Job Scraper and Analyzer

A Python-Based Automated Job Data Extraction and Processing Tool

Dilli Ganesh B
Team 3
(Team member)
Cybernaut intern

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:

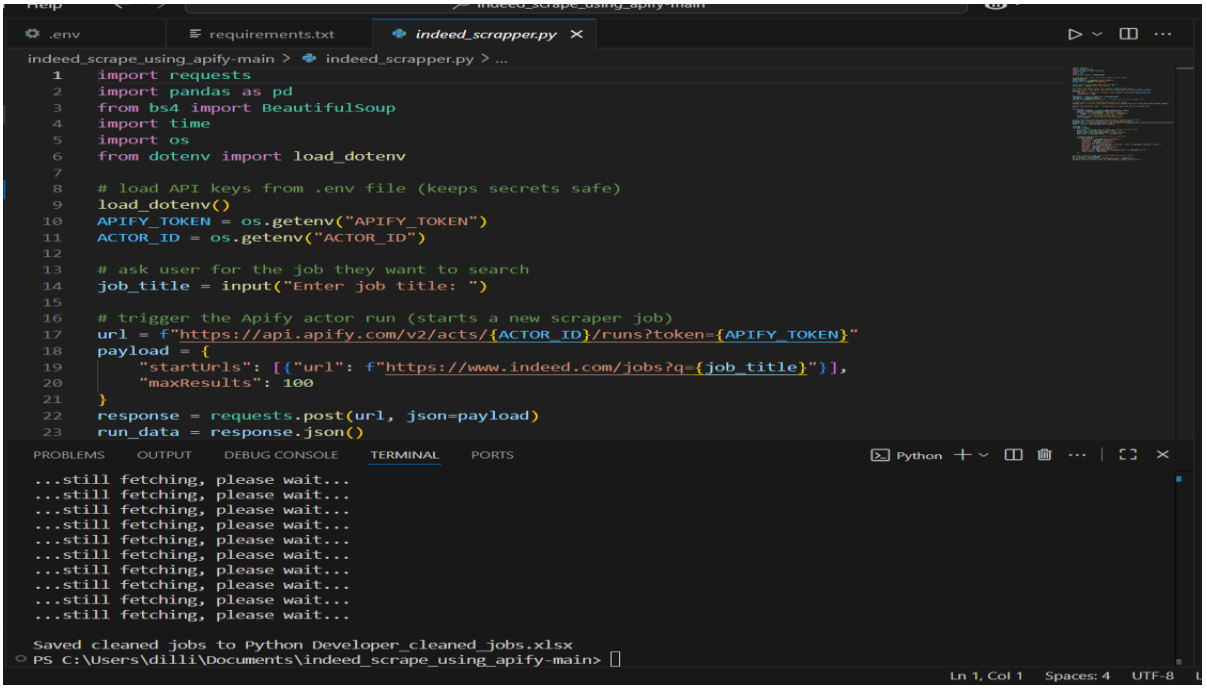
The methodology of the scraper follows these steps:

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.

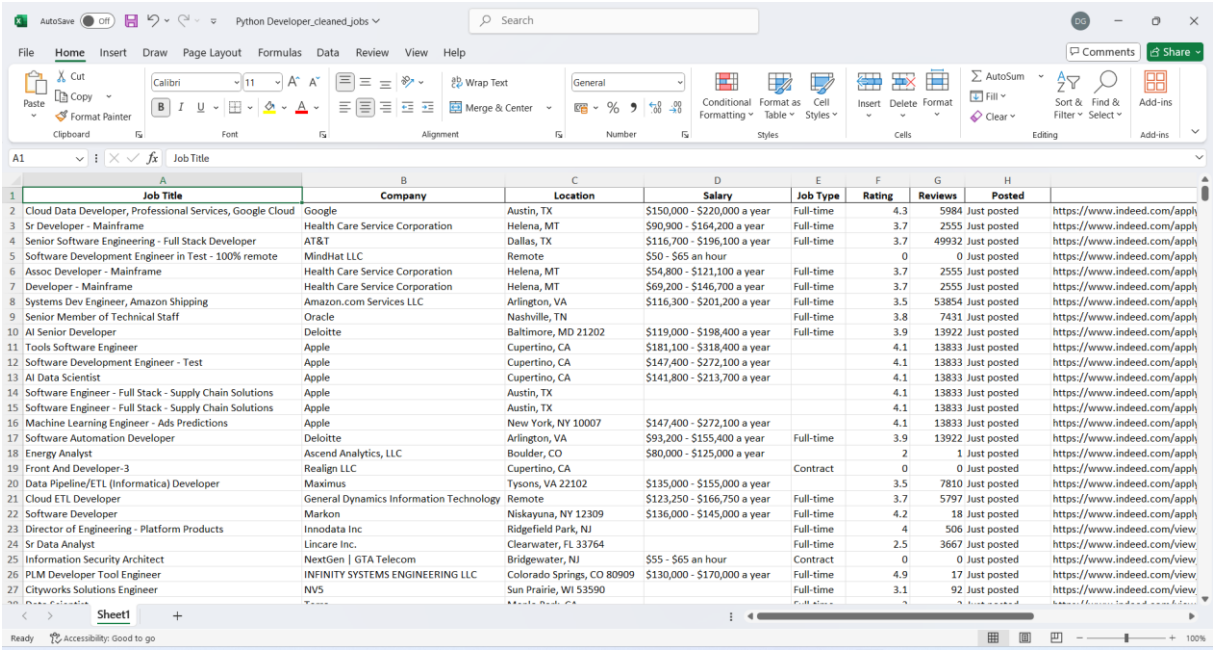
- Sort by posting date.

