CI1030 - Ciência de Dados para Segurança

Exploração e classificação de um dataset de URLs maliciosas

Aluno: Christian Debovi Paim Oliveira (GRR20186713)
Professor: André Gregio

Dataset - Pesquisa

- **Paper:** Detecting Malicious URLs Using Lexical Analysis.
- **Autores:**
  - Mohammad Saiful Islam Mamun
  - Mohammad Ahmad Rathore
  - Arash Habibi Lashkari
  - Natalia Stakhanova
  - Ali A. Ghorbani
- **Universidade:**
  - University of New Brunswick , Fredericton, NB, Canada
- **Objetivo:**
  - Avaliar o uso de características léxicas para a classificação de URLs
  - Avaliar uso de técnicas de ofuscação nas urls coletadas

Datasets – URLs

Arquivos texto com urls:

- Total de 165.366 URLs

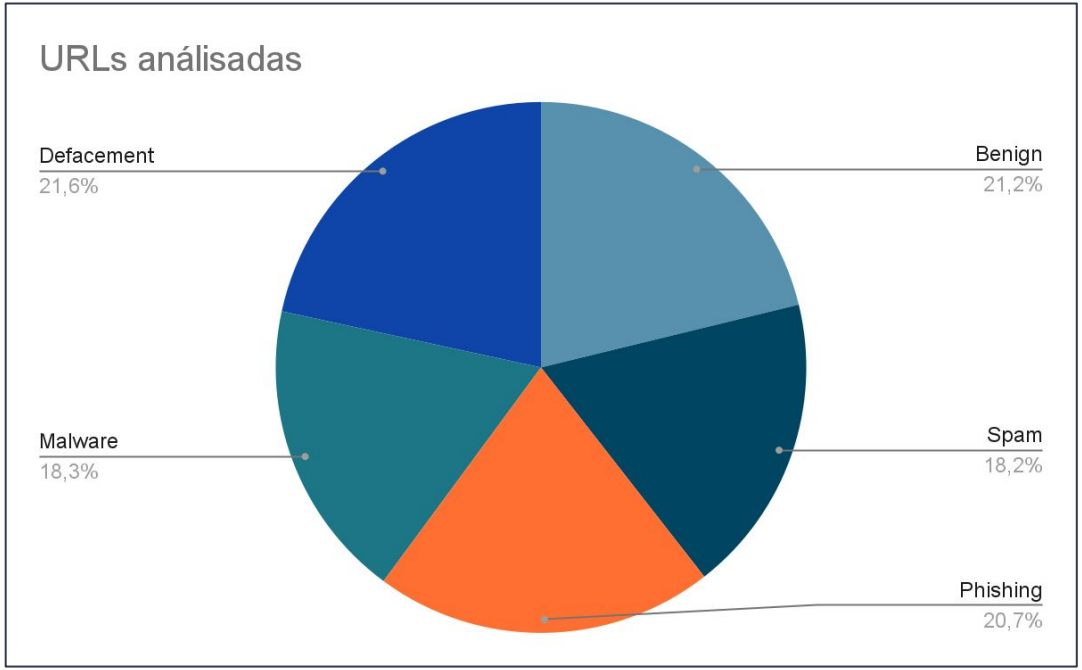
Labels/Classes:

- **Benign:** Alexa top sites + crawler + VirusTotal (35.378 URLs).
- **Spam:** dataset WEBSPAM-UK2007 (12.000 URLs).
- **Phishing:** repositório OpenPhish (9965 URLs).
- **Malware:** lista DNS-BH (11.566 URLs).
- **Defacement:** em Alexa top sites (96.457 URLs).

Dataset – Informações de URLs

CSVs com informações léxicas:

- Extraídas de 36.707 URLs do total
- 79 atributos léxicos + label



Dataset - Atributos

Tipos:

- **Entropy:** variação nos tokens em certas partes da url.
    - Entropy_Domain, Entropy_Extension

- **CharacterContinuityRate:** soma da continuidade dos tokens divididos pelo tamanho da URL.
    - **Ex:** abc567ti = (3 + 3 + 1)/9 = 0.77

- **Ratios:** número de tokens de uma parte da URL divididos pelo número de tokens da outra.
    - argPathRatio, argUrlRatio, argDomainRatio, domainUrlRatio, pathUrlRatio, PathDomainRatio.

