M.F.P Assignment – Report

==

Author: Aditya Kayasth

Date: 30-Nov-25

==

Task:

Create Candidate Search API (Skill + Experience Filter)

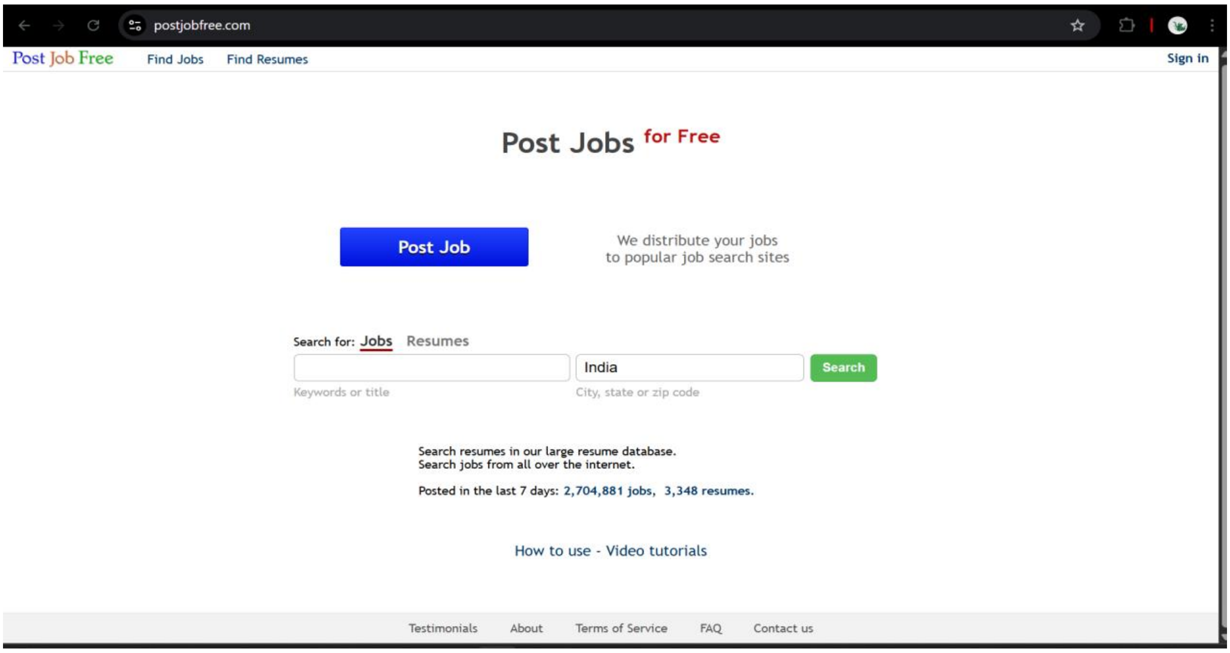
Objective:

Build an API endpoint that accepts skill name and number of years of experience as input parameters and returns a list of matching candidate profiles. The candidate data should be sourced from Naukri.com (via scraping and available APIs) and any other reliable publicly available sources (1 source) that allow candidate data retrieval.

==

Data source used:

1) PostJobFree.com



We distribute your jobs
to popular job search sites

Search resumes in our large resume database.
Search jobs from all over the internet.

