Building a Google for the dark web

January 9 2017, by Christian Mattmann

A geographical map depicting hotbeds of dark web activity related to illegal products. Larger circles indicate more activity. Credit: Christian Mattmann, CC BY-SA

In today's data-rich world, companies, governments and individuals want to analyze anything and everything they can get their hands on – and the World Wide Web has loads of information. At present, the most easily indexed material from the web is text. But as much as 89 to 96 percent of the content on the internet is actually something else – images, video, audio, in all thousands of different kinds of nontextual data types.

Further, the vast majority of online content isn't available in a form that's easily indexed by electronic archiving systems like Google's.

Rather, it requires a user to log in, or it is provided dynamically by a program running when a user visits the page. If we're going to catalog online human knowledge, we need to be sure we can get to and recognize all of it, and that we can do so automatically.

How can we teach computers to recognize, index and search all the different types of material that's available online? Thanks to federal efforts in the global fight against human trafficking and weapons dealing, my research forms the basis for a new tool that can help with this effort.

Understanding what's deep

The "deep web" and the "dark web" are often discussed in the context of scary news or films like "Deep Web," in which young and intelligent criminals are getting away with illicit activities such as drug dealing and human trafficking – or even worse. But what do these terms mean?

The "deep web" has existed ever since businesses and organizations, including universities, put large databases online in ways people could not directly view. Rather than allowing anyone to get students' phone numbers and email addresses, for example, many universities require people to log in as members of the campus community before searching online directories for contact information. Online services such as Dropbox and Gmail are publicly accessible and part of the World Wide Web – but indexing a user's files and emails on these sites does require an individual login, which our project does not get involved with.

The "surface web" is the online world we can see – shopping sites, businesses' information pages, news organizations and so on. The "deep web" is closely related, but less visible, to human users and – in some ways more importantly – to search engines exploring the web to catalog it. I tend to describe the "deep web" as those parts of the public internet

that:

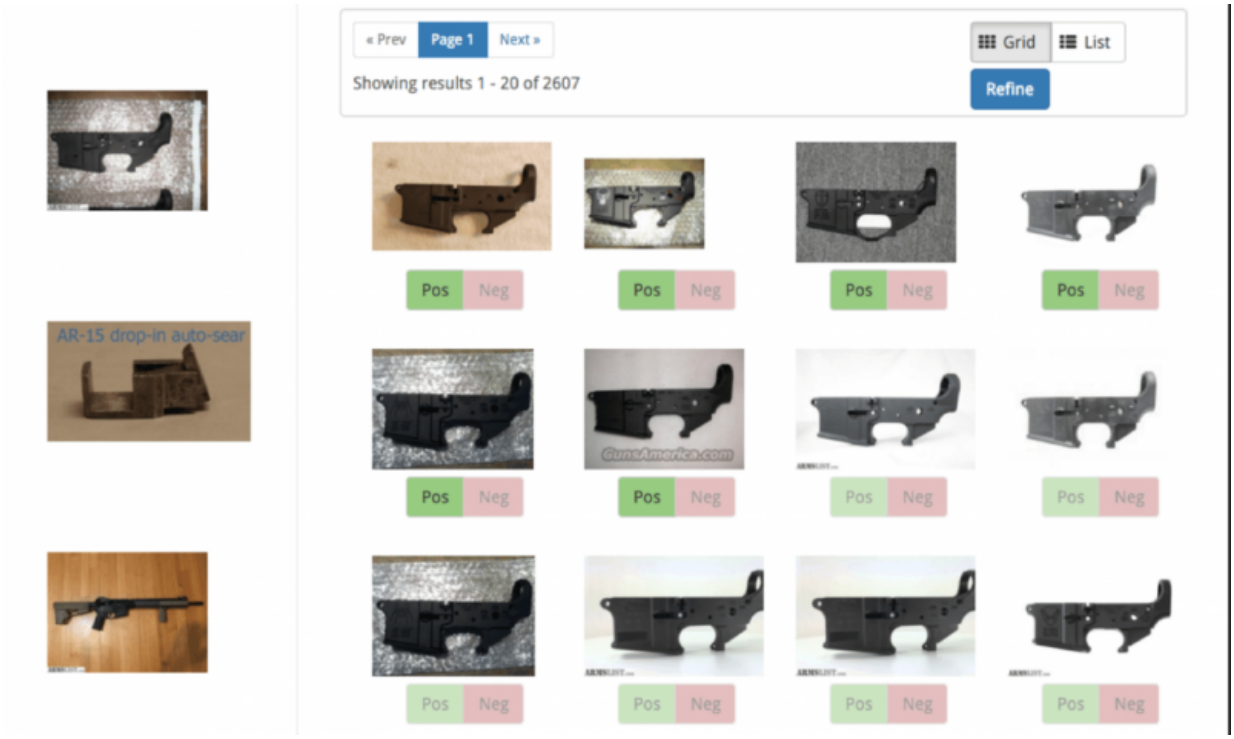
1. Require a user to first fill out a login form,
2. Involve dynamic content like AJAX or Javascript, or
3. Present images, video and other information in ways that aren't typically indexed properly by search services.

