A Lead Generation Tool for Third-Party Data Demand via U.S. Public Job Postings

03 May 2025

Summary

This project is a tool designed to help identify the demand for external data and target potential customers in the public sector using available job posts. The tool analyzes details of U.S. government job postings via USAJobs. First, it creates structured features such as seniority, agency size, and industries. Second, it implements natural language processing (NLP) techniques to analyze the text in the job descriptions to identify positions likely to engage in purchasing external data. Each potential data buyer is assigned a data buyer score, which reflects the likelihood that the position is data-buying; this feature enables easy prioritization of future outreach and targeting.

The framework is fully automated through two modular scripts and a standalone package. The modular scripts enable automatic and real-time lead generation and scoring, allowing demand assessment and trend analysis. The data_buyer_toolkit package, available on PyPI, was developed to integrate the same framework and enhance flexibility and reusability. This package allows users to score specific postings, identify leads by their use case, and map job IDs from titles of interest, providing more control and customization than the end-to-end automated scripts. The package and scripts are easy to reuse, modify to target various industries, and assess the demand in real time, or expand to explore demand seasonality.

Statement of need

Demand for external data in businesses has grown exponentially across industries. Despite this demand, public procurement records are often unavailable or do not identify the roles or departments responsible for the purchasing. This lack of transparency leads to market inefficiencies in which data vendors cannot reliably connect with or identify clients.

This project addresses this inefficiency by creating a reproducible framework to estimate this demand through job descriptions in the available posts. Additionally, it outputs a ranked list of job titles and agencies based on their likelihood of being data buyers, enabling researchers and data vendors to identify likely clients and prioritize their outreach based on these scores.

Acknowledgments

This project was completed during master's degree studies at the University of Pittsburgh. We thank Editage (https://editage.com) for English language editing support and acknowledge the university faculty for their support and instruction throughout the program.

References

Lukauskas, Mantas, Viktorija Šarkauskaitė, Vaida Pilinkienė, Alina Stundziene, Andrius Grybauskas, and Jurgita Bruneckienė. 2023. "Enhancing Skills Demand Understanding Through Job Ad Segmentation Using NLP and Clustering Techniques." *Applied Sciences* 13 (10): 6119. https://doi.org/10.3390/app13106119.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Senger, Elena, Mike Zhang, Rob van der Goot, and Barbara Plank. 2024. "Deep Learning-Based Computational Job Market Analysis: A Survey on Skill Extraction and Classification from Job Postings." https://arxiv.org/abs/2402.05617.

U.S. Office of Personnel Management. 2025. "USAJobs API." https://developer.usajobs.gov/.

Zhang, Yu, Yanyan Zhao, Yangqiu Song, Siheng Liu, and Chao Lin. 2017. "Job Classification Text Corpus (JCTC): A Large-Scale Job Posting Corpus for Text Classification." *IEEE Access* 5: 20020–31. https://doi.org/10.1109/ACCESS.2017.2756278.