Quarkslab

Securing every bit of your data

How to download large datasets of files with CommonCrawl

Philippe Lagadec – Pass-The-Salt 2024 – Rump session

Why a large dataset of files

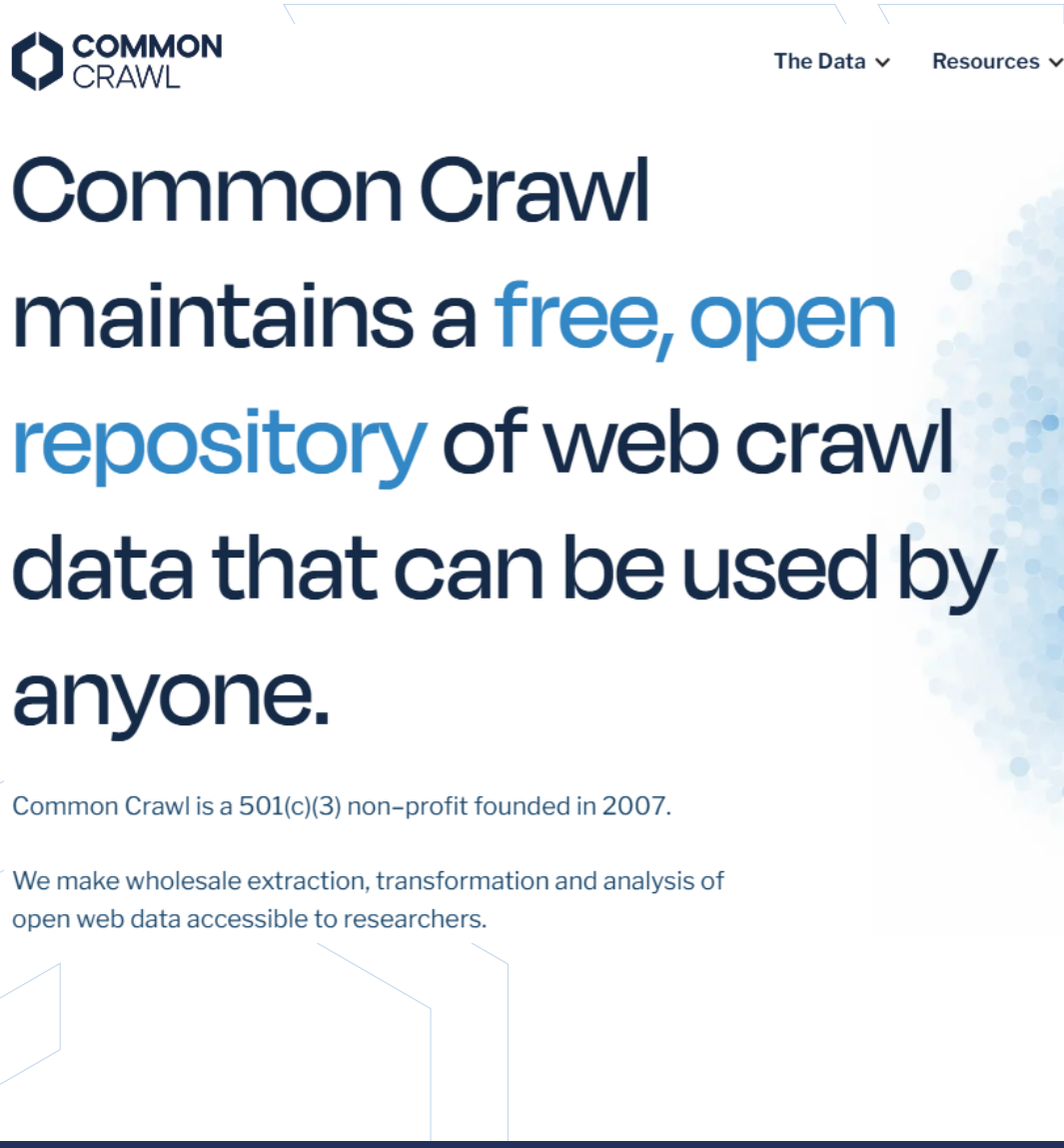
- To test malware detection tools and file scanners/parsers:
  - Detection rate
  - False positives rate
- To test performance of scanning platforms

