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

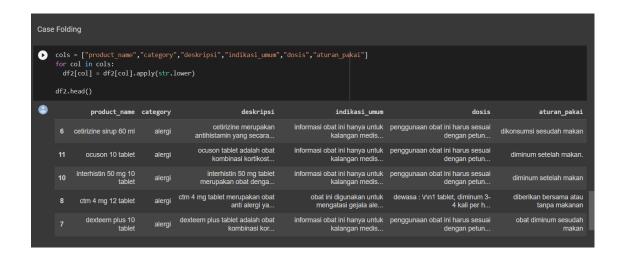
B. Drop Kolom dengan nilai null terbanyak

```
df=df.sort_values('category')
df.info()
Int64Index: 531 entries, 17 to 3
Data columns (total 14 columns):
# Column
                       Non-Null Count Dtype
    product_name
     category
     deskripsi
     indikasi_umum
 4 dosis
                        531 non-null
                        531 non-null
                                          object
    komposisi 40 non-null
perhatian 20 non-null
kontra_indikasi 20 non-null
                        40 non-null
                                          object
                                          object
    efek_samping
                        20 non-null
                                          object
 10 segmentasi
 11 kemasan
 12 manufaktur
                        20 non-null
                                          object
dtypes: object(14)
memory usage: 62.2+ KB
df.drop(df.columns[[6,7,8,9,10,11,12,13]], axis=1, inplace=True)
df.info()
```

C. Case Folding



D. Stopwords Removal



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/1-30ycmppxs07dzDZlyYAMmsOFjqIl6M9?usp=sharing