- **NumberRate:** proporção de dígitos nas partes da URL.
    - NumberRate_Domain, NumberRate_DirectoryName, NumberRate_FileName, NumberRate_URL, NumberRate_AfterPath.

- Outros atributos relacionados ao tamanho e contagem de tokens, dígitos e símbolos de diferentes partes da URL.

Datasets – Seleção de Atributos (WEKA)

Infogain:

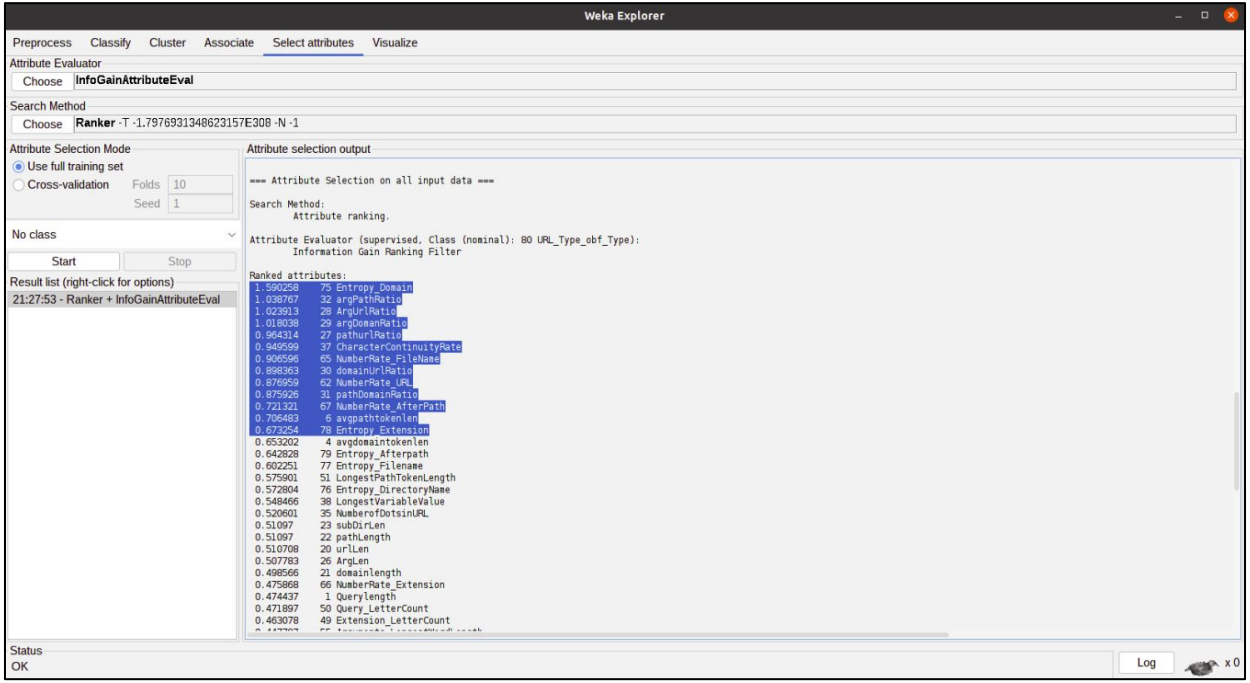
Determina o “peso” de um atributo através da medição do ganho de informação em respeito à classe

InfoGain(Class, Attribute) =

H(Class) - H(Class | Attribute).

Ranker

Ordena atributos de acordo com sua avaliação.



Dataset – Atributos escolhidos

Escolhidos 12 de 79 atributos

- Entropy_Domain
- argPathRatio
- ArgUrlRatio
- argDomanRatio
- pathurlRatio
- CharacterContinuityRate
- NumberRate_FileName
- domainUrlRatio
- NumberRate_URL
- pathDomainRatio
- NumberRate_AfterPath
- avgpathtokenlen

