3. DATA PREPROCESSING

January 23, 2026

Importing Libraries

[10]:
```
import pandas as pd
import numpy as np
```

Reading the data

[11]:
```
data = pd.read_csv('data/transformed/final_raw_data.csv')

data.head()
```

[11]:
```
                                                 url       label
0                        https://www.visitcanada.com  legitimate
1                  http://218.228.19.9/~yossi/9ssfpkz    phishing
2  https://www.msupress.msu.edu/series.php?series…  legitimate
3  https://docs.google.com/presentation/d/e/2PACX…    phishing
4  https://www.c250.columbia.edu/c250_celebrates/…  legitimate
```

[12]:
```
data.duplicated(subset=['url']).sum()
```

[12]:
```
np.int64(0)
```

[13]:
```
# URL Components features
url_components_df = pd.read_csv('data/transformed/1.url_components_data.csv')

url_components_df.head()
```

[13]:
```
                                                 url       label protocol  \
0                        https://www.visitcanada.com  legitimate    https
1                  http://218.228.19.9/~yossi/9ssfpkz    phishing     http
2  https://www.msupress.msu.edu/series.php?series…  legitimate    https
3  https://docs.google.com/presentation/d/e/2PACX…    phishing    https
4  https://www.c250.columbia.edu/c250_celebrates/…  legitimate    https

                  domain     subdomain  tld          sld  \
0    www.visitcanada.com           www  com  visitcanada
1                    NaN           NaN  NaN          NaN
2   www.msupress.msu.edu  www.msupress  edu          msu
3        docs.google.com          docs  com       google
```

```
4  www.c250.columbia.edu      www.c250  edu     columbia

                                                  path  \
0                                                  NaN
1                                       /~yossi/9ssfpkz
2                                          /series.php
3  /presentation/d/e/2PACX-1vRBjV4Bm4UxL3gJ8sCyQx…
4  /c250_celebrates/athletics/athletics_timeline…

                                           query
0                                            NaN
1                                            NaN
2                                     seriesID=17
3  start=false&amp;loop=false&amp;delayms=3000
4                                            NaN
```

[14]:
```
# URL component length features data
len_features_df = pd.read_csv('data/transformed/2.component_len_features_data.
 ↪csv')

len_features_df.head()
```

[14]:
```
                                              url       label  url_len  \
0                         https://www.visitcanada.com  legitimate       27
1                    http://218.228.19.9/~yossi/9ssfpkz    phishing       34
2  https://www.msupress.msu.edu/series.php?series…  legitimate       51
3  https://docs.google.com/presentation/d/e/2PACX…    phishing      175
4  https://www.c250.columbia.edu/c250_celebrates/…  legitimate       79

   domain_len  path_len  query_len  url_depth  subdomain_count
0          19         0          0          0                1
1           0        13          0          2                0
2          20        10         11          1                2
3          15       103         43          5                1
4          21        47          0          3                2
```

[15]:
```
# Domain features data
domain_features_df = pd.read_csv('data/transformed/3.domain_features_data.csv')

domain_features_df.head()
```

[15]:
```
                                              url       label  tld  \
0                         https://www.visitcanada.com  legitimate  com
1                    http://218.228.19.9/~yossi/9ssfpkz    phishing  NaN
2  https://www.msupress.msu.edu/series.php?series…  legitimate  edu
3  https://docs.google.com/presentation/d/e/2PACX…    phishing  com
4  https://www.c250.columbia.edu/c250_celebrates/…  legitimate  edu
```

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

[16]:
```
# SLD features data
sld_features_df = pd.read_csv('data/transformed/4.sld_features_data.csv')

sld_features_df.head()
```

[16]:
```
                                                 url       label          sld  \
0                       https://www.visitcanada.com  legitimate  visitcanada
1                  http://218.228.19.9/~yossi/9ssfpkz    phishing          NaN
2  https://www.msupress.msu.edu/series.php?series…  legitimate          msu
3  https://docs.google.com/presentation/d/e/2PACX…    phishing       google
4  https://www.c250.columbia.edu/c250_celebrates/…  legitimate     columbia

   sld_len  sld_has_digit  sld_has_hyphen  sld_token_count
0       11          False           False                1
1        0          False           False                1
2        3          False           False                1
3        6          False           False                1
4        8          False           False                1
```

