3. DATA PREPROCESSING

January 2, 2026

Importing Libraries

[2]:
```
import pandas as pd
import numpy as np
```

Reading the data

[3]:
```
data = pd.read_csv('data/transformed/final_raw_data.csv')

data.head()
```

[3]:

	url	label
0	https://adoaeco.cn/Loggin	phishing
1	https://gageparkhighschool.com/QTeuUe	phishing
2	https://wwnox.miraltek.cfd/qzxn3	phishing
3	https://halfetitur.com/?token=r2I0IU0FEHfPf5Dn	phishing
4	https://yqcjl.miraltek.cfd/plis0	phishing

[4]:
```
data.duplicated(subset=['url']).sum()
```

[4]:
```
np.int64(0)
```

[5]:
```
# URL Components features
url_components_df = pd.read_csv('data/transformed/1.url_components_data.csv')

url_components_df.head()
```

[5]:

	url	label	protocol	domain	subdomain	tld	sld	path
0	https://adoaeco.cn/Loggin	phishing	https	adoaeco.cn	NaN	cn	adoaeco	/Loggin
1	https://gageparkhighschool.com/QTeuUe	phishing	https	gageparkhighschool.com	NaN	com	gageparkhighschool	/QTeuUe
2	https://wwnox.miraltek.cfd/qzxn3	phishing	https	wwnox.miraltek.cfd	wwnox	cfd	miraltek	/qzxn3
3	https://halfetitur.com/?token=r2I0IU0FEHfPf5Dn	phishing	https	halfetitur.com	NaN	com	halfetitur	/
4	https://yqcjl.miraltek.cfd/plis0	phishing	https	yqcjl.miraltek.cfd	yqcjl	cfd	miraltek	/plis0

```
4      yqcjl.miraltek.cfd     yqcjl  cfd            miraltek   /plis0

                    query
0                     NaN
1                     NaN
2                     NaN
3  token=r2I0IU0FEHfPf5Dn
4                     NaN
```

[6]:
```
# URL component length features data
len_features_df = pd.read_csv('data/transformed/2.component_len_features_data.
 ↪csv')

len_features_df.head()
```

[6]:
```
                                          url     label  url_len  \
0                      https://adoaeco.cn/Loggin  phishing       25
1           https://gageparkhighschool.com/QTeuUe  phishing       37
2                https://wwnox.miraltek.cfd/qzxn3  phishing       32
3  https://halfetitur.com/?token=r2I0IU0FEHfPf5Dn  phishing       46
4                https://yqcjl.miraltek.cfd/plis0  phishing       32

   domain_len  path_len  query_len  url_depth  subdomain_count
0          10         6          0          1                1
1          22         6          0          1                1
2          18         5          0          1                1
3          14         0         22          1                1
4          18         5          0          1                1
```

[7]:
```
# Domain features data
domain_features_df = pd.read_csv('data/transformed/3.domain_features_data.csv')

domain_features_df.head()
```

[7]:
```
                                          url     label  tld  tld_len  \
0                      https://adoaeco.cn/Loggin  phishing   cn        2
1           https://gageparkhighschool.com/QTeuUe  phishing  com        3
2                https://wwnox.miraltek.cfd/qzxn3  phishing  cfd        3
3  https://halfetitur.com/?token=r2I0IU0FEHfPf5Dn  phishing  com        3
4                https://yqcjl.miraltek.cfd/plis0  phishing  cfd        3

   url_has_ipv4  url_has_port
0         False         False
1         False         False
2         False         False
3         False         False
4         False         False
```

[8]:
```
# SLD features data
sld_features_df = pd.read_csv('data/transformed/4.sld_features_data.csv')

sld_features_df.head()
```

[8]:
```
                                          url     label  \
0                      https://adoaeco.cn/Loggin  phishing
1           https://gageparkhighschool.com/QTeuUe  phishing
2                https://wwnox.miraltek.cfd/qzxn3  phishing
3  https://halfetitur.com/?token=r2I0IU0FEHfPf5Dn  phishing
4                https://yqcjl.miraltek.cfd/plis0  phishing

                  sld  sld_len  sld_has_digit  sld_has_hyphen  sld_token_count
0             adoaeco        7          False           False                1
1  gageparkhighschool       18          False           False                1
2            miraltek        8          False           False                1
3          halfetitur       10          False           False                1
4            miraltek        8          False           False                1
```

