DATA SCIENCE AZ IT BIZTONSÁGBAN

6. forduló

SOPHOS

A kategória támogatója: SOPHOS

Ismertető a feladathoz

Az Elastic Malware Benchmark for Empowering Researchers (EMBER) adathalmaz egy Portable Executable (PE) fájlokból álló publikus feature és label gyűjtemény.

Ebben a fordulóban az EMBER featurizációs technikával, ötleteivel és alapvető jellemzőivel fogunk megismerkedni, az EMBER adathalmaz egy mintavételezett részén. A "create_vectorized_features" lépést már végrehajtottuk az adatot, az eredménye megtalálható a tar fileban.

A fordulóhoz tartozó adathalmaz az alábbi linken található:

https://oitm-competition.s3.eu-west-2.amazonaws.com/round6/ember_sample.tar.gz

A fordulóhoz előfeltétel 4GB szabad hely illetve az EMBER python csomag importálása.

```
import ember
```

Felhasznált idő: 39:02/40:00 **Elért pontszám:** 0/8

1. feladat 0/1 pont

Miért nehéz publikus tanító adathalmazt találni kártékony binárisok felismeréséhez?

Válaszok

- [ ] Az ártalmatlan binárisokhoz könnyű a hozzáférés, nem állnak copyright védelem alatt
- [x] A káros mintákat gyakran third party szolgáltatók árulják, jellemzően megosztást tiltó licence alatt
  Ez a válasz helyes, de nem jelölted meg.
- [x] Ismert vírusok megosztása biztonsági kockázat a felhasználók számára
  Ez a válasz helyes, de nem jelölted meg.
- [x] Címkézési nehézségek; annak a meghatározása, hogy egy bináris káros, vagy ártalmatlan-e jóval költségesebb folyamat mint egy kép vagy szöveg címkézése
  Ez a válasz helyes, de nem jelölted meg.

Magyarázat

Az ártalmatlan binárisok jellemzően copyright védelem alatt állnak. A többi állitás igaz.

2. feladat 0/1 pont

Melyik python csomagot használták az EMBER adathalmaz szerzői a bináris fájlok parszolásához?

Válaszok

A helyes válasz:

LIEF

Library to Instrument Executable Formats

LIEF : Library to Instrument Executable Formats

lief=0.9.0

lief 0.9.0

Magyarázat

https://github.com/elastic/ember 2. paragrafus

3. feladat 0/1 pont

Mekkora feature vektort kapunk az EMBERv2 adathalmaznál egy bináris fájlhoz?

Válasz

A helyes válasz:

2381

Magyarázat

```
from ember.features import PEFeatureExtractor
extractor = PEFeatureExtractor(2)
extractor.features

s = 0
for f in extractor.features:
    s += f.dim
```

4. feladat 0/1 pont

Hány fő feature típust (FeatureType) különböztet meg az EMBERv2 featurizáció, ha egy feature típust dimenziótól függetlenül egyszer számolunk?

Válasz

A helyes válasz:

9

Magyarázat

```
import ember
from ember.features import PEFeatureExtractor
extractor = PEFeatureExtractor(2)
extractor.features

>>>
[histogram(256),
 byteentropy(256),
 strings(104),
 general(10),
 header(62),
 section(255),
 imports(1280),
 exports(128),
 datadirectories(30)]
```

5. feladat 0/1 pont

Az alábbiak közül mely PE jellemzőket featurizálja a SectionInfo?

Válaszok

- [x] A szekció neve
  Ez a válasz helyes, de nem jelölted meg.
- [ ] A PE mérete
- [x] A szekció entrópiája
  Ez a válasz helyes, de nem jelölted meg.
- [x] MEM_WRITE hívások száma
  Ez a válasz helyes, de nem jelölted meg.

Magyarázat

https://github.com/elastic/ember/blob/17b459c8a23ac17d7423c2627b837b3e8cb326c2/ember/features.py#L125

6. feladat 0/1 pont

Mekkora a káros / ártalmatlan minták fájl(byte)hosszának aránya a tanítóhalmazban, két tizedesjegyre felkeritve?

Válasz

A helyes válasz:

0.87

Magyarázat

```
import ember
import ujson as json
import pandas as pd
import numpy as np


records = map(json.loads, open('./ember_sample/train_features_samples.jsonl'))


def parse_size(row):
    return row['size']


df = pd.DataFrame.from_records(records)
df[df['label'] == 1]['general'].apply(parse_size).sum() / df[df['label'] == 0]['general'].apply(parse_size).s
```

7. feladat 0/1 pont

A feature vektor melyik indexénél találjuk a fájlok méretét?

Válasz

A helyes válasz:

616

Magyarázat

https://github.com/elastic/ember/blob/17b459c8a23ac17d7423c2627b837b3e8cb326c2/ember/features.py#L283

A fájlok méretét a GeneralFileInfo class 0. dimenziója tárolja.

A GeneralFileInfot megelőzi a ByteHistogram(256), ByteEntropyHistogram(256), StringExtractor(104) a feature vektorban.

256 + 256 + 104 = 616

8. feladat 0/1 pont

Találtunk egy binárist, ami a következő dll-eket importálja:

libraries = ['ORSZAGOS.dll', 'IT.dll', 'MEGMERETTETES.dll]

Ha csak az importált könyvtárak részhalmazát tekintjük feature-öknek, hány ütköző minta van a tanító halmazban?

Válasz

A helyes válasz:

2

Magyarázat

```
import ember
import pandas as pd
import numpy as np


X_train, y_train, X_test, y_test = ember.read_vectorized_features('./ember_sample/')

from sklearn.feature_extraction import FeatureHasher

libraries = ['ORSZAGOS.dll', 'IT.dll', 'MEGMERETTETES.dll']
r = FeatureHasher(256, input_type="string").transform([libraries]).toarray()[0]
print(np.nonzero(r))
print(r[np.nonzero(r)])


# https://github.com/elastic/ember/blob/d97a0b523de02f3fe5ea6089d080abacab6ee931/ember/features.py#L508
# csak az importinfot toltsuk dataframebe
offset = 256 + 256 + 104 + 10 + 62 + 255
df = pd.DataFrame(X_train[: , offset: offset + 1280])
df[(df[32] == 1) & (df[54] == 1) & (df[99] == -1)]
```

KÉSZÍTETTE cone

Megjelenés

Világos