

Nama : Virgie Yunita Salsabil

NIM : 21110022

Kelas : S1-SD02A

RESPONSI

✓ 1. CPMK1- Sub CPMK1.1 (bobot : 10)

- Lakukan crawling data teks dari media sosial/ web dan simpan hasilnya dalam bentuk excel/csv.
- Setiap mahasiswa harus melakukan crawling dengan kata kunci tertentu.
- Kata kunci tidak boleh sama dengan mahasiswa lainnya.

```
1 # Instalasi pustaka nltk (Natural Language Toolkit) utk pemrosesan bahasa alami.  
2 !pip install nltk
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-packages (3.8.1)  
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packages (from nltk) (8.1.7)  
Requirement already satisfied: joblib in /usr/local/lib/python3.10/dist-packages (from nltk) (1.3.2)  
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.10/dist-packages (from nltk) (2023.6.3)  
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from nltk) (4.66.1)
```

```
1 # Instalasi pustaka sastrawi utk pemrosesan b.indo  
2 !pip install Sastrawi
```

```
Collecting Sastrawi  
  Downloading Sastrawi-1.0.1-py2.py3-none-any.whl (209 kB)  
    ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 209.7/209.7 kB 2.0 MB/s eta 0:00:00  
Installing collected packages: Sastrawi  
Successfully installed Sastrawi-1.0.1
```

```
1 # Instalasi paket pandas  
2 !pip install pandas
```

```
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (1.5.3)  
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.10/dist-packages (from pandas) (2.8.2)  
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas) (2023.3.post1)  
Requirement already satisfied: numpy>=1.21.0 in /usr/local/lib/python3.10/dist-packages (from pandas) (1.23.5)  
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.1->pandas) (1.16.0)
```

```
1 # Mengimpor pustaka pandas dg alias 'pd' utk manipulasi dan analisis data.  
2 import pandas as pd  
3  
4 # Mengimpor modul re (regular expression) untuk operasi pemrosesan string.  
5 import re  
6 import nltk  
7  
8 # Mengimpor modul word_tokenize dari nltk untuk pemisahan kata.  
9 from nltk.tokenize import word_tokenize  
10  
11 # Mengimpor modul stopwords dari nltk utk mendapatkan kata-kata yang umumnya diabaikan.  
12 from nltk.corpus import stopwords  
13  
14 # Mengimpor kelas StemmerFactory dari pustaka Sastrawi. Sastrawi adalah pustaka untuk pemrosesan bahasa Indonesia.  
15 from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
```

```

1 # Import required Python package
2
3 # Install Node.js (because tweet-harvest built using Node.js)
4 !sudo apt-get update
5 # Instal paket yg diperlukan utk manajemen sertifikat, pengunduhan, & dukungan GPG.
6 !sudo apt-get install -y ca-certificates curl gnupg
7 # Membuat direktori utk menyimpan kunci GPG yg digunakan oleh Node.js.
8 !sudo mkdir -p /etc/apt/keyrings
9 # Instal kunci GPG dari Node.js & simpan dlm bentuk yg dpt digunakan oleh apt.
10 !curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key | sudo gpg --dearmor -o /etc/apt/keyrings/nodesource.gpg
11
12 # Menetapkan variabel lingkungan NODE_MAJOR ke 20 & menambahkan repository Node.js ke daftar sumber apt.
13 !NODE_MAJOR=20 && echo "deb [signed-by=/etc/apt/keyrings/nodesource.gpg] https://deb.nodesource.com/node_${NODE_MAJOR}.x nodistro main"
14
15 # Melakukan pembaruan lg setelah menambahkan sumber baru.
16 !sudo apt-get update
17 # Menginstal Node.js dg opsi -y utk menyetujui otomatis semua permintaan unduhan.
18 !sudo apt-get install nodejs -y
19
20 # Menampilkan versi Node.js yg telah diinstal.
21 !node -v