[9]:
```
# Character features data
char_feature_df = pd.read_csv('data/transformed/5.char_features_data.csv')

char_feature_df.head()
```

[9]:
```
                                          url     label  dot_count_domain  \
0                      https://adoaeco.cn/Loggin  phishing                 1
1           https://gageparkhighschool.com/QTeuUe  phishing                 1
2                https://wwnox.miraltek.cfd/qzxn3  phishing                 2
3  https://halfetitur.com/?token=r2I0IU0FEHfPf5Dn  phishing                 1
4                https://yqcjl.miraltek.cfd/plis0  phishing                 2

   hyphen_count_domain_path  underscore_count_path_query  slash_count  \
0                         0                            0            3
1                         0                            0            3
2                         0                            0            3
3                         0                            0            3
4                         0                            0            3

   digit_count  alphabet_count  spl_char_count
0            0              20               5
1            0              32               5
2            1              25               6
3            4              35               7
4            1              25               6
```

[10]:
```
# Entropy features data
entropy_feature_df = pd.read_csv('data/transformed/6.entropy_feature_data.csv')
```

```
entropy_feature_df.head()
```

[10]:

	url	label	url_entropy	domain_entropy	sld_entropy	path_entropy
0	https://adoaeco.cn/Loggin	phishing	3.863465	2.721928	2.235926	2.521641
1	https://gageparkhighschool.com/QTeuUe	phishing	4.208925	3.629220	3.419382	2.521641
2	https://wwnox.miraltek.cfd/qzxn3	phishing	4.452820	3.947703	3.000000	2.584963
3	https://halfetitur.com/?token=r2I0IU0FEHfPf5Dn	phishing	4.760096	3.664498	3.121928	-0.000000
4	https://yqcjl.miraltek.cfd/plis0	phishing	4.241729	3.836592	3.000000	2.584963

[11]:
```
# Token features data
token_feature_df = pd.read_csv('data/transformed/7.token_features_data.csv')

token_feature_df.head()
```

[11]:

	url	label	domain_token_count	path_token_count	total_tokens	avg_token_length
0	https://adoaeco.cn/Loggin	phishing	2	1	3	5.00
1	https://gageparkhighschool.com/QTeuUe	phishing	2	1	3	9.00
2	https://wwnox.miraltek.cfd/qzxn3	phishing	3	1	4	5.25
3	https://halfetitur.com/?token=r2I0IU0FEHfPf5Dn	phishing	2	1	3	8.50
4	https://yqcjl.miraltek.cfd/plis0	phishing	3	1	4	5.25

[12]:
```
# Hexadecimal feature data
hex_feature_df = pd.read_csv('data/transformed/8.hex_features_data.csv')

hex_feature_df.head()
```

[12]:

	url	label	has_hex
0	https://adoaeco.cn/Loggin	phishing	False
1	https://gageparkhighschool.com/QTeuUe	phishing	False
2	https://wwnox.miraltek.cfd/qzxn3	phishing	False
3	https://halfetitur.com/?token=r2I0IU0FEHfPf5Dn	phishing	False
4	https://yqcjl.miraltek.cfd/plis0	phishing	False

```
   hex_char_count  hex_ratio
0               0        0.0
1               0        0.0
2               0        0.0
3               0        0.0
4               0        0.0
```

[13]:
```
df_dict = {
    'URL components' : url_components_df,
    'Length features' : len_features_df,
    'Domain features' : domain_features_df,
    'SLD features' : sld_features_df,
    'Character features' : char_feature_df,
    'Entropy features' : entropy_feature_df,
    'Token features' : token_feature_df,
    'Hexadecimal features' : hex_feature_df
}
```

Handling null values

[14]:
```
def null_cols(df):
    null_counts = df.isnull().sum()
    null_cols = null_counts[null_counts > 0]

    if not null_counts.empty:
        print(null_cols)
    else:
        print('No null values found')
```

