Memex Human Trafficking Summary

Summary

- Human trafficking is a serious humanitarian problem. Traffickers use physical or financial force to coerce victims into providing labor. Such labor is often advertised online.
- These online advertisements contain crucial information about workers: services offered, location, pay rates, and other attributes. However, the advertisements are often textual, like Craigslist ads. They contain useful data but cannot be queried using SQL databases or statistical tools such as R.
- DeepDive is a tool for structured data extraction that allows us to transform the raw set of advertisements into a single clean structured database table. Analysts can now use standard tools to exploit this information that was previously inaccessible, and thereby identify and help human trafficking victims.

Problem Description

Human trafficking is the use of physical, financial, cultural, or other methods to coerce labor from individuals. Human trafficking is not only a serious humanitarian problem; it is also a serious challenge to law enforcement and military institutions who must grapple with the organizations who engage in human trafficking. If we could use data-centric tools to better identify human trafficking victims and understand the activities of traffickers, we could help people in need as well as help governmental institutions trying to address an ugly and difficult problem.

Dataset Availability

Many trafficking victims are channeled into forced factory labor or sex work. The trafficker benefits by taking some part of the victim's pay for engaging in this work. Since human trafficking is substantially an economic phenomenon, it is believed that traffickers advertise their victims' services, disguised as a standard commercial offering. Of course, like advertising in general, many of these trafficking-driven advertisements are likely to be on the Web.

If we could analyze online advertisements for factory or sex workers, we might be able to help human trafficking victims in several important ways:

- Identifying which advertisements are likely to describe trafficked individuals, as opposed to non-trafficked factory or sex workers
- Using advertisement-embedded data to shed light on the habits and internal structure of traffickers or trafficking organizations
- Helping law enforcement prioritize their scarce resources to focus on helping individuals who are in the greatest danger, or on traffickers who are the most dangerous

Advertisement data is surprisingly informative. Imagine a large set of text advertisements for sex work. These advertisements --- a mixture of structured HTML fields and unstructured text --- are similar in form to advertisements for more conventional goods. An advertisement may contain the name of the worker, contact information if the customer is interested, physical characteristics, and price. Like advertisements on, say, Craigslist, any one data value might be missing. But there are some values that are commonly observed, and which are crucial for a sex worker (whether trafficked or not) interested in making a sale.

The Analytics Opportunity

The set of advertisements for this work is surprisingly large. Our collaborators have obtained more than 30 million advertisements for sex work from online sources. These advertisements contain extremely rich data. In principle, we could do the following with the advertisement data:

- Traffickers may move their victims from place to place in order to keep them more isolated and easier to control. If we detect individuals in the advertisement data who post multiple advertisements from different physical locations, it might be a signal that the individual is being trafficked
- Non-trafficked sex workers show economic rationality: they charge as much as possible for services, and avoid engaging in risky services. Charging non-market rates or engaging in risky services may be a signal the individual is being trafficked.
- Traffickers may have multiple victims simultaneously. If the contact information for multiple workers across multiple advertisements shows contains consecutive phone numbers, it might suggest one individual purchased several phones at one time. The existence of such a person may be a signal that the individual is being trafficked.

These analytics questions embody information about human trafficking, and to someone new in the area they might sound surprising. But they are entirely straightforward from a data management point of view. Given the right dataset, an analyst or data scientist could easily answer these questions using SQL plus Python, or R, or Stata, or any number of other products.

Unfortunately, in their textual form these advertisements cannot be queried using a standard database or analyzed using standard statistical or business intelligence tools. The only common tool for managing unstructured data is a Google-style search engine, which retrieves a document-at-a-time and cannot enable the business-intelligence-style analytics that we require. This is a shame: the advertisement data is a huge opportunity that conventional data management tools cannot exploit.

That is where DeepDive comes in.

Structured Data Extraction with DeepDive

We need to transform the collection of raw advertisement texts into a structured table suitable for analyst processing. Instead of millions of documents, we want a single table with millions of rows. Crucially, the analyst wants that table to contain the following five columns:

- URL where the advertisement was found
- Phone number of the person in the advertisement
- Name of the person in the advertisement
- Location where the person offers services
- Rates for services offered

If we had such a table in a traditional SQL database, the analyst could write the query that would enable the analytics questions listed above.

DeepDive solves this problem. It is a tool for structured data extraction. It transforms the raw text corpus into a high-quality database that downstream tools can process. After DeepDive has populated the table, the analyst can ignore the fact the data was ever derived from documents in the first place. From the analyst's perspective, there is simply a new and remarkable dataset to analyze.