[17]:
```
# Character features data
char_feature_df = pd.read_csv('data/transformed/5.char_features_data.csv')

char_feature_df.head()
```

[17]:
```
                                                 url       label  \
0                       https://www.visitcanada.com  legitimate
1                  http://218.228.19.9/~yossi/9ssfpkz    phishing
2  https://www.msupress.msu.edu/series.php?series…  legitimate
3  https://docs.google.com/presentation/d/e/2PACX…    phishing
4  https://www.c250.columbia.edu/c250_celebrates/…  legitimate

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

   slash_count  digit_count  alphabet_count  spl_char_count
```

```
0              2              0              22              5
1              4             10              15              9
2              3              2              39             10
3              7             19             135             21
4              5              6              61             12
```

[18]:
```
# Entropy features data
entropy_feature_df = pd.read_csv('data/transformed/6.entropy_features_data.csv')

entropy_feature_df.head()
```

[18]:
```
                                                 url       label  url_entropy  \
0                       https://www.visitcanada.com  legitimate     3.856196
1                  http://218.228.19.9/~yossi/9ssfpkz    phishing     3.962032
2  https://www.msupress.msu.edu/series.php?series…  legitimate     3.965393
3  https://docs.google.com/presentation/d/e/2PACX…    phishing     5.569700
4  https://www.c250.columbia.edu/c250_celebrates/…  legitimate     4.274946

   domain_entropy  sld_entropy  path_entropy
0        3.431624     2.845351      0.000000
1        0.000000     0.000000      3.240224
2        3.008695     1.584963      2.913977
3        2.973557     1.918296      5.540696
4        3.748995     3.000000      3.845213
```

[19]:
```
# Token features data
token_feature_df = pd.read_csv('data/transformed/7.token_features_data.csv')

token_feature_df.head()
```

[19]:
```
                                                 url       label  \
0                       https://www.visitcanada.com  legitimate
1                  http://218.228.19.9/~yossi/9ssfpkz    phishing
2  https://www.msupress.msu.edu/series.php?series…  legitimate
3  https://docs.google.com/presentation/d/e/2PACX…    phishing
4  https://www.c250.columbia.edu/c250_celebrates/…  legitimate

   domain_token_count  path_token_count  total_tokens  avg_token_length
0                   3                 0             3          5.666667
1                   0                 1             1          3.666667
2                   4                 2             6          4.500000
3                   3                 4             7          8.882353
4                   4                 4             8          6.200000
```

[20]:
```
# Hexadecimal feature data
hex_feature_df = pd.read_csv('data/transformed/8.hex_features_data.csv')
```

```
hex_feature_df.head()
```

[20]:

	url	label	has_hex	hex_char_count	hex_ratio
0	https://www.visitcanada.com	legitimate	False	0	0.0
1	http://218.228.19.9/~yossi/9ssfpkz	phishing	False	0	0.0
2	https://www.msupress.msu.edu/series.php?series…	legitimate	False	0	0.0
3	https://docs.google.com/presentation/d/e/2PACX…	phishing	False	0	0.0
4	https://www.c250.columbia.edu/c250_celebrates/…	legitimate	False	0	0.0

[21]:
```
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

[22]:
```
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')
```

[23]:
```
for df_name,df in df_dict.items():
    print(df_name)
    null_cols(df)
    print()
```

```
URL components
domain         2283
subdomain     64972
tld            2435
sld            2286
```

```
path          48387
query        214413
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