[15]:
```
for df_name,df in df_dict.items():
    print(df_name)
    null_cols(df)
    print()
```

```
URL components
domain         2283
subdomain     64887
tld            2435
sld            2286
path          48380
query        214330
dtype: int64

Length features
Series([], dtype: int64)

Domain features
```

```
tld    2435
dtype: int64

SLD features
sld    2286
dtype: int64

Character features
Series([], dtype: int64)

Entropy features
Series([], dtype: int64)

Token features
Series([], dtype: int64)

Hexadecimal features
Series([], dtype: int64)
```

The URL Components data, Domain features and SLD features consists of null values

[16]:
```
domain_features_df[domain_features_df['tld'].isnull()]
```

[16]:
```
                                                    url     label  tld  \
2994                              https://91.92.241.186  phishing  NaN
3667                            https://140.99.164.68/x0  phishing  NaN
6300                            https://31.172.87.101/x0  phishing  NaN
14536                             https://43.153.99.18  phishing  NaN
16733                           https://185.187.56.126  phishing  NaN
…                                                  …       …  …
252716                 http://191.101.7.221/fire/aasdqwe  phishing  NaN
252811                          http://91.239.25.38:6892  phishing  NaN
252961  http://178.217.186.224/panel/etc/info/toke/cp…  phishing  NaN
252990               http://185.75.46.73/information.cgi  phishing  NaN
252997  http://38.118.40.209/CFIDE/debug/serveur.html?…  phishing  NaN

        tld_len  url_has_ipv4  url_has_port
2994          0          True         False
3667          0          True         False
6300          0          True         False
14536         0          True         False
16733         0          True         False
…           …             …             …
252716        0          True         False
252811        0          True          True
252961        0          True         False
252990        0          True         False
```

```
252997        0          True         False

[2435 rows x 6 columns]
```

[17]:
```
sld_features_df[sld_features_df['sld'].isnull()]
```

[17]:
```
                                                      url     label  sld  \
2994                                https://91.92.241.186  phishing  NaN
3667                              https://140.99.164.68/x0  phishing  NaN
6300                              https://31.172.87.101/x0  phishing  NaN
14536                                 https://43.153.99.18  phishing  NaN
16733                               https://185.187.56.126  phishing  NaN
…                                                      …       …  …
252716                   http://191.101.7.221/fire/aasdqwe  phishing  NaN
252811                             http://91.239.25.38:6892  phishing  NaN
252961  http://178.217.186.224/panel/etc/info/toke/cp…  phishing  NaN
252990                 http://185.75.46.73/information.cgi  phishing  NaN
252997  http://38.118.40.209/CFIDE/debug/serveur.html?…  phishing  NaN

        sld_len  sld_has_digit  sld_has_hyphen  sld_token_count
2994          0          False           False                1
3667          0          False           False                1
6300          0          False           False                1
14536         0          False           False                1
16733         0          False           False                1
…         …              …               …              …
252716        0          False           False                1
252811        0          False           False                1
252961        0          False           False                1
252990        0          False           False                1
252997        0          False           False                1

[2286 rows x 7 columns]
```

The URLs where TLDs & SLDs having null values are mostly IP address based URLs. So, the numerical features dependent on TLD & SLD will be 0. In URL components data, there are many null values in Domain, SLD and TLD. These are IP address based URLs. Other features also have many null values since we are considering numerical features, we will ignore those values.

Combining all the features into a single dataset

[18]:
```
df = pd.DataFrame()     # dataframe to store all the processed features
```

[19]:
```
# URL components data

df['has_https'] = url_components_df['protocol'].map(lambda x: 1 if x == 'https'␣
 ↪else 0)
```

[20]:
```
# URL Length features data

df[['url_len','domain_len','path_len','query_len','url_depth','subdomain_count']]␣
 ↪= len_features_df.select_dtypes('number')
```

[21]:
```
# Domain features data

df['tld_len'] = domain_features_df['tld_len']
df[['url_has_ipv4','url_has_port']] =␣
 ↪domain_features_df[['url_has_ipv4','url_has_port']].astype('int64')
```

[22]:
```
# SLD features data

df['sld_len'] = sld_features_df['sld_len']
df[['sld_has_digit','sld_has_hyphen']] =␣
 ↪sld_features_df[['sld_has_digit','sld_has_hyphen']].astype('int64')
df['sld_token_count'] = sld_features_df['sld_token_count']
```