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:



Output:



Job Title	Company	Location	Salary	Job Type	Rating	Reviews	Posted	
Cloud Data Developer, Professional Services, Google Cloud	Google	Austin, TX	$150,000 - $220,000 a year	Full-time	4.3	5984	Just posted	https://www.indeed.com/appl
Sr Developer - Mainframe	Health Care Service Corporation	Helena, MT	$90,900 - $164,200 a year	Full-time	3.7	2555	Just posted	https://www.indeed.com/appl
Senior Software Engineering - Full Stack Developer	AT&T	Dallas, TX	$116,700 - $196,100 a year	Full-time	3.7	49932	Just posted	https://www.indeed.com/appl
Software Development Engineer in Test - 100% remote	MindHat LLC	Remote	$50 - $65 an hour		0	0	Just posted	https://www.indeed.com/appl
Assoc Developer - Mainframe	Health Care Service Corporation	Helena, MT	$54,800 - $121,100 a year	Full-time	3.7	2555	Just posted	https://www.indeed.com/appl
Developer - Mainframe	Health Care Service Corporation	Helena, MT	$69,200 - $146,700 a year	Full-time	3.7	2555	Just posted	https://www.indeed.com/appl
Systems Dev Engineer, Amazon Shipping	Amazon.com Services LLC	Arlington, VA	$116,300 - $201,200 a year	Full-time	3.5	53854	Just posted	https://www.indeed.com/appl
Senior Member of Technical Staff	Oracle	Nashville, TN		Full-time	3.8	7431	Just posted	https://www.indeed.com/appl
AI Senior Developer	Deloitte	Baltimore, MD 21202	$119,000 - $198,400 a year	Full-time	3.9	13922	Just posted	https://www.indeed.com/appl
Tools Software Engineer	Apple	Cupertino, CA	$181,100 - $318,400 a year		4.1	13833	Just posted	https://www.indeed.com/appl
Software Development Engineer - Test	Apple	Cupertino, CA	$147,400 - $272,100 a year		4.1	13833	Just posted	https://www.indeed.com/appl
AI Data Scientist	Apple	Cupertino, CA	$141,800 - $213,700 a year		4.1	13833	Just posted	https://www.indeed.com/appl
Software Engineer - Full Stack - Supply Chain Solutions	Apple	Austin, TX			4.1	13833	Just posted	https://www.indeed.com/appl
Software Engineer - Full Stack - Supply Chain Solutions	Apple	Austin, TX			4.1	13833	Just posted	https://www.indeed.com/appl
Machine Learning Engineer - Ads Predictions	Apple	New York, NY 10007	$147,400 - $272,100 a year		4.1	13833	Just posted	https://www.indeed.com/appl
Software Automation Developer	Deloitte	Arlington, VA	$93,200 - $155,400 a year	Full-time	3.9	13922	Just posted	https://www.indeed.com/appl
Energy Analyst	Ascend Analytics, LLC	Boulder, CO	$80,000 - $125,000 a year		2	1	Just posted	https://www.indeed.com/appl
Front And Developer-3	Realign LLC	Cupertino, CA		Contract	0	0	Just posted	https://www.indeed.com/appl
Data Pipeline/ETL (Informatica) Developer	Maximus	Tysons, VA 22102	$135,000 - $155,000 a year		3.5	7810	Just posted	https://www.indeed.com/appl
Cloud ETL Developer	General Dynamics Information Technology	Remote	$123,250 - $166,750 a year	Full-time	3.7	5797	Just posted	https://www.indeed.com/appl
Software Developer	Markon	Niskayuna, NY 12309	$136,000 - $145,000 a year	Full-time	4.2	18	Just posted	https://www.indeed.com/appl
Director of Engineering - Platform Products	Innodata Inc	Ridgefield Park, NJ		Full-time	4	506	Just posted	https://www.indeed.com/view
Sr Data Analyst	Lincare Inc.	Clearwater, FL 33764		Full-time	2.5	3667	Just posted	https://www.indeed.com/view
Information Security Architect	NextGen | GTA Telecom	Bridgewater, NJ	$55 - $65 an hour	Contract	0	0	Just posted	https://www.indeed.com/view
PLM Developer Tool Engineer	INFINITY SYSTEMS ENGINEERING LLC	Colorado Springs, CO 80909	$130,000 - $170,000 a year	Full-time	4.9	17	Just posted	https://www.indeed.com/view
Cityworks Solutions Engineer	NV5	Sun Prairie, WI 53590		Full-time	3.1	92	Just posted	https://www.indeed.com/view

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/