[24]:
```
domain_features_df[domain_features_df['tld'].isnull()]
```

[24]:

	url	label	tld	tld_len	\
1	http://218.228.19.9/~yossi/9ssfpkz	phishing	NaN	0	
38	http://91.239.24.133:6892	phishing	NaN	0	
249	http://72.230.82.80/ase5.png	phishing	NaN	0	
304	http://185.102.136.127	phishing	NaN	0	
455	http://208.75.241.246:443/msearch.php	phishing	NaN	0	
…	…	…	…	…	
252844	http://78.157.227.34/weds12.pdf	phishing	NaN	0	
252950	http://185.66.10.57/upd/4	phishing	NaN	0	
252966	http://115.29.165.174:25663/s-3.rar	phishing	NaN	0	
252969	http://61.221.169.31/images/kongj.jpg	phishing	NaN	0	
253094	http://91.239.24.216:6892	phishing	NaN	0	

	url_has_ipv4	url_has_port
1	True	False
38	True	True

```
249             True         False
304             True         False
455             True          True
…               …            …
252844          True         False
252950          True         False
252966          True          True
252969          True         False
253094          True          True

[2435 rows x 6 columns]
```

[25]:
```
sld_features_df[sld_features_df['sld'].isnull()]
```

[25]:
```
                                        url     label  sld  sld_len  \
1          http://218.228.19.9/~yossi/9ssfpkz  phishing  NaN        0
38                  http://91.239.24.133:6892  phishing  NaN        0
249              http://72.230.82.80/ase5.png  phishing  NaN        0
304                   http://185.102.136.127  phishing  NaN        0
455     http://208.75.241.246:443/msearch.php  phishing  NaN        0
…                                         …       …  …      …
252844        http://78.157.227.34/weds12.pdf  phishing  NaN        0
252950              http://185.66.10.57/upd/4  phishing  NaN        0
252966    http://115.29.165.174:25663/s-3.rar  phishing  NaN        0
252969  http://61.221.169.31/images/kongj.jpg  phishing  NaN        0
253094              http://91.239.24.216:6892  phishing  NaN        0

        sld_has_digit  sld_has_hyphen  sld_token_count
1               False           False                1
38              False           False                1
249             False           False                1
304             False           False                1
455             False           False                1
…               …               …                    …
252844          False           False                1
252950          False           False                1
252966          False           False                1
252969          False           False                1
253094          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

[26]:
```
df = pd.DataFrame()     # dataframe to store all the processed features
```

[27]:
```
# URL components data

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

[28]:
```
# URL Length features data

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

[29]:
```
# 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')
```

[30]:
```
# 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']
```

[31]:
```
# Character features data

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

[32]:
```
# Entropy features data

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

[33]:
```
# 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.

[34]:
```
# Adding label

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

[35]:
```
df.head()
```

[35]:

	has_https	url_len	domain_len	path_len	query_len	url_depth	\
0	1	27	19	0	0	0	
1	0	34	0	13	0	2	
2	1	51	20	10	11	1	
3	1	175	15	103	43	5	
4	1	79	21	47	0	3	

	subdomain_count	tld_len	url_has_ipv4	url_has_port	…	spl_char_count	\
0	1	3	0	0	…	5	
1	0	0	1	0	…	9	
2	2	3	0	0	…	10	
3	1	3	0	0	…	21	
4	2	3	0	0	…	12	

	url_entropy	domain_entropy	sld_entropy	path_entropy	domain_token_count	\
0	3.856196	3.431624	2.845351	0.000000	3	
1	3.962032	0.000000	0.000000	3.240224	0	
2	3.965393	3.008695	1.584963	2.913977	4	
3	5.569700	2.973557	1.918296	5.540696	3	
4	4.274946	3.748995	3.000000	3.845213	4	

	path_token_count	total_tokens	avg_token_length	class
0	0	3	5.666667	0
1	1	1	3.666667	1
2	2	6	4.500000	0
3	4	7	8.882353	1
4	4	8	6.200000	0

[5 rows x 30 columns]

[36]:
```
print(f'The combined dataset consists of {df.shape[0]} rows and {df.shape[1]}␣
↪columns')
```

```
The combined dataset consists of 253098 rows and 30 columns
```

[37]:
```
df.columns
```

[37]:
```
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')
```

[38]:
```
df.dtypes
```

[38]:
```
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
```

[39]:
```
df.duplicated().sum()
```

[39]:
```
np.int64(29568)
```

[40]:
```
df.drop_duplicates(keep='first',ignore_index=True,inplace=True)
```

[41]:
```
# Saving the dataset

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