```

```

Hit:1 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
Hit:2 http://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:3 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64 InRelease
Hit:4 https://deb.nodesource.com/node_20.x nodistro InRelease
Hit:5 http://archive.ubuntu.com/ubuntu jammy InRelease
Hit:6 http://archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:7 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:8 https://ppa.launchpadcontent.net/c2d4u.team/c2d4u4.0+/ubuntu jammy InRelease
Hit:9 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:10 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
Hit:11 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Reading package lists... Done
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
ca-certificates is already the newest version (20230311ubuntu0.22.04.1).
curl is already the newest version (7.81.0-1ubuntu1.15).
gnupg is already the newest version (2.2.27-3ubuntu2.1).
0 upgraded, 0 newly installed, 0 to remove and 40 not upgraded.
gpg: cannot open '/dev/tty': No such device or address
curl: (23) Failed writing body
deb [signed-by=/etc/apt/keyrings/nodesource.gpg] https://deb.nodesource.com/node_20.x nodistro main
Hit:1 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
Hit:2 http://archive.ubuntu.com/ubuntu jammy InRelease
Hit:3 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64 InRelease
Hit:4 http://archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:5 https://deb.nodesource.com/node_20.x nodistro InRelease
Hit:6 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:7 http://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:8 https://ppa.launchpadcontent.net/c2d4u.team/c2d4u4.0+/ubuntu jammy InRelease
Hit:9 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:10 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
Hit:11 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Reading package lists... Done
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
nodejs is already the newest version (20.10.0-1nodesource1).
0 upgraded, 0 newly installed, 0 to remove and 40 not upgraded.
v20.10.0

```

✓ Menggunakan tweet-harvest untuk melakukan crawling data dari Twitter

- -o: Nama file output untuk menyimpan hasil crawling (diberikan nilai dari variabel filename)
- -s: Kata kunci pencarian di Twitter (diberikan nilai dari variabel search_keyword)
- -l: Batas jumlah tweet yang akan diambil (diberikan nilai dari variabel limit)

Cara Melihat AUTH TOKEN:

- Buka laman tweeter akun kita.
- Klik kanan > Inspect > Application > Drop-down 'Cookies' > Cari Auth Token

```

1 # Crawl Data
2
3 # Nama file untuk menyimpan hasil crawling
4 filename = '1.Crawling_Data.csv'
5
6 # Kata kunci pencarian di Twitter
7 search_keyword = 'skincare'
8
9 # Batas jumlah tweet yang akan diambil
10 limit = 200
11
12 !npx --yes tweet-harvest@latest -o "{filename}" -s "{search_keyword}" -l {limit}

```

Welcome to the Twitter Crawler 🐦

This script uses Chromium Browser to crawl data from Twitter with *your* Twitter auth token.
Please enter your Twitter auth token when prompted.

Note: Keep your access token secret! Don't share it with anyone else.

Note: This script only runs on your local device.

? What's your Twitter auth token? > 788? What's your Twitter auth token? > *788? What's your Twitter auth token? > **788? What

up to date, audited 4 packages in 563ms

found 0 vulnerabilities

Installing dependencies...

```

Hit:1 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
Hit:2 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64 InRelease
Hit:3 https://deb.nodesource.com/node_20.x nodistro InRelease
Hit:4 http://archive.ubuntu.com/ubuntu jammy InRelease
Hit:5 http://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:6 http://archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:7 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:8 https://ppa.launchpadcontent.net/c2d4u.team/c2d4u4.0+/ubuntu jammy InRelease
Hit:9 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:10 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
Hit:11 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Reading package lists... Done
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done

