HiQ Labs ha portato avanti la sua difesa riguardante la possibilità di scraping di dati pubblici e attraverso il suo accordo per il trattamento dei dati viene descritto come sia rispettosa

di ogni requisito richiesto dal GDPR. In ogni modo la tutela dell'API utilizzata potrebbe non essere così solida dato che la questione del trattamento dei dati è una materia in continua evoluzione e potrebbero esserci nuove svolte future anche nel breve periodo.

Bibliografia

[1] LinkedIn Corporation. *LinkedIn: The World's Largest Professional Network*. 2002. URL: https://www.linkedin.com/.

[2] United States District Court Northern District of California. *22-651 - LinkedIn Corporation v. Mantheos Pte. Ltd.* Government. Mag. 2022. URL: https://www.govinfo.gov/app/details/USCOURTS-cand-4_22-cv-00651.

[3] Joshua Ofoeda, Richard Boateng e John Effah. «Application programming interface (API) research: A review of the past to inform the future». In: *International Journal of Enterprise Information Systems (IJEIS)* 15.3 (2019), pp. 76–95.

[4] Roy Thomas Fielding. «Architectural Styles and the Design of Network-based Software Architectures». Tesi di dott. University of California, Irvine, 2000. URL: https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm.

[5] Abhinav Asthana. *Postman: The Collaboration Platform for API Development*. 2014. URL: https://www.postman.com/.

[6] LinkedIn Corporation. *LinkedIn: User Agreement*. 2002. URL: https://www.linkedin.com/legal/user-agreement?/.

[7] LinkedIn Corporation. *LinkedIn: Privacy Policy*. 2002. URL: https://www.linkedin.com/legal/privacy-policy.

[8] J Alexander Lawrence e Kristina Ehle. «Combatting unauthorized webscraping-the remaining options in the United States for owners of public websites despite the recent hiQ labs v. LinkedIn decision». In: *Computer Law Review International* 20.6 (2019), pp. 171–174.

[9] *Derrick App: Build Leads Lists in Seconds*. 2022. URL: https://derrick-app.com.

[10] *PhantomBuster: A new era of lead generation*. 2021. URL: https://phantombuster.com/.

[11] *Scrapingbot: Web scraping has never been so easy*. 2021. URL: https://www.scraping-bot.io.

[12] Nubela. *Proxycurl: Pull rich data about people and companies*. URL: https://nubela.co/proxycurl/.

Ringraziamenti

Volevo ringraziare tutti i miei amici che mi hanno supportato e mi hanno aiutato per concludere il percorso di studi. In particolare il mio compagno di progetti Delpo e Xhoi per avermi sempre motivato.