NAMA: Zanuar Rikza Aditiya
NIM: 230411100087
MATA KULIAH: EKSTRAKSI INFORMASI A

Code ini dugunakan untuk mengambil URL halaman detail setiap putusan dari list putusan
parameter yang disiapkan adalah URL posisi awal munculnya list putusan sesuai kriteria yang akan di ambil dan URL halaman terakhir yang memuat list putusan
output dari code ini adalah file dengan nama hasilListURLPage.txt yang berisi list URL untuk melihat detail setiap putusan

In [2]:
```
import requests
import urllib.request
import time
import os
import warnings
from bs4 import BeautifulSoup
```

fungsi getURLfromWeb untuk Mendapatkan url putusan dari website

In [3]:
```
def getURLfromWeb(url):

    response = requests.get(url, verify=False)
    print(response)

    htmlCode1 = BeautifulSoup(response.text, 'html.parser')
    result1=htmlCode1.findAll('a')

    urlHasil=[]
    for eachResult1 in result1:
        cariURLawal=str(eachResult1).find('https://putusan3.mahkamahagung.go.id/direktori/putusan/')
        cariURLakhir=str(eachResult1).find('html">Putusan') + 4
        #print(cariURLakhir)
        if cariURLawal == 9 and cariURLakhir >= 4:
            #print(str(eachResult1)[cariURLawal:cariURLakhir])
            cariURL=str(eachResult1)[cariURLawal:cariURLakhir]
            urlHasil.append(cariURL)

    print(urlHasil)
    return(urlHasil)
```

In [ ]:
```
def main():

    warnings.filterwarnings('ignore')

    data_path = "./data/"
    #url1 = 'https://putusan3.mahkamahagung.go.id/direktori/kategori/jenis/pencurian-1.html'
    #url2 = 'https://putusan3.mahkamahagung.go.id/direktori/kategori/jenis/pencurian-1/page/'

    url1 = 'https://putusan3.mahkamahagung.go.id/direktori/index/pengadilan/pn-pekalongan/kategori/pidana-umum-1.html
    url2 = 'https://putusan3.mahkamahagung.go.id/direktori/index/pengadilan/pn-pekalongan/kategori/pidana-umum-1/page

    listHasil = []
    # Saya ingin mengambil putusan hingga page 50
    ulang = 50

    # Saya menambahkan last index
    # agar saat code gagal karena suatu alasan saya tidak mengulangi lagi dari awal
    if not os.path.exists(data_path + "lastindex.txt"):
        with open(data_path + "lastindex.txt", "w", encoding="UTF-8") as file_index:
            file_index.write(str(19))
    try:
        last_index = int(open("./data/lastindex.txt", "r", encoding="UTF-8").read())
    except:
        last_index = 19

    startTime = time.time()

    # file_hasil = open(data_path + "hasilListURLPage.txt", "w", encoding='UTF8')

    # Scraping di mulai dari index terakhir yang telah di cek + 1
    for i in range(last_index, ulang):
        if i == 0:
            url = url1
            print(1)
        else:
            n=i+1
            url = url2+str(n)+'.html'
            print(2)

        listHasil = getURLfromWeb(url)

        print(listHasil)

        # menambahkan list url ke file hasilListURLPage.txt
        # agar jika gagal tidak berulang dari awal
        with open(data_path + "hasilListURLPage.txt", "a", encoding="UTF-8") as file_hasil:
```

```
            for listURL in listHasil:
                file_hasil.write(listURL+"\n")

        print(last_index)
        # Menyimpan index terakhir yang telah di cek + 1
        # agar kita tahu dari mana kita harus mulai lagi tanpa mengulangi dari awal
        with open(data_path + "lastindex.txt", "w", encoding="UTF-8") as file_index:
            file_index.write(str(i + 1))

    # file_hasil.close()
    endTime = time.time()
    print(listHasil)
    print('Time Processing : ', endTime-startTime, ' Second')

main();
```