Indexing

Anggota Kelompok

Ferdiansyah Dwi Nurcahyo
(20/459265/PA/19926)
Muny Safitri
(22/506376/NUGM/01039)
Raden Mas Garda
(20/459275/PA/19996)

```
from tkinter import N
from turtle import title
import requests
from bs4 import BeautifulSoup

categories =['bisnis', 'hak-asasi-manusia','ilmu-hukum','kekayaan-intelektual','keluarga', 'ketenagakerjaan','perdata','perlindungan-kosumen','pertanahan-properti','pidana','profesi-hukum','start-up-umkm','teknologi']
# categories =['bisnis', 'hak-asasi-manusia','ilmu-hukum']#'kekayaan-intelektual','keluarga', 'ketenagakerjaan','perdata','perlindungan-kosumen','pertanahan-properti','pidana','profesi-hukum','start-up-umkm','teknologi']
kolom1 = []
kolom2 = []
kolom3 = []
kolom4 = []
kolom5 = []
kolom6 = []
kolom7= []
nomor = []
k=1

for j in range(len(categories)):

  for i in range(5):#pengulangan 5 kali halaman setiap kategory
    url = 'https://www.hukumonline.com/klinik/'+categories[j]+'/page/'+str(i)+'/'

    #Membuat Requests
    r = requests.get(url)

    #Hasil Response
    request = r.content
    soup = BeautifulSoup(request, 'html.parser')

    #Extract Element
    category = soup.findAll('span', attrs={'class':'d-inline-flex flex-row justify-content-center align-items-center mx-2 badge badge-primary badge-pill'})
    title = soup.findAll('h2', attrs={'class':'header-6 font-weight-bold text-dark'})
    date = soup.findAll('span',attrs={ 'class':'small text-muted mr-1'})
    mitra_author = soup.findAll('a',attrs={'class':"small mx-2 d-none d-md-inline-block"})

    #crawling link
    temp =[]
    for a in soup.find_all('a', href=True):
      if 'href' in a.attrs:
        temp.append(a.attrs['href']) #menambahkan link ke kolom7
```

```
for a in soup.find_all('a', href=True):
  if 'href' in a.attrs:
    temp.append(a.attrs['href']) #menambahkan link ke kolom7
    # print("Found the URL:", a.attrs['href'])

#menambahkan setiap konten ke kolom masing
count = 0

#menambahkan kolom1,kolom2,kolom4,kolom6 dan nomor
for x in range(0, len(title)):
  count += 1
  kolom1.append(title[x].text.strip())
  kolom2.append(date[x].text.strip())
  kolom4.append(category[x].text.strip())

  #melakukan transformasi pada tanggal
  kolom6.append(date[0].text.strip().split()[2])
  nomor.append(str(k))
  k=k+1

#menambahkan kolom link (kolom7)
for link in range(24,len(temp)-47,4):
  kolom7.append('https://www.hukumonline.com'+temp[link])
  # url2 = temp[link]

  # #Membuat Requests
  # r2 = requests.get(url2)

  # #Hasil Response
  # request2 = r2.content
  # soup2 = BeautifulSoup(request2, 'html.parser')

  #


#menambahkan kolom3 dan kolom4
for y in range(0,len(mitra_author)-1,2):
  kolom3.append(mitra_author[y].text.strip())
  kolom5.append(mitra_author[y+1].text.strip())
  # print("{0}. {1}\n  {2}  \n{3}".format(y,mitra_author[y].text.strip(),mitra_author[y+1].text.strip(),category[x].text.strip()))
  y=y+2
```

```
[ ]  #mengambil judul artikel sebagai dokumen
     array = [] #dipake untuk menghitung frequency nanti
     for i in df.index:
       array.append(df['judul_artikel'][i])
       print("Judul: "+df['judul_artikel'][i])
```

```
Judul: Konversi Utang Jadi Setoran Saham, Ini Caranya
Judul: Tanpa NDA, Bisakah Karyawan Digugat Pelanggaran Rahasia Dagang?
Judul: Langkah Jika Mantan Karyawan Membocorkan Rahasia Perusahaan
Judul: Pinjaman Melebihi Modal Dasar PT, Perlukah Persetujuan RUPS?
Judul: Persamaan Asosiasi dengan Perkumpulan dan Prosedur Pendiriannya
Judul: 7 Alasan Mengapa Investor Asing Berinvestasi di Indonesia
Judul: Fungsi Lembaga Pengelola Investasi dan Bedanya dengan PIP atau BPKM
Judul: Dapatkah Aset Kripto Jadi Alat Pembayaran Utang dalam Kepailitan?