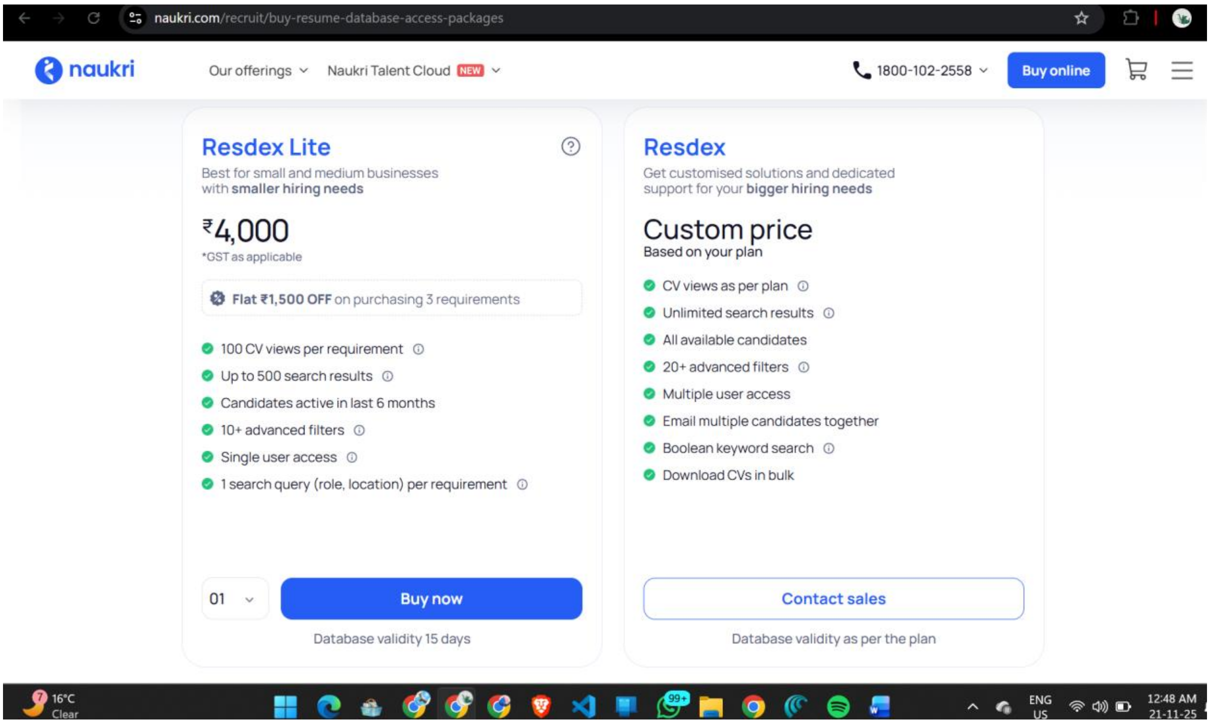
Posted in the last 7 days: 2,704,881 jobs, 3,348 resumes.

Description: It is an open database where candidates upload their resumes specifically to be found by recruiters. They do not use aggressive anti-bot technology. They allow scraping to read their pages using the normal and advanced queries via request module.

Query Structure:

1) Normal: https://www.postjobfree.com/resumes?q={skill}&l={location}
2) Advanced: https://www.postjobfree.com/advanced-resume-search?q=title:({job_title})+({must_have_skills})({excluded_keywords})&l={location}&radius={proximity}&r={limit}

2) Naukri.com (Primary Target):



Resdex Lite

Best for small and medium businesses with **smaller hiring needs**

₹4,000

*GST as applicable

Flat ₹1,500 OFF on purchasing 3 requirements

- 100 CV views per requirement
- Up to 500 search results
- Candidates active in last 6 months
- 10+ advanced filters
- Single user access
- 1 search query (role, location) per requirement

01 | Buy now

Database validity 15 days

Resdex

Get customised solutions and dedicated support for your **bigger hiring needs**

Custom price

Based on your plan

- CV views as per plan
- Unlimited search results
- All available candidates
- 20+ advanced filters
- Multiple user access
- Email multiple candidates together
- Boolean keyword search
- Download CVs in bulk

Contact sales

Database validity as per the plan

Note: Direct scraping of Naukri is strictly blocked as it violates their terms. They monetize this data via "Resdex" (priced at ~₹4,000 for 100 views), making direct unauthorized access difficult and unethical.

===

Challenges:

1) TLS/SSL Fingerprinting:

During testing, Google X-Ray requests were frequently blocked, even when rotating User-Agents. Standard Python requests libraries leak their identity via the TLS Handshake (JA3 Fingerprint). Google detects that the request is coming from a script (OpenSSL) rather than a real browser.

Solution:

Implementing advanced libraries such as **curl_cffi**, **Selenium**, or **Scrapy** allows the scraper to impersonate real browser TLS signatures, significantly reducing detection rates.

2) Data Consistency:

Data returned from public sources like PostJobFree can be inconsistent. Fields like "Years of Experience" vary in format, and sensitive contact information is often redacted in public search views.

Example:

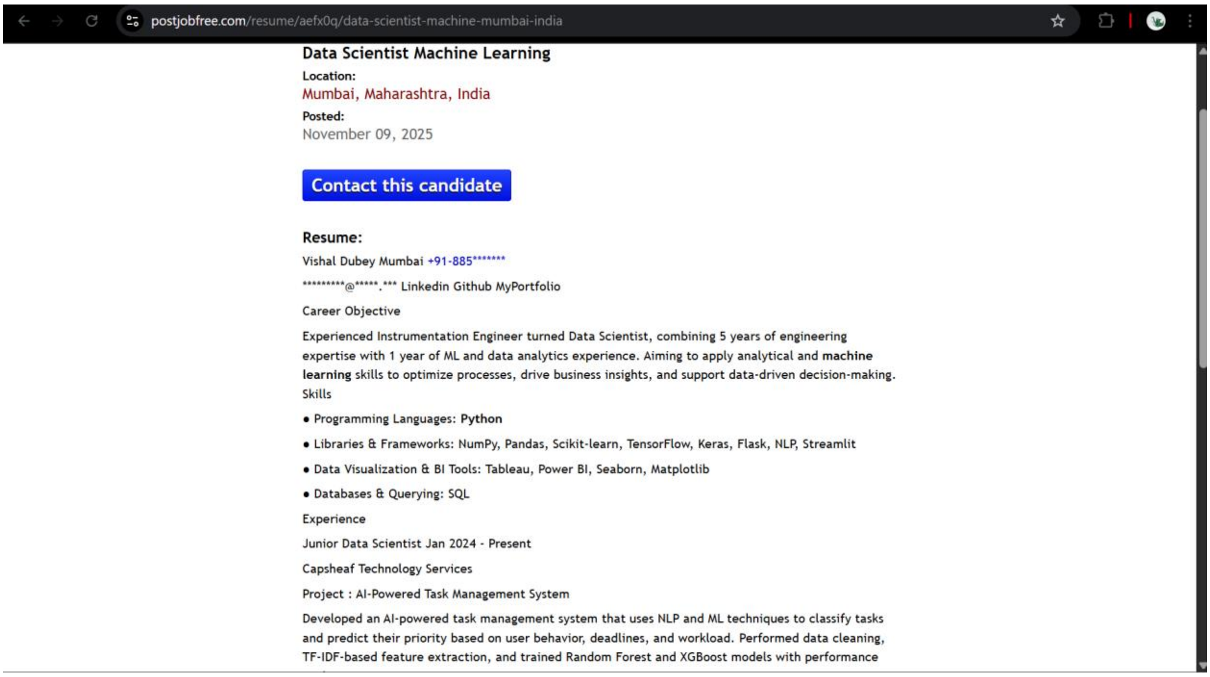
API REQUEST:

```
{
    "skill": "java",
    "experience": 3
}
```

API RESPONSE (Truncated):

```
{
    "status": "success",
    "count": 10,
    "candidates": [
        {
            "source": "PostJobFree",
            "name": "Candidate (Hidden)",
            "current_job_title": "Academic-Driven Assistant Professor with ML & CSE Focus",
      "skills": ["java"],
      "experience_years": "Check Profile",
      "location": "India",
      "resume_url": "https://www.postjobfree.com/resume/aef5wo..."
    }
  ]
}
```

RESUME URL RESULTS:



postjobfree.com/resume/aefx0q/data-scientist-machine-mumbai-india

Data Scientist Machine Learning

Location:
Mumbai, Maharashtra, India

Posted:
November 09, 2025

Resume:

Vishal Dubey Mumbai +91-885*******

*********@*****.*** Linkedin Github MyPortfolio

Career Objective

Experienced Instrumentation Engineer turned Data Scientist, combining 5 years of engineering expertise with 1 year of ML and data analytics experience. Aiming to apply analytical and **machine learning** skills to optimize processes, drive business insights, and support data-driven decision-making.
Skills

- Programming Languages: **Python**
- Libraries & Frameworks: NumPy, Pandas, Scikit-learn, TensorFlow, Keras, Flask, NLP, Streamlit
- Data Visualization & BI Tools: Tableau, Power BI, Seaborn, Matplotlib
- Databases & Querying: SQL

Experience

Junior Data Scientist Jan 2024 - Present

Capsheaf Technology Services

Project : AI-Powered Task Management System

Developed an AI-powered task management system that uses NLP and ML techniques to classify tasks and predict their priority based on user behavior, deadlines, and workload. Performed data cleaning, TF-IDF-based feature extraction, and trained Random Forest and XGBoost models with performance

Solution: I implemented a Regex-based parser to standardize experience data into integers for filtering. To retrieve complete profiles, a secondary scraper that could be deployed to visit individual resume URLs. Integrating a Large Language Model (LLM) would further enhance the extraction of unstructured data from these full resume texts.

===

Improvements planned:

1) **Scraping Efficiency:**

   I plan to fix the blocking issues by upgrading from standard Python requests to curl_cffi. This tool impersonates real browsers, which stops Google and Naukri from blocking the connection. I would also add Proxy Rotation to switch IP addresses automatically, so I can search thousands of times without getting banned.

2) **Data Consistency:**

   I want to fetch better data. Currently, the API retrieves search summaries and no contact details about the candidate can be found at all. I plan to build a scraper that visits every resume link to grab full work history and contact info. To keep this fast, I would use Multithreading to download multiple profiles at the same time, and maybe use an LLM (AI) to clean up the messy text automatically.

===

Gitlab Link: https://gitlab.com/internship-task1/MFP-Task-API-V1/-/tree/main

===