```

```

fonts-freefont-ttf is already the newest version (20120503-10build1).
fonts-liberation is already the newest version (1:1.07.4-11).
libasound2 is already the newest version (1.2.6.1-1ubuntu1).
libatk-bridge2.0-0 is already the newest version (2.38.0-3).
libatk1.0-0 is already the newest version (2.36.0-3build1).
libatspi2.0-0 is already the newest version (2.44.0-3).
libcairo2 is already the newest version (1.16.0-5ubuntu2).
libfontconfig1 is already the newest version (2.13.1-4.2ubuntu5).
libnsspr4 is already the newest version (2:4.32-3build1).
libxcb1 is already the newest version (1.14-3ubuntu3).
libxcomposit1 is already the newest version (1:0.4.5-1build2).
libxdamage1 is already the newest version (1:1.1.5-2build2).
libxext6 is already the newest version (2:1.3.4-1build1).
libxf86vm0 is already the newest version (1:6.0.0-1).
libxkbcommon0 is already the newest version (1.4.0-1).
libxrandr2 is already the newest version (2:1.5.2-1build1).
xfonts-scalable is already the newest version (1:1.0.3-1.2ubuntu1).
fonts-ipafont-gothic is already the newest version (00303-21ubuntu1).
fonts-tlwg-loma-otf is already the newest version (1:0.7.3-1).
fonts-unifont is already the newest version (1:14.0.01-1).
fonts-wqy-zenhei is already the newest version (0.9.45-8).
xfonts-cyrillic is already the newest version (1:1.0.5).
fonts-noto-color-emoji is already the newest version (2.042-0ubuntu0.22.04.1).
libcups2 is already the newest version (2.4.1op1-1ubuntu4.7).
libdbus-1-3 is already the newest version (1.12.20-2ubuntu4.1).
libdrm2 is already the newest version (2.4.113-2~ubuntu0.22.04.1).
libfreetype6 is already the newest version (2.11.1+dfsg-1ubuntu0.2).

```



```

1 # Specify the path to your CSV file
2 file_path = f"tweets-data/{filename}"
3
4 # Read the CSV file into a pandas DataFrame
5 data = pd.read_csv(file_path, delimiter=",")
6
7 # Display the DataFrame
8 display(data)

```

| | created_at | id_str | full_text | quote_count | reply_count | retwee |
|-----|--------------------------------|---------------------|--|-------------|-------------|--------|
| 0 | Fri Dec 29 02:44:14 +0000 2023 | 1740564352913330435 | Jelang akhir tahun ini aku sempet istirahatin ... | 0 | 8 | |
| 1 | Fri Dec 29 06:32:19 +0000 2023 | 1740621752894705726 | hayuukk giveaway skincare rombongan akhir tahu... | 8 | 314 | |
| 2 | Fri Dec 29 09:08:47 +0000 2023 | 1740661129968484849 | Sometimes I frustrated betul bila PMS datang ... | 0 | 1 | |
| 3 | Sun Dec 24 11:41:55 +0000 2023 | 1738887726362177614 | Skincare edisi holiday season!💙💗 kalo lagi li... | 1 | 11 | |
| 4 | Fri Dec 22 13:29:44 +0000 2023 | 1738190082425950706 | 🌟Giveaway Skincare🌟 Untuk 1 orang pemenang. Ca... | 10 | 184 | |
| ... | ... | ... | ... | ... | ... | |
| 214 | Fri Dec 29 09:04:21 +0000 2023 | 1740660013314765103 | Ubel ada ide urebranding imej @pepengbison Gi... | 0 | 11 | |
| 215 | Fri Dec 29 07:00:00 +0000 2023 | 1740628719285203275 | Start from now! 2023 YEAR END SALE 🌟 29-31 Des... | 0 | 0 | |
| 216 | Tue Dec 26 15:14:45 +0000 2023 | 1739666063347073450 | Butuh rekomendasi face moisturizer yang ga mas... | 0 | 17 | |
| 217 | Tue Dec 26 09:38:43 +0000 2023 | 1739581497731137854 | Dr semua bisnis kecil-kecilan yg pernah aku co... | 1 | 3 | |
| | Fri Dec 29 | | 🧐 : What do | | | |

```
1 # Memilih hanya dua kolom "username" dan "full_text"
2 selected_columns = ['username', 'full_text']
3 subset_data = data[selected_columns]
4
5 # Menampilkan DataFrame hasil seleksi
6 display(subset_data)
```

| | username | full_text |  |
|----------------------|---------------|--|---|
| 0 | Far_Rida_ | Jelang akhir tahun ini aku sempet istirahatin ... |  |
| 1 | hmtly | hayuukk giveaway skincare rombongan akhir tahu... | |
| 2 | umiizani | Sometimes I frustrated betul bila PMS datang ... | |
| 3 | irapersa | Skincare edisi holiday season!💙💗 kalo lagi li... | |
| 4 | _mutiaraaaf | 🌟Giveaway Skincare🌟 Untuk 1 orang pemenang. Ca... | |
| ... | ... | ... | |
| 214 | ubel_bluebell | Ubel ada ide urebranding imej @pepengbison Gi... | |
| 215 | Adara_ID | Start from now! 2023 YEAR END SALE 🌟 29-31 Des... | |
| 216 | todayis_pat | Butuh rekomendasi face moisturizer yang ga mas... | |
| 217 | amysorayaa | Dr semua bisnis kecil-kecilan yg pernah aku co... | |
| 218 | msbb_id | 🧐 : What do you want in 2024? 😬 : Alis teb... | |
| 219 rows × 2 columns | | | |

