Laboratorio 5

Laboratorio 5

Security Data Science

Manuel Archila

Importar librerias

```
import pandas as pd
from clasificador import *
import json
```

1. Leer el json

```
file_path = './large_eve.json'


data = []
with open(file_path, 'r') as file:
    for line in file:
        record = json.loads(line)
        data.append(record)


total_records = len(data)
print(f'Total de registros: {total_records}')
```

```
Total de registros: 746909
```

2. Filtrar por DNS

```
dns_records = [record for record in data if record['event_type'] == 'dns']
total_dns_records = len(dns_records)
print(f'Total de registros DNS: {total_dns_records}')
```

```
Total de registros DNS: 15749
```

3. Normalizar el json y hacer el dataset

```
df_normalized = pd.json_normalize(dns_records)
df_normalized.shape
```

```
(15749, 18)
```

4. Filtrar por DNS tipo A

```
dns_a_records = df_normalized[(df_normalized['event_type'] == 'dns') & (df_normalized['d
# Contamos el número de registros después del filtrado
num_dns_a_records = len(dns_a_records)
print(f'Total de registros DNS tipo A: {num_dns_a_records}')
print(dns_a_records.shape)
```

```
Total de registros DNS tipo A: 2849
(2849, 18)
```

5. Filtrar por nombres de dominios unicos

```
unique_domains = dns_a_records['dns.rrname'].unique()
print(unique_domains)

# Contar el número de dominios únicos
num_unique_domains = len(unique_domains)
print(f'Número de dominios únicos: {num_unique_domains}')
```

```
['api.wunderground.com' 'stork79.dropbox.com'
 'hpca-tier2.office.aol.com.ad.aol.aoltw.net'
 'safebrowsing.clients.google.com.home' 'fxfeeds.mozilla.com'
 'www.metasploit.com' 'aolmtcmxm03.office.aol.com'
 'aolmtcmxm02.office.aol.com.ad.aol.aoltw.net'
 'aolmtcmxm02.office.aol.com' 'hpca-tier2.office.aol.com'
 'aolmtcmxm03.office.aol.com.ad.aol.aoltw.net'
 'aolmtcmxm04.office.aol.com' 'safebrowsing.clients.google.com'
 'wpad.home' 'safebrowsing.clients.google.com.stayonline.net'
 'aolmtcmxm04.office.aol.com.ad.aol.aoltw.net'
 'AOLDTCMA04.ad.aol.aoltw.net.office.aol.com' 'AOLDTCMA04.office.aol.com'
 '192.168.22.110phpmyadmin' 'secure.informaction.com'
 'secure.informaction.com.localdomain'
 'safebrowsing.clients.google.com.localdomain' 'ueip.vmware.com'
 'en-us.fxfeeds.mozilla.com' '192.168.22.110phpmyadmin.localdomain'
 'time.windows.com' 'softwareupdate.vmware.com' 'proxim.ntkrnlpa.info'
 'portswigger.net' 'www.offensive-security.com'
 'www.offensive-security.com.stayonline.net' 'www.stopbadware.org'
 'AOLDTCMA04.ad.aol.aoltw.net' 'gg.arrancar.org' 'www.sql-ledger.org'
 'www.backtrack-linux.org' 'www.backtrack-linux.org.stayonline.net'
 'en-us.start3.mozilla.com' 'www.theanime.cn' 'www.theanime.cn. '
```

```
'wpad.aol.aoltw.net' 'wpad.aoltw.net' 'tools.google.com.ad.aol.aoltw.net'
'safebrowsing.clients.google.com.hackerlabs.vpn'
'secure.informaction.com.stayonline.net' 'wpad.ad.aol.aoltw.net'
'knoa-coll.ops.aol.com' 'knoa-coll.ops.aol.com.ad.aol.aoltw.net'
'www.phpmyadmin.net' 'tools.google.com'
'toolbarqueries.clients.google.com' 'teredo.ipv6.microsoft.com'
'secure.informaction.com.home'
'toolbarqueries.clients.google.com.ad.aol.aoltw.net'
'secure.informaction.com.hsd1.pa.comcast.net' 'clients1.google.com'
'clients1.google.com.ad.aol.aoltw.net' 'ntp.ubuntu.com'
'en-us.www.mozilla.com' 'data.alexa.com' 'www.postgresql.org'
'sourceforge.net' 'www.freepbx.org'
'secure.informaction.com.hackerlabs.vpn' 'www.bigflickrfeed.com'
'www.gnu.org' 'wpad' 'safebrowsing.clients.google.com.lan'
'www.google.com' 'safebrowsing.clients.google.com.hsd1.pa.comcast.net'
'phppgadmin.sourceforge.net' '"192.168.206.56"' 'freepbx.org'
'erp.acunetix.com' 'www.acunetix.com' 'go.microsoft.com'
'download.windowsupdate.com' 'www.update.microsoft.com' 'api.flickr.com'
'widgets.alexa.com' 'download.microsoft.com'
'safebrowsing.clients.google.com.office.aol.com' 'www.malwarecity.com'
'api.facebook.com' 'www.sql-ledger.org.hsd1.pa.comcast.net'
'www.offensive-security.com.hsd1.pa.comcast.net' 'www.securityfocus.com'
'sync.xmarks.com' '192.168.26-27.0' 'www.cakephp.org' 'ky.hec.net'
'google.com' 'mirrors.adams.net' 'mirror.its.uidaho.edu'
'mirrors.cat.pdx.edu' 'mirror.clarkson.edu' 'mirror.rackspace.com'
'mirrors.ecvps.com' 'centos.cs.wisc.edu' 'centos.mirror.facebook.net'
'mirrors.easynews.com' 'mirrors.bluehost.com'
'centos.mirror.netriplex.com' 'mirror.stanford.edu' 'mirrors.tummy.com'
'mirror.ash.fastserv.com' 'mirrors.kernel.org' 'mirror.hmc.edu'
'mirrors.liquidweb.com' 'mirror.5ninesolutions.com'
'mirror.san.fastserv.com' 'updates.interworx.info' 'ftp.wallawalla.edu'
'mirror.sanctuaryhost.com' 'mirror.team-cymru.org' 'mirror.umoss.org'
'mirrors.gigenet.com' 'mirrors.xmission.com' 'repo.genomics.upenn.edu'
'centos.mirrors.tds.net' 'mirror.nyi.net' 'mirror.atlantic.net'
'ftp.usf.edu' 'mirror.rocketinternet.net' 'mirrors.rit.edu'
'clients5.google.com' 'FL' 'www.apple.com' 'internalcheck.apple.com'
'cloud.xmarks.com' 'www.metasploit.com.office.aol.com' 'saruman'
'clients2.google.com' '192.168.21.1201.stayonline.net'
'clients2.google.com.ad.aol.aoltw.net' '192.168.21.1201' ''
'fileservices.me.com' 'configuration.apple.com'
'r1s6i7.connectivity.me.com' 'images.apple.com' 'news.google.com'
'gdata.youtube.com' 'aosnotify.me.com' 'dns.msftncsi.com' 'kodapp.com'
'rc.threatspace.net' 'www.msftncsi.com' 'ul.backblaze.com'
'www.social-engineer.org' 'activex.microsoft.com' 'whitecell.localdomain'
'www.arduino.cc' 'secure.informaction.com.office.aol.com' 'www.mac.com'
'gfe.nvidia.com' 'addons.mozilla.org' 'versioncheck.addons.mozilla.org'
'idisk.mac.com' 'www.nagios.org' 'vtlfccmfxlkgifuf.com'
'linkhelp.clients.google.com.ad.aol.aoltw.net' 'update.macromates.com'
'1922.168.22.254' 'linkhelp.clients.google.com' '1922.168.22.254.home'
'192.168.21-28.0' 'ejfodfmfxlkgifuf.xyz' '192.168.21-28.0.home'
'192.168.22.201:' 'aoldtcmds01.office.aol.com'
```

```
 'aoldtcmds01.office.aol.com.ad.aol.aoltw.net'
 'ntp.ubuntu.com.localdomain' 'redir.metaservices.microsoft.com'
 'ocsp.verisign.com' '192.168.22.201:.stayonline.net'
 'client-software.real.com']
Número de dominios únicos: 177
```

