Project Truth Inquery: Identifying fake abortion clinic websites with Python

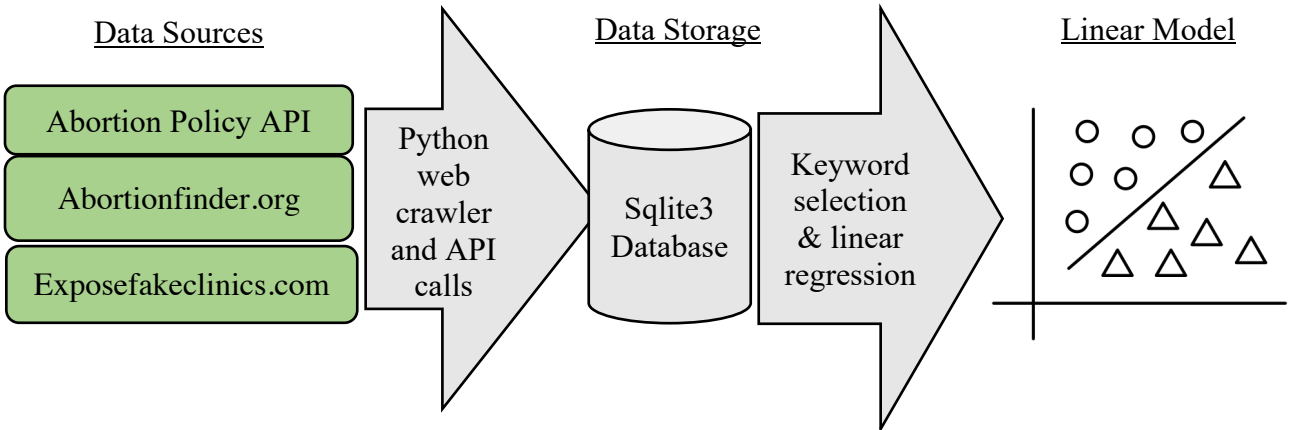
Matt Ryan mattryan@uchicago.edu
Aaron Haefner aaronhaefner@uchicago.edu
Dema Palathingal dema@uchicago.edu

Abstract:

"Crisis pregnancy center" (CPC) websites are often difficult to distinguish from the websites of clinics that provide reproductive healthcare including abortion. This is by design, as attracting and influencing patients who have not yet decided whether to pursue an abortion is a key political goal of the nonprofits that operate and fund CPCs. Identifying CPC websites and maintaining databases of verified abortion-providing clinics has previously been a manual task completed by reproductive rights activists. Even then, the information generated by these activist activities is often not the first search engine result when users query information on abortion access. This project leverages web crawling, database management, and natural language processing in Python to mathematically model and assess the likelihood that current and future clinic websites are real or fake. The current tool takes a clinic URL of unknown status and uses gathered data to assess the utility of a keyword in predicting the likelihood that website is a real or fake abortion clinic. This tool could have utility for individual consumers seeking healthcare, regulatory bodies like the FTC charged with countering fraud, as well as companies seeking to improve search engine results.

Structure:

The project integrates two data sources of interest to reproductive healthcare policy. First, we draw data from the abortion policy API regarding the legal status of provision of reproductive care across the country. We have also integrated data from crawling the websites of both legitimate healthcare provider clinics (HPCs) and "Crisis Pregnancy Centers" (CPCs), or organizations that use deceptive practices to influence the choices of patients seeking reproductive healthcare, historically in ways that are coercive and at times illegal. We access labeled data on the legitimacy of the clinics from abortion access front, one of several advocacy organizations that have sought to maintain verified databases of legitimate clinics that provide reproductive healthcare as well as databases of fake clinics or CPCs. Through the creation of a dataset at the clinic level, we have built a regression model to assess the effect of a keyword on predicting whether a given site is a CPC or HPC. We compare the analysis of tokenized word counts from HCP websites and CPC websites with one expert's view of what words might be used on CPC sites to deceive patients. The current project is in a working state. Future work could expand the dataset that the program and improve the model's accuracy using more labeled data, potentially even moving beyond a linear model to a nonlinear one.



The project is a public repository on github. After reaching out to several advocacy organizations in January 2023 to gain contextual information about HCP and CPC clinics, we received feedback that a tool like this was needed and was something at least one activist organization had been wanting to see developed for years (see email from ineedana submitted with deliverable #2).

Code Responsibilities:

Aaron wrote the web crawler program, graph visualization program and all files within the truth_inquery/crawler module. The output of his code is saved in the truth_inquery/data, truth_inquery/output, and truth_inquery/output_graphs directories.

Dema wrote the api_requests and database program located in the truth_inquery/database_model module.

Matt wrote the lpm and dataframe_cleaner programs located in the truth_inquery/analysis_model module.

Application Guide:

The application can be run completely from the command line. The selenium crawler collects URL-level data on healthcare providers that provide abortion services. Those and the crisis pregnancy center websites are crawled and tokenized. The top 200 tokens are stored in the output directory by state and clinic type.

The application generates SQL .db using these data as well as those queried from the Abortion Policy API. The api_requests.py retrieves data as json files from the API, providing access to key information across states on the following:

1. Gestational limits – i.e.: state restrictions on abortion after certain weeks of pregnancy.

2. Insurance coverage – which is under what types of insurance would cover abortion in the respective state.

3. Minors – how the specific state considers abortions for minors.

4. Waiting periods – how the state considers abortions for minors with respect to waiting periods.

We query both SQL databases, clean and join dataframes in pandas, and add the count of a user-provided keyword for the regression analysis. The mean squared error of the regression is returned to aid in initial comparative analysis between keywords.

Planned accomplishments and current state:

This project began with our interest in the issue of abortion. Once we became aware of the dynamic between crisis pregnancy centers and legitimate healthcare providers, we became interested in how CPCs communicate with the public in ways that obfuscate their identities. Would any methods of how these entities communicated via their websites reflect their differences with legitimate healthcare provider clinics?

To explore that question further, reached out via email to Professor Jenn Holland, an expert on the history of crisis pregnancy centers. She was nice enough to meet with us virtually on Feb. 14th and provided some ideas about how the history and doctrine of CPCs may affect their choices of words or phrases on crisis pregnancy center websites.

After this experience, we pursued a plan that would analyze all words used on CPC and HPC websites, however, data organization of top 200 tokens via web crawling was a choice we made to identify and compare the top tokens and counts across states and between CPCs and HPCs within states. This approach necessarily made comparison across individual words more computationally expensive – as you can see in the dataframe_cleaner program. Ultimately, we settled for a program that can use three policy-related metrics:

- waiting period required by law (hours)
- number of counseling visits required by law
- weeks since last menstrual period (LMP) that abortion was legally permissible until in a given state.

And the frequency count of a single user-provided keyword to run a regression on these variables predictive power of whether a given url was a CPC or an HPC based on the labeled data.

While this is narrower in scope than what we first set out to accomplish, the nature of the program we wrote is expandable well beyond the current analysis, and adding more keywords or adding a more robust NLP module would be a natural extension to the development of the tool.