✓ 2. CPMK1- Sub CPMK1.1. (bobot : 25)

- Lakukan pembersihan meliputi penghapusan punctuation, angka dan karakter yang tidak penting menggunakan menggunakan regex.
- Simpan hasilnya menjadi file csv/excel.

```
1 # Membersihkan kolom full_text
2 def clean_text(text):
3     # Menghapus semua karakter selain huruf (a-z, A-Z) dan spasi dari teks.
4     cleaned_text = re.sub(r'^a-zA-Z\s', '', text)
5     return cleaned_text
6
7 # Mengaplikasikan fungsi clean_text ke kolom full_text dan menambahkan hasilnya sebagai kolom baru 'cleaned_text'
8 subset_data['cleaned_text'] = subset_data['full_text'].apply(clean_text)
9
10 # Tampilkan dataset
11 subset_data
```

<ipython-input-32-ba80c261b5a2>:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stable/u>
subset_data['cleaned_text'] = subset_data['full_text'].apply(clean_text)

| | username | full_text | cleaned_text | |
|-----|---------------|--|---|--|
| 0 | Far_Rida_ | Jelang akhir tahun ini aku sempet istirahatn ... | Jelang akhir tahun ini aku sempet istirahatn ... | |
| 1 | hmtly | hayuukk giveaway skincare rombongan akhir tahu... | hayuukk giveaway skincare rombongan akhir tahu... | |
| 2 | umiizani | Sometimes I frustrated betul bila PMS datang ... | Sometimes I frustrated betul bila PMS datang ... | |
| 3 | irapersa | Skincare edisi holiday season! 🧡💕 kalo lagi li... | Skincare edisi holiday season kalo lagi libur... | |
| 4 | _mutiaraaaf | 🌟Giveaway Skincare🌟 Untuk 1 orang pemenang. Ca... | Giveaway Skincare Untuk orang pemenang Carany... | |
| ... | ... | ... | ... | |
| 214 | ubel_bluebell | Ubel ada ide urebranding imej @pepengbison Gi... | Ubel ada ide urebranding imej pepengbison Gim... | |
| 215 | Adara_ID | Start from now! 2023 YEAR END SALE 🌟 29-31 Des... | Start from now Desember Beli skincare ... | |

```
1 # Membersihkan kolom username
2 def clean_username(username):
3     # Menghapus semua karakter selain huruf (a-z, A-Z) dan spasi dari username.
4     cleaned_username = re.sub(r'^a-zA-Z\s', '', username)
5     return cleaned_username
6
7 # Mengaplikasikan fungsi clean_username ke kolom username dan menambahkan hasilnya sebagai kolom baru 'cleaned_username'
8 subset_data['cleaned_username'] = subset_data['username'].apply(clean_username)
9
10 # Menampilkan DataFrame yang sudah dimodifikasi
11 subset_data
```

```
<ipython-input-33-76f12a761967>:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/u](https://pandas.pydata.org/pandas-docs/stable/user_guide/1d_ops.html)
 subset_data['cleaned_username'] = subset_data['username'].apply(clean_username)

```
1 # Membuat DataFrame baru tanpa kolom 'full_text' dan 'username'
2 CLEAN = subset_data.drop(['full_text', 'username'], axis=1)
3
4 # Menampilkan DataFrame baru yang telah dibuat
5 CLEAN
```

| | cleaned_text | cleaned_username |
|-----|---|------------------|
| 0 | Jelang akhir tahun ini aku sempat istirahatin ... | FarRida |
| 1 | hayuukk giveaway skincare rombongan akhir tahu... | hmtly |
| 2 | Sometimes I frustrated betul bila PMS datang ... | umiizani |
| 3 | Skincare edisi holiday season kalo lagi libur... | irapersa |
| 4 | Giveaway Skincare Untuk orang pemenang Carany... | mutiaraaaf |
| ... | ... | ... |
| 214 | Ubel ada ide urebranding imej pepengbison Gim... | ubelbluebell |
| 215 | Start from now Desember Beli skincare ... | AdaraID |
| 216 | Butuh rekomendasi face moisturizer yang ga mas... | todayispat |
| 217 | Dr semua bisnis kecilkecilan yg pernah aku cob... | amysorayaa |
| 218 | What do you want in Alis tebal dan bulu m... | msbbid |

219 rows × 2 columns