6. Corregir los tlds

```
def obtener_tld_corregido(domain):
    parts = domain.split('.')
    for i in range(len(parts) - 1, 0, -1):
        tld_potencial = '.'.join(parts[i:])
        if tld_potencial in ["com", "net", "org", "gov", "edu", "mil", "arpa", "int"]:
            continue
        else:
            return tld_potencial
    return domain
```

```
tlds = [obtener_tld_corregido(dominio) for dominio in unique_domains]
```

```
tlds_df =  pd.DataFrame({'domain_tld': tlds})
tlds_df.head()
```

	domain_tld
0	wunderground.com
1	dropbox.com
2	aoltw.net
3	home
4	mozilla.com

7. Clasificar usando la funcion del archivo clasificador

```
df_final = clasificacion(tlds_df)
```

```
c:\Users\aleja\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.py:299:
UserWarning: Trying to unpickle estimator DecisionTreeClassifier from version 1.0.2 when using
version 1.2.1. This might lead to breaking code or invalid results. Use at your own risk. For
more info please refer to:
https://scikit-learn.org/stable/model_persistence.html#security-maintainability-limitations
  warnings.warn(
```

8. Filtrar por nombre de dominios nuevamente

```
dominios_dga = df_final[df_final['isDGA'] == 1]

dominios_dga_unicos = dominios_dga.drop_duplicates(subset=['domain_tld'])

print(f'Total de dominios DGA únicos: {dominios_dga_unicos.shape[0]}')
```

```
Total de dominios DGA únicos: 32
```