- **Malware datasets**: already well covered
  - MalwareBazaar, VT, VirusShare
  - Mlget tool

- **Legitimate file datasets**: not so easy
  - Existing public datasets are often old, partial: FUSE (Excel only), ENRON (old)

CommonCrawl

- Huge index of billions of Internet pages and files: https://commoncrawl.org/
- Internet crawled every 2-3 months since 2008
- June 2024 crawl: contains 2.7 billion web pages (or 382 TiB of uncompressed content)
- Every page or file in the index has attributes than can be used for queries:
    - URL
    - Mimetype
    - Content up to 1MB (which is a problem for files)



commoncrawl-fetcher-lite

- https://github.com/tballison/commoncrawl-fetcher-lite
- A simple tool to dowload CommonCrawl indexes, then to look for files with specifc mimetypes, and download them
- How I use it to build a dataset:
    1. Get the list of available mimetypes per CommonCrawl crawls from https://commoncrawl.github.io/cc-crawl-statistics/plots/mimetypes
    2. Use mimetype-detected, more accurate than mimetype reported by web servers
    3. Use the CSV file, complete list of crawls

crawl	mimetype_detected	pages
CC-MAIN-2024-22	text/html	2428084407
CC-MAIN-2024-22	application/xhtml+xml	246672291
CC-MAIN-2024-22	application/pdf	20978860
CC-MAIN-2024-22	text/plain	2892301
CC-MAIN-2024-22	application/atom+xml	2742522
CC-MAIN-2024-22	application/rss+xml	2282800
CC-MAIN-2024-22	application/xml	1578558
CC-MAIN-2024-22	text/calendar	887657
CC-MAIN-2024-22	application/json	684559
CC-MAIN-2024-22	<other>	353916
CC-MAIN-2024-22	application/octet-stream	309689
CC-MAIN-2024-22	application/x-bibtex-text-file	263606
CC-MAIN-2024-22	application/rdf+xml	259530
CC-MAIN-2024-22	text/prs.lines.tag	213997
CC-MAIN-2024-22	application/vnd.google-earth.kml+xml	108563
CC-MAIN-2024-22	text/x-php	93325
CC-MAIN-2024-22	text/csv	91774
CC-MAIN-2024-22	text/x-vcard	88973
CC-MAIN-2024-22	application/pgp-signature	75101
CC-MAIN-2024-22	application/x-tika-ooxml	73528
CC-MAIN-2024-22	application/x-tika-msoffice	63367
CC-MAIN-2024-22	application/vnd.openxmlformats-officed	60427
CC-MAIN-2024-22	application/msword	59996

Config file

```
{   "dryRun" : false,
    "indices": {
     "paths": [
       "crawl-data/CC-MAIN-2018-34/cc-index.paths.gz"
          ]
    },
    "docs": {
          "path": "EXE"
    },
    "recordSelector": {
       "must": {
          "status": [
             {
                "match": "200"
             }
          ]
       },
       "should": {
          "mime_detected": [
             {
                "match": "application/vnd.microsoft.portable-executable",
                "match": "application/x-ms-dos-executable",
                "match": "application/x-dosexec"
             }
```

Post-processing

1. Lots of truncated files of 1 048 578 bytes (CommonCrawl index limitation):
    1. Remove them
    2. Or re-fetch them using curl/wget from the list of URLs
2. Sometimes EXE files are corrupt: extra byte added before MZ => script to clean them up
3. Some files are actual malware => run an antivirus to separate them
4. Some files have a wrong format => check with file/magika to filter out wrong samples
5. Filename is only a hash: rename to add extension based on mimetype

Results after a few hours of downloading:

- 19000+ clean PDF files, 4.8GB
- 5500+ clean EXE files, 2.2GB / filtered out 167 malicious EXE files

Other issues

- Some file formats are rarely published on the Internet as legit files
  - For example LNK files
- Some file formats do not have a specific mimetype:
  - For example recent file formats like MSIX
  - Script files that are difficult to recognize by content are often classified as plain text: PowerShell, BAT, VBScript, …

Any questions:

@decalage2 on X

@decalage@mastodon.social