KEMENTERIAN RISET TEKNOLOGI DAN PENDIDIKAN TINGGI

PROGRAM STUDI ILMU KOMPUTER DEPARTEMEN ILMU KOMPUTER DAN ELEKTRONIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS GADJAH MADA

TEMU KEMBALI INFORMASI

Tugas 4 Metode Transformasi Data

DISUSUN OLEH:

ADAM YOGISYAH PUTRA	**20/455439/PA/19654**
MUHAMMAD ARSYA PUTRA	**20/462186/PA/20158**
HIZKYA FIRSTADIPA HARTOKO	**20/455447/PA/19662**

DOSEN:

Dr. Lukman Heryawan, S.T., M.T.

Metode Transformasi Data

I. Metode

Metode yang digunakan pada saat melakukan transformasi data adalah dengan memanfaatkan bahasa pemrograman python dengan library Pandas. Dataset mula-mula kami gabungkan menjadi kesatuan dokumen yang berasal dari hasil web crawling situs halodoc. Kemudian dilakukan cleaning data untuk menghilangkan data dengan nilai kosong. Selanjutnya kami melakukan Case Folding yaitu mengubah seluruh data menjadi huruf kecil agar mempermudah pemrosesan. Selain itu kami juga melakukan Stopwords Removal yang menghilangkan tanda baca yang ada pada data.

II. Proses Transformasi

A. Menyatukan Dataset

```
#import library
import pandas as pd
import numpy as np
import requests
from bs4 import BeautifulSoup, SoupStrainer
                Follow link (ctrl + click)
html = requests.get('https://github.com/HizkiFirst/tki-scraping-obat')
#merge semua data csv
dfs = []
for link in BeautifulSoup(html.text, parse_only=SoupStrainer('a')):
    if hasattr(link, 'href') and link['href'].endswith('.csv'):
        url = 'https://github.com'+link['href'].replace('/blob/', '/raw/')
        dfs.append(pd.read_csv(url, encoding = 'ISO-8859-1'))
df = pd.concat(dfs)
```

B. Drop Kolom dengan nilai null terbanyak

```
#check kategori yang memiliki banyak nilai null
df=df.sort_values('category')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 531 entries, 17 to 3
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   product_name     531 non-null    object
 1   category         531 non-null    object
 2   deskripsi        531 non-null    object
 3   indikasi_umum    531 non-null    object
 4   dosis            531 non-null    object
 5   aturan_pakai     531 non-null    object
 6   komposisi        40 non-null     object
 7   perhatian        20 non-null     object
 8   kontra_indikasi  20 non-null     object
 9   efek_samping     20 non-null     object
 10  segmentasi       20 non-null     object
 11  kemasan          20 non-null     object
 12  manufaktur       20 non-null     object
 13  no_reg           20 non-null     object
dtypes: object(14)
memory usage: 62.2+ KB
```

```
[ ]  #drop kolom dengan jumlah null terbanyak
     df.drop(df.columns[[6,7,8,9,10,11,12,13]], axis=1, inplace=True)
     df.info()
```

C. Case Folding

Case Folding

```
cols = ["product_name","category","deskripsi","indikasi_umum","dosis","aturan_pakai"]
for col in cols:
  df2[col] = df2[col].apply(str.lower)

df2.head()
```

	product_name	category	deskripsi	indikasi_umum	dosis	aturan_pakai
6	cetirizine sirup 60 ml	alergi	cetirizine merupakan antihistamin yang secara...	informasi obat ini hanya untuk kalangan medis...	penggunaan obat ini harus sesuai dengan petun...	dikonsumsi sesudah makan
11	ocuson 10 tablet	alergi	ocuson tablet adalah obat kombinasi kortikost...	informasi obat ini hanya untuk kalangan medis...	penggunaan obat ini harus sesuai dengan petun...	diminum setelah makan.
10	interhistin 50 mg 10 tablet	alergi	interhistin 50 mg tablet merupakan obat denga...	informasi obat ini hanya untuk kalangan medis...	penggunaan obat ini harus sesuai dengan petun...	diminum setelah makan
8	ctm 4 mg 12 tablet	alergi	ctm 4 mg tablet merupakan obat anti alergi ya...	obat ini digunakan untuk mengatasi gejala ale...	dewasa : \r\n1 tablet, diminum 3-4 kali per h...	diberikan bersama atau tanpa makanan
7	dexteem plus 10 tablet	alergi	dexteem plus tablet adalah obat kombinasi kor...	informasi obat ini hanya untuk kalangan medis...	penggunaan obat ini harus sesuai dengan petun...	obat diminum sesudah makan

D. Stopwords Removal

Case Folding

```
cols = ["product_name","category","deskripsi","indikasi_umum","dosis","aturan_pakai"]
for col in cols:
  df2[col] = df2[col].apply(str.lower)

df2.head()
```

	product_name	category	deskripsi	indikasi_umum	dosis	aturan_pakai
6	cetirizine sirup 60 ml	alergi	cetirizine merupakan antihistamin yang secara...	informasi obat ini hanya untuk kalangan medis...	penggunaan obat ini harus sesuai dengan petun...	dikonsumsi sesudah makan
11	ocuson 10 tablet	alergi	ocuson tablet adalah obat kombinasi kortikost...	informasi obat ini hanya untuk kalangan medis...	penggunaan obat ini harus sesuai dengan petun...	diminum setelah makan.
10	interhistin 50 mg 10 tablet	alergi	interhistin 50 mg tablet merupakan obat denga...	informasi obat ini hanya untuk kalangan medis...	penggunaan obat ini harus sesuai dengan petun...	diminum setelah makan
8	ctm 4 mg 12 tablet	alergi	ctm 4 mg tablet merupakan obat anti alergi ya...	obat ini digunakan untuk mengatasi gejala ale...	dewasa : \r\n1 tablet, diminum 3-4 kali per h...	diberikan bersama atau tanpa makanan
7	dexteem plus 10 tablet	alergi	dexteem plus tablet adalah obat kombinasi kor...	informasi obat ini hanya untuk kalangan medis...	penggunaan obat ini harus sesuai dengan petun...	obat diminum sesudah makan

E. Stopwords Removal

Stopwords Removal

```
import string

#cols = ["category","deskripsi","indikasi_umum","dosis","aturan_pakai"]
#df2["deskripsi"] = df2["deskripsi"].str.replace('[{}]'.format(string.punctuation), '')

cols = ["product_name","category","deskripsi","indikasi_umum","dosis","aturan_pakai"]
for col in cols:
  df2[col] = df2[col].str.replace('[{}]'.format(string.punctuation), ' ')
```

III. Hasil

Hasil yang diperoleh adalah dataset yang sudah siap untuk diproses dan tidak memiliki nilai yang kosong sehingga mempermudah pemrosesan.

Link dataset: https://github.com/HizkiFirst/tki-scraping-obat/blob/main/clean-dataset.csv

IV. Lampiran

Repository: https://github.com/HizkiFirst/tki-scraping-obat

Source code: https://colab.research.google.com/drive/1-30ycmppxsO7dzDZlyYAMmsOFjqIl6M9?usp=sharing