```
1 # Menyimpan DataFrame CLEAN ke dalam file CSV dengan nama '2.Data_Cleaned.csv' tanpa menyertakan indeks
2 CLEAN.to_csv('2.Data_Cleaned.csv', index=False)
```

✓ 3. CPMK1- Sub CPMK1.2.

a. Lakukan parsing dan simpan hasilnya dalam bentuk excel/csv (bobot:10)

```
1 # Mengimpor pustaka nltk
2 import nltk
3
4 # Mengunduh "punkt" dari nltk
5 nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Unzipping tokenizers/punkt.zip.
True
```

```
1 # Membaca file CSV dengan nama '2.Data_Cleaned.csv' ke dalam DataFrame 'Data_3A'
2 Data_3A = pd.read_csv('2.Data_Cleaned.csv')
3
4 # Menampilkan DataFrame 'Data_3A'
5 Data_3A
```

```

1 # Menggunakan metode word_tokenize dari nltk untuk memecah teks pada kolom 'cleaned_text' menjadi token (kata-kata)
2 Data_3A['parse_text'] = Data_3A['cleaned_text'].apply(word_tokenize)
3
4 # Menggunakan metode word_tokenize dari nltk untuk memecah teks pada kolom 'cleaned_username' menjadi token (kata-kata)
5 Data_3A['parse_name'] = Data_3A['cleaned_username'].apply(word_tokenize)
6
7 # Menampilkan DataFrame 'Data_3A' yang telah dimodifikasi dengan tambahan kolom 'parse_text' dan 'parse_name'
8 Data_3A

```

| | cleaned_text | cleaned_username | parse_text | parse_name |
|-----|---|------------------|---|----------------|
| 0 | Jelang akhir tahun ini aku sempet istirahatin ... | FarRida | [Jelang, akhir, tahun, ini, aku, sempet,istir... | [FarRida] |
| 1 | hayuukk giveaway skincare rombongan akhir tahu... | hmtly | [hayuukk, giveaway, skincare, rombongan, akhir... | [hmtly] |
| 2 | Sometimes I frustrated betul bila PMS datang ... | umiizani | [Sometimes, I, frustrated, betul, bila, PMS, d... | [umiizani] |
| 3 | Skincare edisi holiday season kalo lagi libur... | irapersa | [Skincare, edisi, holiday, season, kalo, lagi,... | [irapersa] |
| 4 | Giveaway Skincare Untuk orang pemenang Carany... | mutiaraaaf | [Giveaway, Skincare, Untuk, orang, pemenang, C... | [mutiaraaaf] |
| ... | ... | ... | ... | ... |
| 214 | Ubel ada ide urebranding imej pepengbison Gim... | ubelbluebell | [Ubel, ada, ide, urebranding, imej, pepengbiso... | [ubelbluebell] |
| 215 | Start from now Desember Beli skincare ... | AdaraID | [Start, from, now, Desember, Beli, skincare, h... | [AdaraID] |
| 216 | Butuh rekomendasi face moisturizer yang ga mas... | todayispat | [Butuh, rekomendasi, face, moisturizer, yang, ... | [todayispat] |
| 217 | Dr semua bisnis kecilkecilan yg pernah aku cob... | amysorayaa | [Dr, semua, bisnis, kecilkecilan, yg, pernah, ... | [amysorayaa] |
| 218 | What do you want in Alis tebal dan bulu m... | msbbid | [What, do, you, want, in, Alis, tebal, dan, bu... | [msbbid] |

219 rows × 4 columns

