

O dataset escolhido foi: Evasive-PDFMal2022

Descrição: <https://www.unb.ca/cic/datasets/pdfmal-2022.html>

Referência: Maryam Issakhani, Princy Victor, Ali Tekeoglu, and Arash Habibi Lashkari1, "PDF Malware Detection Based on Stacking Learning", The International Conference on Information Systems Security and Privacy, February 2022

Link do artigo: <https://www.scitepress.org/PublicationsDetail.aspx?ID=VibgIHYeOxw=&t=1>

=====

Arquivos deste diretório:

- dataset_info.pdf: este arquivo;
- PDFMalware2022.csv: arquivo contendo o nome dos arquivos, atributos e classe atribuída a cada arquivo.

Observação: O diretório com os arquivos em si, os PDFs maliciosos e benignos, não estão aqui mas podem ser encontrados na referência. (Acesso em março de 2022)

=====

Descrição do dataset (fonte: descrição e artigo do dataset):

O dataset escolhido contém 10,025 amostras de PDFs, 5557 maliciosas e 4468 benignas, que fogem do padrão da classe semelhante a eles. Ou seja, amostras maliciosas com atributos semelhantes à benignas e benignas com atributos semelhantes à maliciosas.

O arquivo PDFMalware2022.csv (obtido pelo dataset) contém uma lista dos arquivos de amostra, um valor para cada um dos 31 atributos (presentes no csv) e uma classificação baseada em conhecimento prévio sobre a fonte das amostras e na execução de um algoritmo de clusterização. Resumidamente, as amostras são compostas por testes que apresentaram resultado "falso positivo" e "falso negativo" na clusterização.

Classes:

- Malicious: Amostras maliciosas (confirmada pela origem), que na classificação foram relacionadas a amostras benignas;
- Benign: Amostras benignas (confirmada pela origem), que na classificação foram relacionadas a amostras maliciosas.

Atributos:

- pdfsize: Tamanho do arquivo;
- metadata size: Tamanho da região de metadata;
- pages: Diferença com o atributo pageno não é clara;
- xref Length: Número de Xrefs;
- title characters: Número de caracteres no título;
- isEncrypted: Diferença com o atributo encrypt não é clara;
- embedded files: Diferença com o atributo EmbeddedFile não é clara;
- images: Número de imagens;
- text: Presença de texto ou não;
- header: Tipo do cabeçalho;

- obj: Número de palavras-chave indicando o início de objetos;
- endobj: Número de palavras-chave indicando o fim de objetos;
- stream: Número de palavras-chave "streams" (sequências de dados binários);
- endstream: Número de palavras-chave que indicam o fim de streams, "endstreams";
- xref: Número de palavras-chave "/Xref";
- trailer: Número de palavras-chave "/Trailer";
- startxref: Número de palavras-chave "/Startxref";
- **pageno: Diferença com o atributo pages não é clara;**
- **encrypt: Diferença com o atributo isEncrypted não é clara;**
- ObjStm: Número de objetos stream;
- JS: Número de palavras-chave "/JS";
- Javascript: Número de palavras-chave "/JavaScript";
- AA: Número de palavras-chave "/AA";
- OpenAction: Número de palavras-chave "/OpenAction";
- Acroform: Número de palavras-chave "/Acroform";
- JBIG2Decode: Número de palavras-chave "/JBIG2Decode";
- RichMedia: Número de palavras-chave "/Richmedia";
- launch: Número de palavras-chave "/launch";
- **EmbeddedFile: Diferença com o atributo embedded files não é clara;**
- XFA: Número de palavras-chave "/XFA";
- Colors: Número de palavras-chave "/Colors";