9. Verificar si el dominio se encuentra en la lista de dominios comunes

Prompt: “Necesito una funcion que que utilice la lista de un millón de TLD que te proporcione y que devuelva 0 si el TLD se encuentra en la lista y 1 si no está. Evita cargar la lista cada vez que se busca un TLD.”

```
import pandas as pd

# Cargar la lista de TLDs desde el archivo CSV
# Asegúrate de ajustar la ruta del archivo según donde tengas guardado 'top-1m.csv'
lista_tld_df = pd.read_csv('./top-1m.csv', header=None, names=['rank', 'tld'])
lista_tld = set(lista_tld_df['tld'])

# Definir la función que verifica si un TLD está en la lista
def verificar_tld(tld):
    return 0 if tld in lista_tld else 1
```

```
dominios_dga_unicos['es_dga_sospechoso'] = dominios_dga_unicos['domain_tld'].apply(verif
dominios_dga_unicos.head()
```

```
C:\Users\aleja\AppData\Local\Temp\ipykernel_29224\3854646683.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  dominios_dga_unicos['es_dga_sospechoso'] =
dominios_dga_unicos['domain_tld'].apply(verificar_tld)
```

	domain_tld	isDGA	es_dga_sospechoso
1	dropbox.com	1	0
2	aoltw.net	1	1
5	metasploit.com	1	0
18	110phpmyadmin	1	1
25	windows.com	1	0

```
dominios_no_sospechosos = dominios_dga_unicos[dominios_dga_unicos['es_dga_sospechoso'] =

dominios_no_sospechosos_unicos = dominios_no_sospechosos.drop_duplicates(subset=['domain_

print(f'Total de dominios no DGA sospechosos únicos: {dominios_no_sospechosos_unicos.sha
print(dominios_no_sospechosos_unicos[['domain_tld', 'es_dga_sospechoso']])
```

```
Total de dominios no DGA sospechosos únicos: 10
               domain_tld  es_dga_sospechoso
2                aoltw.net                  1
18           110phpmyadmin                  1
34          sql-ledger.org                  1
35    backtrack-linux.org                   1
64      bigflickrfeed.com                   1
82         malwarecity.com                  1
89             cakephp.org                  1
97               ecvps.com                  1
160  vtlfccmfxlkgifuf.com                   1
162         macromates.com                  1
```

10. Obtener fecha de creacion de los dominios

Prompt: “Necesito una función qué en base al TLD, devuelva la fecha de creación de dicho dominio”

```
import whois
from datetime import datetime

def obtener_fecha_creacion_tld(tld):

    try:
        # Realizar la consulta WHOIS para el TLD proporcionado
        w = whois.whois(tld)

        # Obtener la fecha de creación desde el objeto WHOIS
        fecha_creacion = w.creation_date

        # La fecha de creación puede ser una lista o un solo valor, dependiendo del TLD
        if isinstance(fecha_creacion, list):
            fecha_creacion = fecha_creacion[0]

        # Formatear la fecha de creación como string en formato 'YYYY-MM-DD'
        if fecha_creacion:
            if isinstance(fecha_creacion, datetime):
                return fecha_creacion.strftime('%Y-%m-%d')
            else:
                return str(fecha_creacion)
        else:
            return "No se pudo encontrar la fecha de creación."
```

```
            except Exception as e:
                return "Error al buscar la fecha de creación: " + str(e)
```

```
dominios_fecha_creacion = dominios_no_sospechosos_unicos.apply(lambda x: obtener_fecha_c
```

```
dominios_fecha_creacion
```

```
2                                          2000-01-10
18            No se pudo encontrar la fecha de creación.
34                                         2000-09-08
35                                         2009-04-29
64     Error al buscar la fecha de creación: No match...
82                                         2008-02-06
89                                         2005-06-13
97                                         2009-05-21
160    Error al buscar la fecha de creación: No match...
162                                        2003-02-05
dtype: object
```

Conclusiones

Los dominios que tienen una fecha de creación muy antigua son comunmente de entidades conocidas y respetables debido a que eran los principios de la internet publica. Es por esto que el objetivo principal eran los dominios con fechas de creación relativamente recientes. Al observar las fechas de creación de los dominios los que más llaman la atención, debido a su fecha de creación reciente, son: - ecvps.com - backtrack-linux.org - malwarecity.com

Sin embargo, hay algunos dominios que no tienen fecha de creación asociada. Esto puede ser un indicador de que el dominio fue maliciosa o fraudulento en algun momento y fue eliminado. Estos dominios son: - 110phpmyadmin - bigflickrfeed.com - vtlfccmfxlkgifuf.com

Basandose en la fechas de creación y los patrones de letras aleatorias encontrados en los tlds se puede inferir que los dominios que son maliciosos o fraudulentos son: - ecvps.com - vtlfccmfxlkgifuf.com