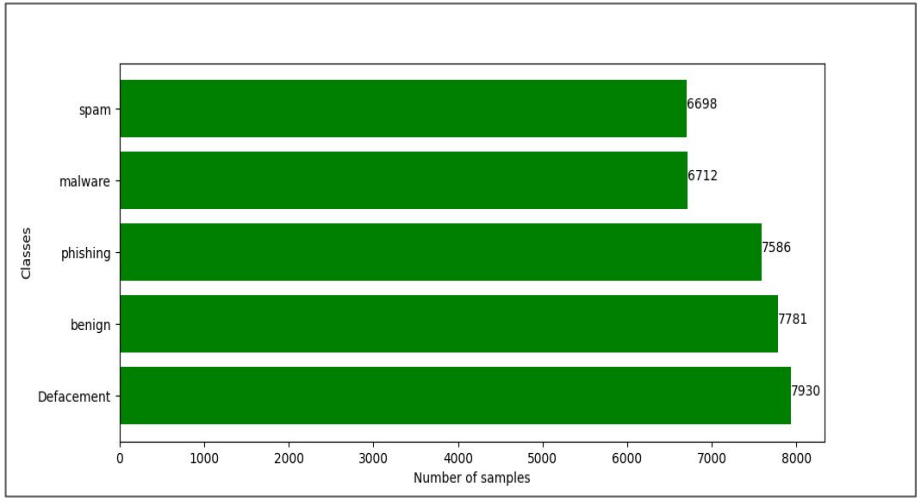
Dataset – Atributos escolhidos

	Entropy_Domain	argPathRatio	ArgUrlRatio	argDomanRatio	pathurlRatio	...	NumberRate_URL	pathDomainRatio	NumberRate_AfterPath	avgpathtokenlen	URL_Type_obf_Type
Querylength						...					
0	0.784493	0.076923	0.034483	0.080000	0.448276	...	0.017241	1.040000	-1.000000	4.400000	Defacement
0	0.784493	0.058824	0.030303	0.080000	0.515151	...	0.000000	1.360000	-1.000000	6.000000	Defacement
0	0.784493	0.060606	0.030769	0.080000	0.507692	...	0.000000	1.320000	-1.000000	5.800000	Defacement
0	0.784493	0.025974	0.018349	0.080000	0.706422	...	0.000000	3.080000	-1.000000	5.500000	Defacement
0	0.784493	0.040816	0.024691	0.080000	0.604938	...	0.000000	1.960000	-1.000000	7.333334	Defacement
...	...	...	...	...	...	...	...	...	...	...	...
29	0.791265	0.752212	0.582192	3.269231	0.773973	...	0.212329	4.346154	0.066667	3.666667	spam
0	0.820010	0.016393	0.013605	0.111111	0.829932	...	0.142857	6.777778	-1.000000	8.461538	spam
58	0.801139	0.838710	0.739837	8.272727	0.882114	...	0.231707	9.863636	0.029412	3.375000	spam
35	0.897617	0.755319	0.612069	4.733333	0.810345	...	0.215517	6.266667	0.418182	3.600000	spam
40	0.801139	0.828283	0.722467	7.454546	0.872247	...	0.229075	9.000000	0.060000	3.250000	spam

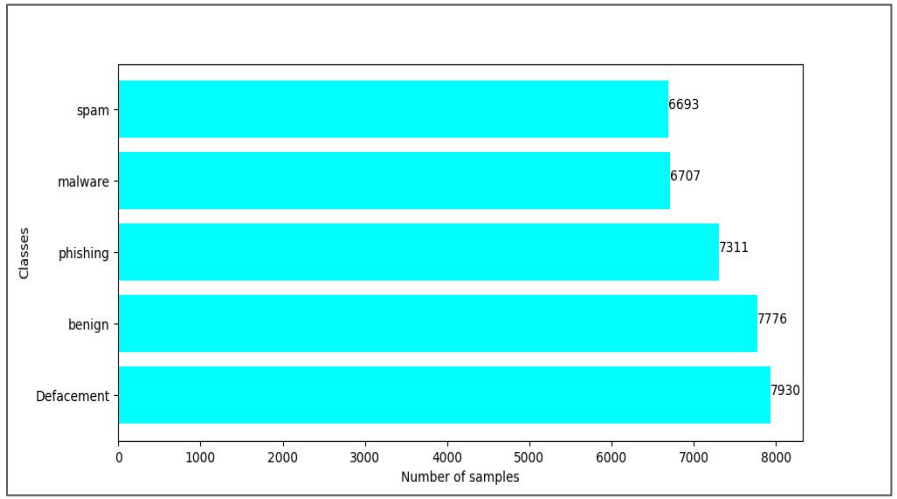
[36707 rows x 13 columns]

