Using Similarity Scores to Identify Organizations of Interest by Website

SUMMARY

In 2022, the Chief Evaluation Office (CEO) commissioned Insight Policy Research, which has since been subsumed by Westat, to explore the development of an automated approach for identifying the websites of organizations that DOL may wish to reach for activities such as data collection, outreach, compliance, or enforcement. This approach can be used when the characteristics of interest (such as the type of organization or the types of occupations the organization employs) are not available in other datasets that could help with identification.

This Department of Labor-funded study contributes to the labor evidence-base to inform data, methods, and tools as well as addresses Departmental strategic goals and priorities.

Multiple agencies and programs within DOL may have a need to identify different categories of organizations they work with. For example, they may seek to identify employment service providers, benefits providers, local unions, or even specific types of employers. Such identification can support data collection, outreach, compliance, and enforcement activities. However, characteristics of organizations relevant to the activity are not always available in datasets, making it difficult to identify the organizations needed for contact. An automated approach can make identifying websites of potentially relevant organizations more efficient, while still allowing a human reviewer to make the final decision of whether an organization is relevant for contact.

This brief describes an automated approach using web scraping and natural language processing to identify websites of interest, provides a hypothetical example using the approach, and summarizes the lessons learned in applying this process.

KEY TAKEAWAYS

- The developed approach includes the following steps:
    1. Identify search terms that will be used by a web crawler.
    2. Automate the search process that crawls Google search results.
    3. Process website text to standardize the text for comparison.
    4. Calculate similarity scores by automatically comparing the text from websites identified to the websites of organizations already known to have the characteristics of interest.
    5. Conduct a manual review of the sites sorted by similarity scores (i.e., those most likely to be of interest).
- While this approach is more efficient than a human performing a Google search to identify sites, it could be improved by incorporating scraping subpages in addition to the root pages used in this approach. Additionally, the process will

capture irrelevant websites that use similar terminology, which means human review is a required part of the process.

SEE FULL STUDY

TIMEFRAME: 2022-2024
SUBMITTED BY: Insight Policy Research
DATE PREPARED: December 2024
SPONSOR: Chief Evaluation Office
CEO CONTACT: ChiefEvaluationOffice@dol.gov

The Department of Labor's (DOL) Chief Evaluation Office (CEO) sponsors independent evaluations and research, primarily conducted by external, third-party contractors in accordance with the Department of Labor Evaluation Policy. CEO's research development process includes extensive technical review at the design, data collection and analysis stage, including: external contractor review and OMB review and approval of data collection methods and instruments per the Paperwork Reduction Act (PRA), Institutional Review Board (IRB) review to ensure studies adhere to the highest ethical standards, review by academic peers (e.g., Technical Working Groups), and inputs from relevant DOL agency and program officials and CEO technical staff. Final reports undergo an additional independent expert technical review and a review for Section 508 compliance prior to publication. The resulting reports represent findings from this independent research and do not represent DOL positions or policies.