

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 6 Metode dan Hasil Kompresi Indeks

Sistem Rekomendasi Obat Berdasarkan Gejala



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.

I. Metode

Metode yang digunakan untuk kompresi indeks sistem rekomendasi obat berdasarkan gejala adalah dengan metode Dictionary as a String, yaitu mengompres term list dengan menggabungkan semua term menjadi satu string. Dengan tidak menyimpan term pada setiap kolom di term, kita dapat menghemat file dari dictionary. Metode Dictionary as a String ini termasuk dalam metode Lossless compression karena tidak ada informasi yang terbuang. Selain itu, metode Lossy compression juga digunakan dengan cara membuang semua term angka karena term tersebut tidak digunakan pada sistem rekomendasi obat.

II. Proses Kompresi Indeks

1. Menghilangkan data yang tidak perlu

▼ Import Library dan Dataset

```
import pandas as pd
import numpy as np

url = 'https://raw.githubusercontent.com/HizkiFirst/tki-scraping-obat/main/inverted-index-dataset.csv'
df = pd.read_csv(url)
df.head(5)
```

	term	docFrequency	postingList
0	0	123	[47, 172, 210, 63, 155, 51, 51, 63, 63, 53, 32...
1	00	1	[139]
2	000	9	[309, 309, 309, 21, 21, 309, 21, 21, 314]
3	03	2	[204, 215]
4	035	1	[202]

▼ Penghapusan term berupa bilangan

▼ Mencari index term huruf pertama

```
df[df['term'] >= 'a']
```

	term	docFrequency	postingList
113	a	8	[273, 344, 344, 273, 45, 18, 276, 29]
114	aa	2	[421, 421]
115	abdomen	1	[381]
116	abnormal	6	[252, 250, 254, 250, 252, 254]
117	aborsi	4	[210, 210, 210, 210]
...
3972	zinkid	2	[385, 385]
3973	zn	2	[160, 146]
3974	zollinger	4	[388, 388, 390, 390]
3975	zone	4	[342, 384, 363, 387]
3976	zoster	2	[233, 233]

3864 rows × 3 columns

▼ Mengambil data dengan term huruf saja

```
[ ] df_clean = df.iloc[113:]
```

```
[ ] df_clean = df_clean.reset_index(drop=True)
```

```
[ ] df_clean
```

	term	docFrequency	postingList
0	a	8	[273, 344, 344, 273, 45, 18, 276, 29]
1	aa	2	[421, 421]
2	abdomen	1	[381]
3	abnormal	6	[252, 250, 254, 250, 252, 254]
4	aborsi	4	[210, 210, 210, 210]
...
3859	zinkid	2	[385, 385]
3860	zn	2	[160, 146]
3861	zollinger	4	[388, 388, 390, 390]
3862	zone	4	[342, 384, 363, 387]
3863	zoster	2	[233, 233]

3864 rows × 3 columns

2. Pembuatan dictionary

▼ Pembuatan Dictionary

```
[ ] df_dict = pd.DataFrame(columns=['Freq','Posting_ptr','Term_ptr'])
```

```
[ ] df_dict
```

	Freq	Posting_ptr	Term_ptr
--	------	-------------	----------

```
[ ] df_dict['Freq'] = df_clean['docFrequency']
```

```
[ ] df_dict['Posting_ptr'] = df_clean['postingList']
```

```
[ ] df_dict
```

	Freq	Posting_ptr	Term_ptr
0	8	[273, 344, 344, 273, 45, 18, 276, 29]	NaN
1	2	[421, 421]	NaN
2	1	[381]	NaN
3	6	[252, 250, 254, 250, 252, 254]	NaN
4	4	[210, 210, 210, 210]	NaN
...
3859	2	[385, 385]	NaN
3860	2	[160, 146]	NaN

3. Penggabungan semua term menjadi 1 term string dan menentukan term pointer

▼ Menentukan Term Pointer dan Membuat Term String

```
[ ] term_str = ""
term_counter = 0
for i in range(3864):
    term = df_clean['term'].iloc[i]
    term_str += term
    df_dict['Term_ptr'].iloc[i] = term_counter
    term_counter += len(term)
```

/usr/local/lib/python3.7/dist-packages/pandas/core/indexing.py:1732: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
self._setitem_single_block(indexer, value, name)

```
[ ] term_str
```

```
'aaaabdomenabnormalaborsiabsesabsolutabsopsiabsorpsinyaaacetateacetoneideacetylcycteineacetylsalicylicacidnacareacneacneactacquireacsbpactifedactingactionacycloviradadaadalahadany  
aadekuatadenosinadenosinaadenosineadhesiadhesiveadidasadipositadpadrenergicadrenergikadrenoceptoradrenokortikaladrenoreceptorsadrenoreseptoradukadvanceaereusaerobaerogenesaeruginosaaf  
ricanumafetosagaakagalactiaeagaragenagentagonisagonistagregasiahliainsairakanakarakhirakhirnyaakibatakilenakneaksiaktifaktifitasaktivasiaktivatoraktivitaasaktivitasaktivitas  
inataakurasiakuratakutalamalamlamialamiahalasatalbicansalbothylalcoholaldosteronaldosteronealergenalerghalfaalialiranalkalosisalkoholallergyallorisaloclairaloealparaalphaalternatifalte  
rnatifnyaaluminiumalumuniumamanambillambroxolamenoreamenoreeaminefronaminoaminoacylaminoglikosidaaminopenisillinamlodipinamlodipineammoniumamoebiasisisamoksisillinamoxicillinamoxiclavamoxs  
anampulanaerobanafilaksisanakanalgesikanalgetikanaloganalognyaanastesiandaandalananemi...'
```



df_dict



	Freq	Posting_ptr	Term_ptr
0	8	[273, 344, 344, 273, 45, 18, 276, 29]	0
1	2	[421, 421]	1
2	1	[381]	3
3	6	[252, 250, 254, 250, 252, 254]	10
4	4	[210, 210, 210, 210]	18
...
3859	2	[385, 385]	29481
3860	2	[160, 146]	29487
3861	4	[388, 388, 390, 390]	29489
3862	4	[342, 384, 363, 387]	29498
3863	2	[233, 233]	29502

3864 rows × 3 columns

4. Export Dictionary ke CSV dan Term

▼ Export Hasil Dataset dan Term String

```
[ ] df_dict.to_csv('compressed-dictionary-dataset.csv', index=False)
```



```
text_file = open("term.txt", "w")  
text_file.write(term_str)  
text_file.close()
```

III. Hasil

Link Dataset Hasil Kompresi Indeks:

https://github.com/HizkiFirst/tki-scraping-obat/blob/kumpul/tugas-6_7/compressed-dictionary-dataset.csv

Link Term String:

https://github.com/HizkiFirst/tki-scraping-obat/blob/kumpul/tugas-6_7/term.txt

IV. Lampiran

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

Link Google Collab Source Code:

<https://colab.research.google.com/drive/1LLpyk1KC3retW3YzhQMabFoTOVaYRGPG?usp=sharing>