Lead Scraping and Scoring Tool — Report

Objective

The primary goal of this project was to develop a lightweight and intuitive web-based tool that automates LinkedIn lead generation. By leveraging Google Search results through SerpAPI, the tool identifies relevant professional profiles and ranks them based on keyword relevance using a rule-based scoring system. This solution aims to streamline the process of identifying potential candidates, collaborators, or clients without violating LinkedIn's terms of service.

Approach

- **Frontend Framework:** The user interface was developed using **Streamlit**, a Python-based framework that supports quick deployment of interactive web apps with minimal effort.
- **Backend Engine:** The **SerpAPI Google Search API** was used to retrieve search results for LinkedIn profiles based on user-defined queries.
- **Search Input:** The user enters search parameters such as role/skill (e.g., "Data Scientist"), location (e.g., "Pune"), Company (e.g., "Google") and optionally a company name to further refine the search results.
- **Data Flow:** Search results (title, snippet, and URL) are parsed and presented in a clean tabular format alongside dynamically computed lead scores.

Model Selection & Scoring Logic

- The system employs a **rule-based model** rather than machine learning. This was chosen for simplicity, transparency, and speed.
- **Scoring Strategy:** A fixed set of domain-specific keywords (e.g., "AI", "ML", "Recruiter", "Engineer", "Manager", 'Talent Acquisition', 'Recruiter', 'HR', 'Hiring Manager' etc.) is used to score each result.
- Each occurrence of a keyword in the **title** or **snippet** adds **2 points** to the lead score.
- This approach allows real-time ranking of leads based on textual relevance, without needing training data or complex infrastructure.

Data Preprocessing

- **Normalization:** Profile title and snippet are converted to lowercase to ensure uniform keyword matching.
- **Validation:** Results are filtered to ensure only complete and relevant entries (with all necessary fields) are displayed.
- **Ranking:** The output table is sorted in descending order of scores, highlighting the most relevant leads at the top.

Performance & Evaluation

- The rule-based method achieves **fast, consistent performance** suitable for small to medium-scale lead generation tasks.
- It eliminates the need for large annotated datasets while maintaining interpretability.
- **Evaluation:** While no quantitative accuracy metric is used, qualitative assessment shows the tool surfaces relevant LinkedIn profiles effectively.
- Future scope includes integrating NLP-based ranking using **TF-IDF**, **BERT**, or other embeddings for improved contextual understanding and lead relevance.

Summary

This project demonstrates how combining API-based search with keyword-driven scoring can create a practical lead generation tool. Its strengths lie in being **lightweight**, **deployable**, and **customizable**. The use of Streamlit ensures accessibility to non-technical users, while the rule-based system allows easy extension. The tool can be further scaled or adapted for use in recruitment, business development, or market research scenarios.