**SHERLOCK: A Tailorable Web Scraper for Government Use**

**DRAFT — PRELIMINARY TASK OUTLINE**

Eduard Hovy

CMU

January 2017

**Goal**

Several Government agencies could benefit from information that is posted on the web, if only they could (1) locate the useful postings and (2) filter out all irrelevant information from them (to enable rapid review and (semi-)automated analysis).

These agencies potentially include ICE (searching for economic fraud, such as online offers for money laundering or false medications for the elderly); TSA and CBP (searching for background information about individuals with undesirable personal profiles who might seek to enter the country); FBI (searching for sale of illegal weapons or drugs); other intel agencies (search for sources of instructions for manufacturing bombs and other weapons); etc.

The vast amount of information online and the difficulty of extracting just useful and relevant aspects prevent such activity.

The goal of this project is to build a demonstration web scraper system, toward a possible eventual scraper system that can be configured by government staff.

**Background: ISI’s WAT Scraper and Other Technology**

Sophisticated web scraping systems exist and are available commercially; see

* <http://inbound.connotate.com/automated-web-crawler-b>
* <http://yipitdata.com/?gclid=CPvJiurKydECFZiCswodDZ8E3w>
* <https://www.networkautomation.com/solutions/automate-data-extraction/>
* <https://scrapy.org/> (a very nice free web-scraper package)
* <http://webscraper.io/> (another free start-up package)

However, they are expensive and not necessarily suited to government purposes.

A web scraper for combating underage sex trafficking was built at USC/ISI in 2007–2012 under funding initially from DHS S&T and then the FBI. This scraper, called WAT (Web Archival Tool), regularly harvested over 10,000 postings a day from over 60 regions across the US, discarded all those not relevant to underage sex trafficking (about 98.7% of it), and formatted and stored the contents of the remainder. The growing database was then queried for analyses of patterns in time, location, and other occurrences, and led (for example) to arrests of pimps and return of minors at the 2014 Superbowl.

The FBI funded a classified rebuild of WAT (by Sotera Defense Inc.; <http://www.soteradefense.com/>) that went operational mid-2016 and is used across the country.

The source code of WAT is freely available online at GitHub, under an Apache license, at

<https://github.com/philpot/trafficcop-wat.git>

This code contains two distinct classifying capabilities: a rule-based keyword matcher and a machine-learning trained classifier. It does not contain an actual crawler, but that is easy to obtain or build. It also does not contain the code for deconstruction and extraction from postings (including phone number deobfuscator, scoring, and database integration), or older code previously used for image duplicate detection, image feature, and a web-based user interface tailored for FBI users. This repository can be cloned by anyone interested.

This is where we would start. We estimate it would take a few hours to download and install the GitHub source code, using a standard data stream/page archive. The code is relatively straightforward Python.

We are fortunate in having obtained the interest of the original WAT designer and builder, Mr. Andrew Philpot. He is willing to serve as external consultant to the project.

**SHERLOCK: Plan**

The vision for SHERLOCK is to be a scraper that can be easily installed and configured by government personnel and tailored as appropriate for their task.

To demonstrate this capability and assess government interest, we propose to build an initial prototype version of SHERLOCK that demonstrates the key capabilities. Demonstration is planned for June 2017.

Our plan is to re-use as much existing code as possible and build a (partly ‘canned’) demonstration that addresses at least two domains. We would employ the open-source Scrapy ecosystem (<https://scrapy.org/>) for generic webpage crawling/scraping, followed by Selenium ([seleniumhq.org](http://seleniumhq.org/)) to perform initial analysis of the page to determine how to process it next, followed then by the reconstructed WAT source code for certain types of extraction. We would need to build the user interaction/parameter setup interface.

This includes the following steps:

* January 2017:
  + Identify relevant personnel
  + Obtain and install WAT code
  + Obtain and install Scrapy and Selenium
* February 2017:
  + Design SHERLOCK architecture and interface
  + Pre-script the demonstration
  + Revive as much WAT code as is feasible for SHERLOCK
  + Build the web crawler using Scrapy
* March — April 2017:
  + Build the early analysis stage in Selenium
  + Deploy and connect to WAT code
  + Build and integrate additional functionality, as needed for demo
  + Implement and integrate the interface
* May 2017:
  + Configure and test demo system
  + Finalize the tailoring interface
  + Search for potential users among government contacts
* June 2017:
  + Finalize demonstration(s) data
  + Perform demonstration(s) to potential DHS clients
  + If interest is found, plan out the next steps

**Personnel and Financials**

A task with a hard deadline cannot be left to a graduate student, because of the demands of classes and course projects. Also, student-produced software is often not up to professional standards, and we wish to avoid having to re-program modules later. Fortunately, there is an abundance of available programming talent at CMU. We suggest employing a short-term programmer with the appropriate web and information extraction and analysis skills for 6 months.

Personnel:

* Project leader: Eduard Hovy, CMU (10% effort)
* Implementer: TBD (100% FTE)
* System co-design and expert advisor: Andrew Philpot (external consultant)

Funds requested, 6-month effort:

* Hovy ($20K) + programmer ($60K) + consultant ($10K) = $90K