# Lab 7

In [36]:
```
import pandas as pd
import tldextract
import wordsegment
wordsegment.load()
```

## Task 1

In [37]:
```
url_data = pd.read_csv('inputurls.csv')
url_data.head(10)
```

Out[37]:

| | urls |
|---|---|
| 0 | https://youtube.com |
| 1 | https://github.com |
| 2 | https://stackoverflow.com |
| 3 | https://nordvpn.com |
| 4 | https://myanimelist.net |
| 5 | https://amazon.com |
| 6 | https://napavalleytoukraine.org |
| 7 | https://sunflower-ukraine.org |
| 8 | https://theukrainewar.org |
| 9 | https://goalsforukraine.hockey |

## Task 2

In [38]:
```
wordDic ={}
```

```
urls    object
dtype: object
```

## Task 3

In [39]:
```
for index, row in url_data.iterrows():
    url = row['urls']
    ext = tldextract.extract(url)
    domain = ext.domain

    keywords = wordsegment.segment(domain)
    for keyword in keywords:
        keyword = keyword.strip().lower()
        if keyword not in wordDic:
            wordDic[keyword] = 1
        else:
            wordDic[keyword] += 1

print(wordDic)
```

```
{'youtube': 1, 'git': 1, 'hub': 1, 'stack': 1, 'overflow': 1, 'nord': 1, 'vp
n': 1, 'my': 1, 'anime': 1, 'list': 1, 'amazon': 1, 'napa': 1, 'valley': 1,
'to': 1, 'ukraine': 11, 'sunflower': 1, 'the': 1, 'war': 1, 'goals': 1, 'fo
r': 4, 'donation': 1, 'dev': 1, 'clevedon': 1, 'ukrainian': 1, 'support': 1,
'hearts': 1, 'mission': 1, 'pucks': 1, 'people': 1, 'assist': 1}
```

Task 4

In [40]:
```
data_output = pd.DataFrame(list(wordDic.items()), columns=['Keyword', 'Count']
data_output = data_output.sort_values('Count', ascending=False)
data_output.to_csv('output.csv', index=False)
print("Output written to output.csv")
```

```
Output written to output.csv
```