Job Scraper and Analyzer

A Python-Based Automated Job Data Extraction and Processing Tool

AUTHOR,
Prabanjan S

Abstract:

With the exponential growth of online job postings across platforms like Indeed, the need for automated systems to extract, clean, and analyze job market data has become critical. This project, *Job Scraper and Analyzer*, is a Python-based application that leverages the **Apify API**, **BeautifulSoup**, and **pandas** to dynamically scrape live job data from Indeed. The system retrieves essential details such as job title, company, location, salary, job type, and posting date. Furthermore, it extracts job descriptions, identifies key technical skills (Python, Java, SQL, Excel, AWS, Django, Flask, Machine Learning), and formats the data for meaningful analysis.

The processed data is exported into **Excel (XLSX)** and **CSV** files, enhanced with automatic formatting, headers styling, and column resizing using **openpyxl**. The solution enables job seekers, HR professionals, and data analysts to gain structured insights into the job market while reducing the time and effort required for manual job tracking.

Introduction:

The online job market has rapidly expanded, with platforms like Indeed offering millions of job postings updated in real-time. Manually collecting and analyzing such postings is inefficient, error-prone, and infeasible at scale. Automated job scraping tools address these challenges by programmatically extracting and processing job postings, enabling continuous and structured monitoring of the labor market.

The *Job Scraper and Analyzer* project is built using Python, **Apify Actor API**, and **web scraping libraries**. It retrieves up to 50 job postings for a given search query, cleans and structures the data, detects relevant skills from job descriptions, and saves the results into formatted Excel and CSV files. This system is designed for students, professionals, and recruiters who require quick access to actionable job market insights.

Existing Methods:

Currently, job seekers and recruiters use three main approaches to track and analyze job postings:

1. **Manual Browsing on Job Sites**
    - Users visit platforms like Indeed or LinkedIn and check postings manually.
    - Limitations: time-consuming, error-prone, and difficult to track large numbers of postings over time.
2. **API-Based Job Data Retrieval**
    - Some platforms offer APIs for retrieving job postings.
    - Limitations:
        - Rate limits restrict data collection.
        - Advanced access often requires paid subscriptions.
        - API changes can break dependent systems.
        - APIs may not provide full job descriptions or all metadata.
3. **Third-Party Job Analytics Tools**
    - Tools like Glassdoor Insights or LinkedIn Premium provide analytics dashboards.
    - Limitations:
        - Closed systems with limited customization.
        - Data export options are restricted or premium-only.
        - No direct access to raw job postings for further analysis.

Limitations of Existing Methods:

- No continuous or automated logging of job postings.
- Heavy dependency on APIs or third-party services with restrictions.
- Limited filtering, alerting, and customization options.
- Manual steps required for deeper data processing.

Limitations Of Manual Job Data Collection:

Manual approaches and basic tools have several shortcomings:

1. **Manual Browsing**
    - Very slow and impractical for large-scale monitoring.
    - Human errors in recording details.
    - Cannot track frequent updates or trends.
2. **API-Based Retrieval**
    - Restricted by rate limits and paid access tiers.
    - Susceptible to API endpoint changes.
    - Limited flexibility in extracting extra information like benefits or job type.
3. **Third-Party Platforms**
    - Export and analysis options often restricted.
    - No access to raw description text for skill extraction.
    - Reliance on vendor-specific features.

Overall Gaps:

- Lack of automated logging and historical tracking.
- Minimal customization in filtering and alerts.
- Inability to run independently in the background.

These challenges highlight the need for a self-contained, automated solution like the *Job Scraper and Analyzer*, which can dynamically scrape jobs, clean and store them locally, detect relevant skills, and run without dependency on restricted APIs or third-party platforms.

Proposed Solution:

This project introduces an automated job scraper and analyzer powered by Apify API, BeautifulSoup, and pandas. The system resolves existing limitations by:

1. **Automated Real-Time Scraping**
    - Uses Apify Actor API to scrape job postings from Indeed dynamically.
    - Retrieves job title, company, location, salary, job type, and description.
2. **Skill Detection in Descriptions**
    - Extracts job description text using BeautifulSoup.
    - Detects common technical skills like Python, SQL, Java, Excel, AWS, Django, Flask, and Machine Learning.
3. **Data Cleaning & Transformation**
    - Removes duplicates and trims results to a maximum of 50 jobs per search query.
    - Sorts results by posting date.
4. **Excel and CSV Export**
    - Saves cleaned data to both CSV and Excel formats.
    - Excel files are enhanced with styled headers, column resizing, and alignment for readability.
5. **Automation and Customization**
    - Allows custom job titles to be input by the user.
    - Limits job results while still fetching more for thorough cleaning.