[23]:
```
# Character features data

df[['dot_count_domain','hyphen_count_domain_path','underscore_count_path_query','slash_count',
 ↪= char_feature_df.select_dtypes('number')
```

[24]:
```
# Entropy features data

df[['url_entropy','domain_entropy','sld_entropy','path_entropy']] =␣
 ↪entropy_feature_df.select_dtypes('number')
```

[25]:
```
# Token features data

df[['domain_token_count','path_token_count','total_tokens','avg_token_length']]␣
 ↪= token_feature_df.select_dtypes('number')
```

We will ignore Hexadecimal-based features since the hexadecimal features in the data are very low and its contribution is very less in predictions.

[26]:
```
# Adding label

df['class'] = url_components_df['label'].apply(lambda x: 1 if x == 'phishing'␣
 ↪else 0)
```

[27]:
```
df.head()
```

[27]:

	has_https	url_len	domain_len	path_len	query_len	url_depth	\
0	1	25	10	6	0	1	
1	1	37	22	6	0	1	
2	1	32	18	5	0	1	
3	1	46	14	0	22	1	

```
4          1       32          18         5          0          1

   subdomain_count  tld_len  url_has_ipv4  url_has_port  …  spl_char_count  \
0                1        2             0             0  …               5
1                1        3             0             0  …               5
2                1        3             0             0  …               6
3                1        3             0             0  …               7
4                1        3             0             0  …               6

   url_entropy  domain_entropy  sld_entropy  path_entropy  domain_token_count  \
0     3.863465        2.721928     2.235926      2.521641                   2
1     4.208925        3.629220     3.419382      2.521641                   2
2     4.452820        3.947703     3.000000      2.584963                   3
3     4.760096        3.664498     3.121928     -0.000000                   2
4     4.241729        3.836592     3.000000      2.584963                   3

   path_token_count  total_tokens  avg_token_length  class
0                 1             3              5.00      1
1                 1             3              9.00      1
2                 1             4              5.25      1
3                 1             3              8.50      1
4                 1             4              5.25      1

[5 rows x 30 columns]
```

[28]:
```
print(f'The combined dataset consists of {df.shape[0]} rows and {df.shape[1]}␣
↪columns')
```

```
The combined dataset consists of 253051 rows and 30 columns
```

[29]:
```
df.columns
```

[29]:
```
Index(['has_https', 'url_len', 'domain_len', 'path_len', 'query_len',
       'url_depth', 'subdomain_count', 'tld_len', 'url_has_ipv4',
       'url_has_port', 'sld_len', 'sld_has_digit', 'sld_has_hyphen',
       'sld_token_count', 'dot_count_domain', 'hyphen_count_domain_path',
       'underscore_count_path_query', 'slash_count', 'digit_count',
       'alphabet_count', 'spl_char_count', 'url_entropy', 'domain_entropy',
       'sld_entropy', 'path_entropy', 'domain_token_count', 'path_token_count',
       'total_tokens', 'avg_token_length', 'class'],
      dtype='object')
```

[30]:
```
df.dtypes
```

[30]:
```
has_https                    int64
url_len                      int64
domain_len                   int64
path_len                     int64
```

```
query_len                        int64
url_depth                        int64
subdomain_count                  int64
tld_len                          int64
url_has_ipv4                     int64
url_has_port                     int64
sld_len                          int64
sld_has_digit                    int64
sld_has_hyphen                   int64
sld_token_count                  int64
dot_count_domain                 int64
hyphen_count_domain_path         int64
underscore_count_path_query      int64
slash_count                      int64
digit_count                      int64
alphabet_count                   int64
spl_char_count                   int64
url_entropy                    float64
domain_entropy                 float64
sld_entropy                    float64
path_entropy                   float64
domain_token_count               int64
path_token_count                 int64
total_tokens                     int64
avg_token_length               float64
class                            int64
dtype: object
```

[31]:
```
df.duplicated().sum()
```

[31]:
```
np.int64(29531)
```

[32]:
```
df.drop_duplicates(keep='first',ignore_index=True,inplace=True)
```

[33]:
```
# Saving the dataset

df.to_csv(r'data/processed/processed_data.csv',index=False)
```