In [1]:
```
def topLevelDomain(domain):
    string = domain.split('.')
    return string[len(string)-2]
```

In [2]:
```
import pandas as pd

dataFrame1 = pd.read_csv("top-1m.csv", names=['Domain'])
print (dataFrame1.shape)
dataFrame1.head()
```

```
(1000000, 1)
```

Out[2]:

| | Domain |
|---|---|
| 1 | google.com |
| 2 | microsoft.com |
| 3 | www.google.com |
| 4 | facebook.com |
| 5 | doubleclick.net |

In [3]:
```
domain_names = [topLevelDomain(dataFrame1.iloc[i,0]) for i in range(dataFrame1.shape[0])]
```

In [4]:
```
dataFrame1['Top_Level_Domain'] = domain_names
```

In [5]:
```
dataFrame1
```

Out[5]:

| | Domain | Top_Level_Domain |
|---|---|---|
| 1 | google.com | google |
| 2 | microsoft.com | microsoft |
| 3 | www.google.com | google |
| 4 | facebook.com | facebook |
| 5 | doubleclick.net | doubleclick |
| ... | ... | ... |
| 999996 | www.supermommyclub.com | supermommyclub |
| 999997 | www.supportcentre-rbs.co.uk | co |
| 999998 | www.supratraderonline.com | supratraderonline |
| 999999 | www.supremecourt.ohio.gov | ohio |
| 1000000 | www.suteba.org.ar | org |

1000000 rows × 2 columns

In [6]:
```
import sys
path = "C:\\Users\\sahil\\OneDrive\\Desktop\\Fall 2020\\freq-master"
sys.path.insert(1, path)
from freq import FreqCounter

#create FreqCounter object
fc = FreqCounter()

#load the defualt frequency table
fc.load(path+'/freqtable2018.freq')
```

In [7]:
```
freqScore = [fc.probability(dataFrame1.iloc[i,1])[0] for i in range(dataFrame1.shape[0])]

dataFrame1['Frequency Scores'] = freqScore
```

In [9]:
```
dataFrame1.shape
```

Out[9]:
```
(1000000, 3)
```

In [10]:
```
dataFrame1.iloc[0,1]
```

Out[10]:
```
'google'
```

In [10]:
```
dataFrame1["Frequency Scores"].max()
```

Out[10]:
```
99.8891
```

In [11]:
```
dataFrame1["Frequency Scores"].min()
```

Out[11]:
```
0.0
```

In [12]:
```
dataFrame1["Frequency Scores"].mean()
```

Out[12]:
```
7.535460673504287
```

In [13]:
```
dataFrame1["Frequency Scores"].median()
```

Out[13]:
```
7.0937
```

In [11]:
```
dataFrame1[dataFrame1["Frequency Scores"] < 3.5].drop_duplicates(subset='Top_Level_Domain', keep='first', inplace=False)
```

Out[11]:

| | Domain | Top_Level_Domain | Frequency Scores |
|---|---|---|---|
| 13 | fbcdn.net | fbcdn | 0.1321 |
| 28 | akadns.net | akadns | 2.3848 |
| 38 | ytimg.com | ytimg | 2.9440 |
| 45 | yahoo.com | yahoo | 3.1215 |
| 50 | msn.com | msn | 1.4956 |
| ... | ... | ... | ... |
| 997542 | wotsmqt.com | wotsmqt | 2.7802 |
| 997561 | wpwcnshsq.info | wpwcnshsq | 1.3107 |
| 997562 | wqed.org | wqed | 3.0055 |
| 997585 | wskonnekt.com | wskonnekt | 3.4247 |
| 997592 | wtmmn.com | wtmmn | 0.5644 |

31120 rows × 3 columns

In [23]:
```
fc.probability("CH1833l7")[0]
```

Out[23]:
```
4.9379
```

In [24]:
```
fc.probability("RT4x78")[0]
```

Out[24]:
```
1.8976
```

In [ ]:
```

```