Dataset – Distribuição

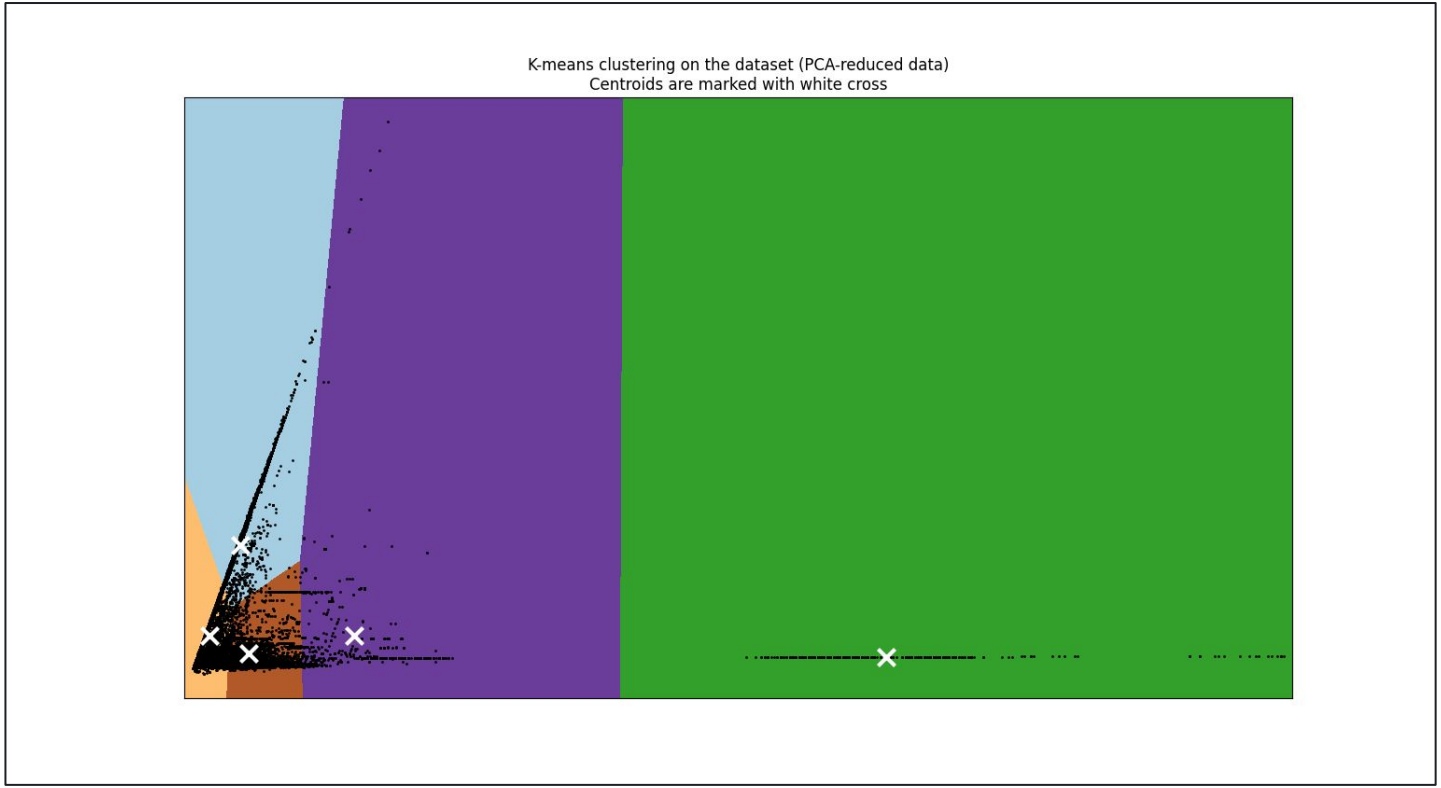
Todos os dados



Infogain + Dropnan



Dataset – Clusters



Fontes

- Dataset utilizado:
  - https://www.unb.ca/cic/datasets/url-2016.html
- Paper “Detecting Malicious URLs Using Lexical Analysis”:
  - https://www.researchgate.net/publication/308365207_Detecting_Malicious_URLs_Using_Lexical_Analysis

Obrigado pela atenção!