Advantages Over Existing Methods

- No dependency on platform-specific APIs or subscription limits.
- Fully automated scraping and formatting with minimal user input.
- Customizable for different job roles, skills, and filters.
- Cleaned and structured outputs in both Excel and CSV for easy analysis.
- Skill detection adds extra insights beyond raw job postings.
- Local data storage ensures privacy and independence from third-party tools.

The *Job Scraper and Analyzer* thus offers a practical, customizable, and efficient solution for job seekers, recruiters, and analysts who need continuous, structured, and reliable job market insights.

Technologies To Be Used:

1. **Programming Language:** Python
    - Chosen for simplicity, library support, and data processing capabilities.
2. **Libraries and Modules:**
    - **requests** → For interacting with the Apify API.
    - **pandas** → For structured data handling and exporting.
    - **BeautifulSoup (bs4)** → For parsing job descriptions.
    - **openpyxl** → For Excel formatting.
    - **dotenv** → For secure API key handling.
    - **time** → For status polling and waiting.
3. **Data Storage:**
    - CSV and Excel files for easy accessibility and visualization.
4. **External Service:**
    - **Apify API** → Cloud-based scraping infrastructure to extract job listings.

Methods:

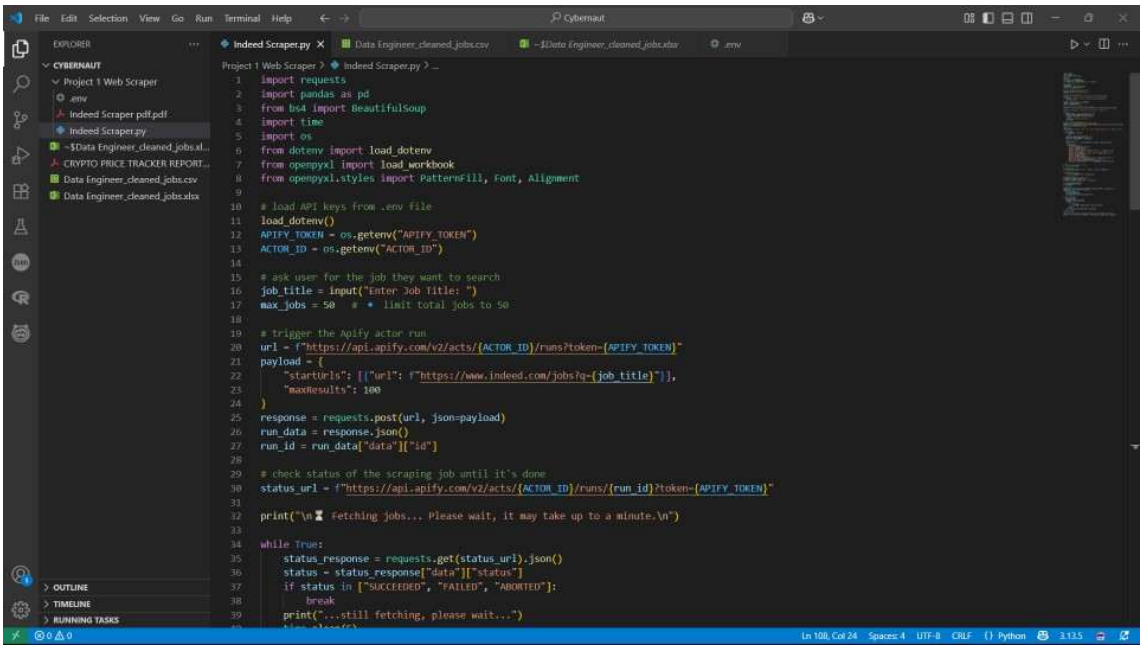
The methodology of the scraper follows these steps:

1. **Initialization**
    - Load API tokens from environment variables.
    - Ask user for job title input.
    - Define maximum job limit (50).
2. **Trigger API Scraper**
    - Send request to Apify Actor with job search query.
    - Monitor run status until completion.
3. **Data Retrieval**
    - Extract dataset ID from Apify.
    - Download job postings as JSON.
4. **Data Cleaning and Skill Detection**
    - Parse job descriptions with BeautifulSoup.
    - Detect skills based on keyword matching.
    - Remove duplicates and limit to 50 records.
5. **Data Transformation**
    - Organize into pandas DataFrame.