What's dark?

The "dark web," by contrast, are pages – some of which may also have "deep web" elements – that are hosted by web servers using the anonymous web protocol called Tor. Originally developed by U.S. Defense Department researchers to secure sensitive information, Tor was released into the public domain in 2004.

Like many secure systems such as the WhatsApp messaging app, its original purpose was for good, but has also been used by criminals hiding behind the system's anonymity. Some people run Tor sites handling illicit activity, such as drug trafficking, weapons and human trafficking and even murder for hire.

The U.S. government has been interested in trying to find ways to use modern information technology and computer science to combat these criminal activities. In 2014, the Defense Advanced Research Projects Agency (more commonly known as DARPA), a part of the Defense Department, launched a program called Memex to fight human trafficking with these tools.



Tika extracting information from images of weapons curated from the deep and dark web. Stolen weapons are classified automatically for further follow-up.

Specifically, Memex wanted to create a search index that would help law enforcement identify human trafficking operations online – in particular by mining the deep and dark web. One of the key systems used by the project's teams of scholars, government workers and industry experts was one I helped develop, called Apache Tika.

The 'digital Babel fish'

Tika is often referred to as the "digital Babel fish," a play on a creature called the "Babel fish" in the "Hitchhiker's Guide to the Galaxy" book series. Once inserted into a person's ear, the Babel fish allowed her to understand any language spoken. Tika lets users understand any file and

the information contained within it.

When Tika examines a file, it automatically identifies what kind of file it is – such as a photo, video or audio. It does this with a curated taxonomy of information about files: their name, their extension, a sort of "digital fingerprint. When it encounters a file whose name ends in ".MP4," for example, Tika assumes it's a video file stored in the MPEG-4 format. By directly analyzing the data in the file, Tika can confirm or refute that assumption – all video, audio, image and other files must begin with specific codes saying what format their data is stored in.

Once a file's type is identified, Tika uses specific tools to extract its content such as Apache PDFBox for PDF files, or Tesseract for capturing text from images. In addition to content, other forensic information or "metadata" is captured including the file's creation date, who edited it last, and what language the file is authored in.

From there, Tika uses advanced techniques like Named Entity Recognition (NER) to further analyze the text. NER identifies proper nouns and sentence structure, and then fits this information to databases of people, places and things, identifying not just whom the text is talking about, but where, and why they are doing it. This technique helped Tika to automatically identify offshore shell corporations (the things); where they were located; and who (people) was storing their money in them as part of the Panama Papers scandal that exposed financial corruption among global political, societal and technical leaders.

Identifying illegal activity

Improvements to Tika during the Memex project made it even better at handling multimedia and other content found on the deep and dark web. Now Tika can process and identify images with common human

trafficking themes. For example, it can automatically process and analyze text in images – a victim alias or an indication about how to contact them – and certain types of image properties – such as camera lighting. In some images and videos, Tika can identify the people, places and things that appear.

Additional software can help Tika find automatic weapons and identify a weapon's serial number. That can help to track down whether it is stolen or not.

Employing Tika to monitor the deep and dark web continuously could help identify human- and weapons-trafficking situations shortly after the photos are posted online. That could stop a crime from occurring and save lives.

Memex is not yet powerful enough to handle all of the content that's out there, nor to comprehensively assist law enforcement, contribute to humanitarian efforts to stop human trafficking and even interact with commercial search engines.

It will take more work, but we're making it easier to achieve those goals. Tika and related software packages are part of an open source software library available on DARPA's Open Catalog to anyone – in law enforcement, the intelligence community or the public at large – who wants to shine a light into the deep and the dark.

This article was originally published on The Conversation. Read the original article.

Provided by The Conversation

https://phys.org/news/2017-01-google-dark-web.html