# 🧠 People Vector — Energy & Regulatory Job Trend Scraper

\*Part of the MISO Policy Forecast Project\*

This repository powers the People Vector dataset — a data pipeline that scrapes, cleans, and visualizes public job postings related to energy, utilities, and regulation. The dataset is used to track emerging workforce and policy trends that may precede energy sector policy shifts.

## ⚙️ Quick Setup

1. Clone the repository:  
git clone https://github.com/your-username/miso-policy-forecast-people-vector.git  
cd miso-policy-forecast-people-vector/people\_vector  
  
2. Create & activate a virtual environment:  
python3 -m venv ../.venv  
source ../.venv/bin/activate  
  
3. Install dependencies:  
pip install -r requirements.txt

## 🚀 Running the Pipeline

Run each script in sequence from inside the people\_vector folder:  
python3 scrape\_indeed.py  
python3 scrape\_usajobs.py  
python3 scrape\_greenhouse.py  
python3 clean\_merge.py  
python3 plot\_quick.py  
  
Expected runtime: ~5–7 minutes total. Outputs will appear in data\_clean/.

## 🧾 Example Run Output (October 7, 2025)

|  |  |
| --- | --- |
| Total rows (clean) | 329 |
| Sources | USAJobs: 231 • Indeed: 98 |
| Date range | up to 2025-10-07 |
| Top companies | Veterans Health Administration, Bureau of Prisons, DOE, BPA, CBP |
| Files generated | people\_vector\_clean.csv, people\_vector\_clean.parquet, daily\_counts.csv, daily\_counts.png, top\_keywords.png |
|  |  |

## 🧩 Data Sources

Indeed — Playwright headless Chromium — Public job listings by keyword (“energy”, “regulatory”, “utility”, etc.)  
USAJobs — Official API — Federal openings in DOE, EPA, and related agencies  
Greenhouse — JSON API (optional) — Scrapes energy-sector employers that use Greenhouse boards

## 🗂️ Output Files

data\_clean/people\_vector\_clean.csv — Final deduped dataset (329 rows, filtered by keywords)  
data\_clean/daily\_counts.csv — Time-series of postings per day  
data\_clean/daily\_counts.png — Daily posting trend visualization  
data\_clean/top\_keywords.png — Top detected energy/regulatory keywords  
data\_raw/\*.jsonl — Raw unprocessed scraper outputs

## 🔑 Environment Variables

You’ll need an API key for USAJobs. Get it from https://developer.usajobs.gov and export it as:  
export USAJOBS\_API\_KEY="your\_api\_key\_here"

## 📦 Requirements

Create a file named requirements.txt in the project root with:  
requests  
beautifulsoup4  
pandas  
pyarrow  
python-dateutil  
tqdm  
matplotlib  
playwright  
lxml

## 🧠 Project Context

This scraper supports the MISO Policy Forecast initiative, where multiple “vectors” (Money, People, Paper Trail) feed into a model predicting regulatory and energy policy movements. This folder contains the People Vector — tracking how shifts in job postings reflect early market and policy signals.

## 👥 Contributors

Martin Vizcaíno – Data Scraping Lead

## 🗺️ Folder Structure

people\_vector/  
├── scrape\_indeed.py  
├── scrape\_usajobs.py  
├── scrape\_greenhouse.py  
├── clean\_merge.py  
├── plot\_quick.py  
├── keywords.txt  
│  
├── data\_raw/  
└── data\_clean/