- Sort by posting date.

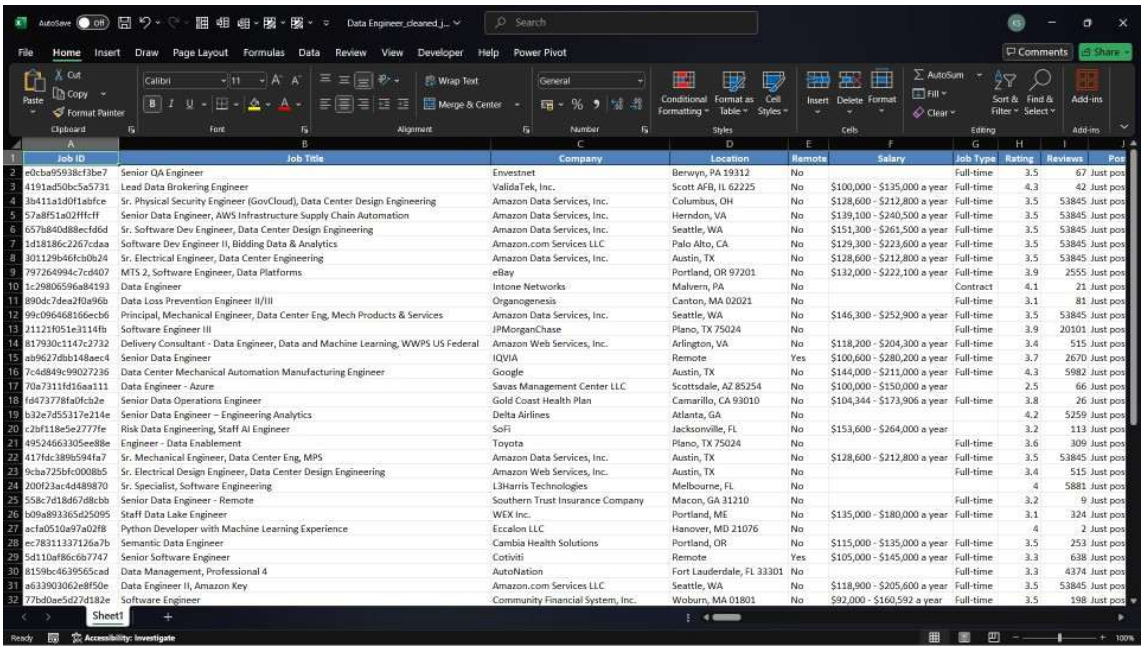
6. **Export to Files**
   - Save to Excel and CSV formats.
   - Apply header styling, column resizing, and alignment in Excel.
7. **Final Output**
   - Provide summary of total jobs collected and number of unique companies.

Screenshots Of Code Snippets:



```python
import requests
import pandas as pd
from bs4 import BeautifulSoup
import time
import os
from dotenv import load_dotenv
from openpyxl import load_workbook
from openpyxl.styles import PatternFill, Font, Alignment

# load API keys from .env file
load_dotenv()
APIFY_TOKEN = os.getenv("APIFY_TOKEN")
ACTOR_ID = os.getenv("ACTOR_ID")

# ask user for the job they want to search
job_title = input("Enter Job Title: ")
max_jobs = 50  # ✨ limit total jobs to 50

# trigger the Apify actor run
url = f"https://api.apify.com/v2/acts/{ACTOR_ID}/runs?token={APIFY_TOKEN}"
payload = {
    "startUrls": [{"url": f"https://www.indeed.com/jobs?q={job_title}"}],
    "maxResults": 100
}
response = requests.post(url, json=payload)
run_data = response.json()
run_id = run_data["data"]["id"]

# check status of the scraping job until it's done
status_url = f"https://api.apify.com/v2/acts/{ACTOR_ID}/runs/{run_id}?token={APIFY_TOKEN}"

print("\n⏳ Fetching jobs... Please wait, it may take up to a minute.\n")

while True:
    status_response = requests.get(status_url).json()
    status = status_response["data"]["status"]
    if status in ["SUCCEEDED", "FAILED", "ABORTED"]:
        break
    print("...still fetching, please wait...")
```

