World's largest PDF archive to aid malware research has been created

June 14 2023

Credit: Pixabay/CC0 Public Domain

As part of DARPA's SafeDocs program, JPL data scientists have amassed 8 million PDFs that can now be used for further study in order to make the internet more secure.

NASA's Jet Propulsion Laboratory is well known for landing rovers on Mars, exploring the solar system with robotic probes, and developing sensitive science instruments that observe Earth and other planets. But less well-known is the lab's cutting-edge work in the digital world.

In support of a wider effort to make the internet more secure, JPL data scientists have created the largest single publicly available open-source archive (corpus) of PDFs. Short for portable document format, a PDF is a complex type of file that looks like a printed document and can contain images, movie files, interactive forms, 3D models, and much more.

The new PDF corpus is part of a Defense Advanced Research Projects Agency (DARPA) program called Safe Documents (SafeDocs) that aims to deal with online threats while anticipating the security needs of PDF users. By working with the nonprofit PDF Association, which seeks to establish open specifications and standards for the technology, JPL is helping to develop several tools to confront these challenges.

When building the corpus, the team didn't evaluate the actual subject matter of the PDFs. Their goal was to gather a large representative sample of PDFs that exist on the internet so experts can search for malicious software that could be hidden in the files' code. That work will then be used to help anticipate emerging online threats and improve PDF technology.

"PDFs are used everywhere and are important for contracts, legal documents, 3D engineering designs, and many other purposes. Unfortunately, they are complex and can be compromised to hide malicious code or render different information for different users in a malicious way," said Tim Allison, a data scientist at JPL in Southern California. "To confront these and other challenges from PDFs, a large sample of real-world PDFs needs to be collected from the internet to create a shared, freely available resource for software experts."

A digital feat

Building the corpus was no easy task. As a starting point, Allison's team used Common Crawl, an open-source public repository of web-crawl data, to identify a wide variety of PDFs to be included in the corpus—files that are publicly available and not behind firewalls or in private networks. Conducted between July and August 2021, the crawl identified roughly 8 million PDFs.

Common Crawl limits downloaded data to 1 megabyte per file, meaning larger files were incomplete. But researchers need the entire PDF, not a truncated version, in order to conduct meaningful research on them. The file-size limit reduced the number of complete, untruncated files extracted directly from Common Crawl to 6 million. To get the other 2 million PDFs and ensure the corpus was complete, the JPL team re-fetched the truncated files using specialized software that downloaded the whole files from the incomplete PDFs' web addresses.

Various metadata, such as the software used to create each PDF, was extracted and is included with the corpus. The JPL team also relied on free, publicly available geolocation software to identify the server location of the source website for each PDF. The complete data set totals about 8 terabytes, making it the largest publicly available corpus of its kind.

The corpus will do more than help researchers identify threats. Privacy researchers, for example, could study these files to determine how file-creation and editing software can be improved to better protect personal information. Software developers could use the files to find bugs in their code and to check if old versions of software are still compatible with newer versions of PDFs.

"This is open and repeatable science. Researchers need to have a common data set to work with so that they can compare results of different analysis techniques and experiments," said Simson Garfinkel, who created a corpus of 1 million files, including thousands of PDFs, called GOVDOCS1 in 2008 when he was an associate professor at the Naval Postgraduate School in Monterey, California. "PDF is one of the most important file types on the internet today, and this contribution of roughly 8 terabytes of data provides faculty, students, and corporations with up-to-date reference data that will power research for years to come."

Provided by JPL/NASA