Judul: Biar Tak Diblokir, Ini Cara Pendaftaran PSE Lingkup Privat
Judul: PT Tak Punya IUP Eksplorasi? Ini Sanksi Pidananya
Judul: Perbedaan Layanan Pinjaman Bank Digital dengan Pinjol
Judul: Anak Berkendara Hingga Tabrak Orang, Bagaimana Proses Hukumnya?
Judul: Dapatkah Tender Terbatas Dikategorikan Persekongkolan Tender?
Judul: Izin Usaha dan Kode KBLI Perdagangan Batu Kapur
Judul: Begini Hak Opsi dalam Sewa Guna Usaha (Leasing)
Judul: Jatuh Talak Suami, Ini Hak-hak Istri dan Anak
Judul: Dana Sumbangan Diselewengkan, Dapatkah Yayasan Dibubarkan?
Judul: 3 Perbedaan Mediasi dan Arbitrase
Judul: Bingung Cari Nama PT yang Bagus? Simak Aturannya
Judul: Luas Rumah KPR Berbeda dengan di IMB, Ini Konsekuensinya
Judul: Dapatkah Menyimpangi Isi Homologasi dengan Novasi?
Judul: Pembayaran Mandek, Bagaimana Status Uang dalam Perjanjian Akuisisi?
Judul: Syarat Perusahaan Menjadi Trader Batubara
Judul: Sengketa yang Tak Dapat Diselesaikan Melalui Penyelesaian Sengketa Alternatif
Judul: Kenali Bank Digital di Indonesia dan Syarat Pendiriannya
Judul: Klasifikasi Bahan Baku, Barang Konsumsi, dan Bahan Penolong
Judul: Pemindahan Kepemilikan Perusahaan (Akuisisi) oleh Pemegang Saham
Judul: Haruskah Pengumuman Tender Swasta Melalui Media Massa?
Judul: Skema Pembayaran Utang Debitur Tak Boleh di Luar Homologasi
Judul: Pembayaran ke Kreditur Pasca Homologasi, Haruskah Izin Pengurus?
```

```
[ ] #preprocessing and to lower (tidak menghapus angka)
    punc = '''!()-[]{};:'"\, <>./?@#$%^&*_~'''
    imbuhan = ["peng","pen","pel","per","pe","meng","men","me","ber","ke","di","an",\
               "ke an","me an",
               "pen an","pe an","di kan","kan"]
    imbuhanLebur = ["meny"]
    exception = ["didik","pena","penuh","kepala","kelapa","peluh","dirgahayu","pel","per","beranda",\
                 "menang",""]
    array2 = []
    for char in punc:
      for ele in range(0,len(array)):
        array[ele] = array[ele].replace(char," ").lower()
    print(array)
```

```
['konversi utang jadi setoran saham  ini caranya', 'tanpa nda  bisakah karyawan digugat pelanggaran rahasia dagang ', 'langkah jika mantan karyawan membocorkan rahasia perusahaan', 'pinjaman melebihi modal dasar pt  perlukah persetu
```

```
[ ] from nltk.tokenize import word_tokenize
    import nltk
    from nltk.corpus import stopwords
    nltk.download('stopwords')
    import numpy as np
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
[ ] #tokenisasi
    token = []
    for ele in range(0,len(array)):
      token.append(array[ele].split())
    print((token))
    #output array of array
```

```
[['konversi', 'utang', 'jadi', 'setoran', 'saham', 'ini', 'caranya'], ['tanpa', 'nda', 'bisakah', 'karyawan', 'digugat', 'pelanggaran', 'rahasia', 'dagang'], ['langkah', 'jika', 'mantan', 'karyawan', 'membocorkan', 'rahasia', 'perus
```

```
[ ]  #daftar term(token) berserta id document(dictiniories)
     ele = {}
     kolom1=[]
     kolom2=[]
     index = 1
     for i in range(len(token)):
       for a in token[i]:
         ele[a] = index
         kolom1.append(a)
         kolom2.append(str(index))
         print(a +',' + str(index))#token dan id dokumen
       index = index+1
     # print(ele['karyawan'])
```

```
keluarga,82
di,82
lapas,82
selama,82
pandemi,82
covid,82
19,82
aspek,83
perlindungan,83
data,83
pribadi,83
dalam,83
jasa,83
pengiriman,83
barang,83
hukumnya,84
mewajibkan,84
karyawan,84
share,84
live,84
location,84
di,84
luar,84
jam,84
kerja,84
```