Output:



Job ID	Job Title	Company	Location	Remote	Salary	Job Type	Rating	Reviews	Pos
e0cba95938cf3be7	Senior QA Engineer	Envestnet	Berwyn, PA 19312	No		Full-time	3.5	67	Just pos
4191ad50bc5a5731	Lead Data Brokering Engineer	ValidaTek, Inc.	Scott AFB, IL 62225	No	$100,000 - $135,000 a year	Full-time	4.3	42	Just pos
3b411a1d0f1abfce	Sr. Physical Security Engineer (GovCloud), Data Center Design Engineering	Amazon Data Services, Inc.	Columbus, OH	No	$128,600 - $212,800 a year	Full-time	3.5	53845	Just pos
57a8f51a02fffcff	Senior Data Engineer, AWS Infrastructure Supply Chain Automation	Amazon Data Services, Inc.	Herndon, VA	No	$139,100 - $240,500 a year	Full-time	3.5	53845	Just pos
657b840d88ecfd6d	Sr. Software Dev Engineer, Data Center Design Engineering	Amazon Data Services, Inc.	Seattle, WA	No	$151,300 - $261,500 a year	Full-time	3.5	53845	Just pos
1d18186c2267cdaa	Software Dev Engineer II, Bidding Data & Analytics	Amazon.com Services LLC	Palo Alto, CA	No	$129,300 - $223,600 a year	Full-time	3.5	53845	Just pos
301129b46fcb0b24	Sr. Electrical Engineer, Data Center Engineering	Amazon Data Services, Inc.	Austin, TX	No	$128,600 - $212,800 a year	Full-time	3.5	53845	Just pos
797264994c7cd407	MTS 2, Software Engineer, Data Platforms	eBay	Portland, OR 97201	No	$132,000 - $222,100 a year	Full-time	3.9	2555	Just pos
1c29806596a84193	Data Engineer	Intone Networks	Malvern, PA	No		Contract	4.1	21	Just pos
890dc7dea2f0a96b	Data Loss Prevention Engineer II/III	Organogenesis	Canton, MA 02021	No		Full-time	3.1	81	Just pos
99c096468166becb6	Principal, Mechanical Engineer, Data Center Eng, Mech Products & Services	Amazon Data Services, Inc.	Seattle, WA	No	$146,300 - $252,900 a year	Full-time	3.5	53845	Just pos
21121f051e3114fb	Software Engineer III	JPMorganChase	Plano, TX 75024	No		Full-time	3.9	20101	Just pos
817930c1147c2732	Delivery Consultant - Data Engineer, Data and Machine Learning, WWPS US Federal	Amazon Web Services, Inc.	Arlington, VA	No	$118,200 - $204,300 a year	Full-time	3.4	515	Just pos
ab9627dbb148aec4	Senior Data Engineer	IQVIA	Remote	Yes	$100,600 - $280,200 a year	Full-time	3.7	2670	Just pos
7c4d849c99027236	Data Center Mechanical Automation Manufacturing Engineer	Google	Austin, TX	No	$144,000 - $211,000 a year	Full-time	4.3	5982	Just pos
70a7311fd16aa111	Data Engineer - Azure	Savas Management Center LLC	Scottsdale, AZ 85254	No	$100,000 - $150,000 a year		2.5	66	Just pos
fd473778fa0fcb2e	Senior Data Operations Engineer	Gold Coast Health Plan	Camarillo, CA 93010	No	$104,344 - $173,906 a year	Full-time	3.8	26	Just pos
b32e7d55317e214e	Senior Data Engineer – Engineering Analytics	Delta Airlines	Atlanta, GA	No			4.2	5259	Just pos
c2bf118e5e2777fe	Risk Data Engineering, Staff AI Engineer	SoFi	Jacksonville, FL	No	$153,600 - $264,000 a year		3.2	113	Just pos
49524663305ee88e	Engineer - Data Enablement	Toyota	Plano, TX 75024	No		Full-time	3.6	309	Just pos
417fdc389b594fa7	Sr. Mechanical Engineer, Data Center Eng, MPS	Amazon Data Services, Inc.	Austin, TX	No	$128,600 - $212,800 a year	Full-time	3.5	53845	Just pos
9cba725bfc0008b5	Sr. Electrical Design Engineer, Data Center Design Engineering	Amazon Web Services, Inc.	Austin, TX	No		Full-time	3.4	515	Just pos
200f23ac4d489870	Sr. Specialist, Software Engineering	L3Harris Technologies	Melbourne, FL	No			4	5881	Just pos
558c7d18d67d8cbb	Senior Data Engineer - Remote	Southern Trust Insurance Company	Macon, GA 31210	No		Full-time	3.2	9	Just pos
b09a893365d25095	Staff Data Lake Engineer	WEX Inc.	Portland, ME	No	$135,000 - $180,000 a year	Full-time	3.1	324	Just pos
acfa0510a97a02f8	Python Developer with Machine Learning Experience	Eccalon LLC	Hanover, MD 21076	No			4	2	Just pos
ec78311337126a7b	Semantic Data Engineer	Cambia Health Solutions	Portland, OR	No	$115,000 - $135,000 a year	Full-time	3.5	253	Just pos
5d110af86c6b7747	Senior Software Engineer	Cotiviti	Remote	Yes	$105,000 - $145,000 a year	Full-time	3.3	638	Just pos
8159bc4639565cad	Data Management, Professional 4	AutoNation	Fort Lauderdale, FL 33301	No		Full-time	3.3	4374	Just pos
a633903062e8f50e	Data Engineer II, Amazon Key	Amazon.com Services LLC	Seattle, WA	No	$118,900 - $205,600 a year	Full-time	3.5	53845	Just pos
77bd0ae5d27d182e	Software Engineer	Community Financial System, Inc.	Woburn, MA 01801	No	$92,000 - $160,592 a year	Full-time	3.5	198	Just pos

Future Enhancements:

1. **User Interface (UI)** → Develop a simple desktop or web interface for easier interaction without editing code.
2. **Smarter Skill Extraction** → Use Natural Language Processing (NLP) to detect skills and keywords more accurately.
3. **Dashboards** → Add live charts and insights using Plotly Dash, Power BI, or Grafana.
4. **Database Integration** → Store job data in SQL/NoSQL databases for large-scale analysis and querying.
5. **Alerts & Notifications** → Send email or SMS alerts for jobs that match specific criteria.
6. **Cross-Platform Scraping** → Expand support to LinkedIn, Glassdoor, Naukri, and Monster for broader coverage.
7. **Machine Learning Insights** → Predict job market trends, demand for skills, and salary benchmarks.
8. **Cloud Deployment** → Run the scraper on AWS/Azure/Google Cloud for 24/7 automation.
9. **Mobile App** → Provide instant job insights and alerts on smartphones.
10. **Data Export Options** → Support PDF, Google Sheets, and API endpoints for data sharing and integration.

Conclusion:

The *Job Scraper and Analyzer* successfully demonstrates the power of Python in automating job market research. By combining API-based scraping, HTML parsing, skill detection, and structured data export, the system offers a flexible and reliable way to analyze the job market. It eliminates manual tracking, provides enriched insights with skill detection, and delivers professional-grade outputs in Excel and CSV formats. With future enhancements such as visualization dashboards, NLP-based skill extraction, and database integration, this tool has the potential to evolve into a comprehensive job analytics platform.

References:

1. Indeed – Job Search Platform. Available at: https://www.indeed.com/
2. Apify Documentation – Actor and API Usage. Available at: https://docs.apify.com/
3. BeautifulSoup Documentation. Available at: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
4. pandas – Python Data Analysis Library. Available at: https://pandas.pydata.org/docs/
5. openpyxl Documentation. Available at: https://openpyxl.readthedocs.io/en/stable/