```

1 # Data parsing
2 data_parsing = Data_3A.drop(['cleaned_text', 'cleaned_username'], axis = 1)
3
4 # Tampilkan data
5 data_parsing

```

| | parse_text | parse_name |
|-----|---|----------------|
| 0 | [Jelang, akhir, tahun, ini, aku, sempet,istir... | [FarRida] |
| 1 | [hayuukk, giveaway, skincare, rombongan, akhir... | [hmtly] |
| 2 | [Sometimes, I, frustrated, betul, bila, PMS, d... | [umiizani] |
| 3 | [Skincare, edisi, holiday, season, kalo, lagi,... | [irapersa] |
| 4 | [Giveaway, Skincare, Untuk, orang, pemenang, C... | [mutiaraaaf] |
| ... | ... | ... |
| 214 | [Ubel, ada, ide, urebranding, imej, pepengbiso... | [ubelbluebell] |
| 215 | [Start, from, now, Desember, Beli, skincare, h... | [AdaraID] |
| 216 | [Butuh, rekomendasi, face, moisturizer, yang, ... | [todayispat] |
| 217 | [Dr, semua, bisnis, kecilkecilan, yg, pernah, ... | [amysorayaa] |
| 218 | [What, do, you, want, in, Alis, tebal, dan, bu... | [msbbid] |

219 rows × 2 columns

```

1 # Menyimpan DataFrame 'data_parsing' ke file CSV dengan nama '3A.Data_Parsing.csv' tanpa menyertakan indeks
2 data_parsing.to_csv('3A.Data_Parsing.csv', index=False)

```

✓ b. Carilah kata-kata slangword yang ada dalam dataset Anda, dengan cara mencocokkan dengan kamus KBBI (terlampir).

- Simpan hasilnya dalam bentuk csv/excel.
- Tampilkan 100 kata slang yang Anda dapatkan dan tampilkan dalam bentuk Gunakan (bobot:25)

```

1 # Membaca file CSV yang berisi kata-kata slang
2 slang = pd.read_csv('stopwords.csv')
3
4 # Tampilkan data
5 slang

```

| Slangwords | |
|------------|---------|
| 0 | Adam |
| 1 | Aga |
| 2 | Agustus |
| 3 | Ahad |
| 4 | Ajisaka |
| ... | ... |
| 47295 | zulu |
| 47296 | zurafah |
| 47297 | zuriah |
| 47298 | zuriat |
| 47299 | zus |

```
1 # Mengambil daftar kata-kata slang dari kolom 'Slangwords' dalam DataFrame 'slang' dan mengonversikannya menjadi list
2 slang_words = slang['Slangwords'].tolist()
3
4 # Fungsi untuk mencari kata-kata slang dalam teks
5 def find_slang(text):
6     # Mengambil token (kata-kata) dari teks
7     words = text
8
9     # Mencari kata-kata slang yang ada dalam teks
10    found_slang = [word for word in words if word in slang_words]
11
12    # Menggabungkan kata-kata slang yang ditemukan menjadi satu string, dipisahkan oleh koma
13    return ', '.join(found_slang)
14
15 # Menerapkan fungsi find_slang ke kolom 'parse_text' dalam DataFrame 'data_parsing' dan menambahkan hasilnya sebagai kolom baru 'found_slang_in_text'
16 data_parsing['found_slang_in_text'] = data_parsing['parse_text'].apply(find_slang)
17
18 # Menampilkan DataFrame 'data_parsing' yang telah dimodifikasi dengan tambahan kolom 'found_slang_in_text'
19 data_parsing
```

| | parse_text | parse_name | found_slang_in_text |
|-----|---|----------------|---|
| 0 | [Jelang, akhir, tahun, ini, aku, sempet, istir... | [FarRida] | akhir, tahun, ini, aku, kulit, wajah, aku, dar... |
| 1 | [hayuukk, giveaway, skincare, rombongan, akhir... | [hmtly] | rombongan, akhir, tahun, mau, nomor, berapa, p... |
| 2 | [Sometimes, I, frustrated, betul, bila, PMS, d... | [umiizani] | betul, bila, datang, tak, pernah, bagi, nak, r... |
| 3 | [Skincare, edisi, holiday, season, kalo, lagi,... | [irapersa] | edisi, kalo, lagi, liburan, biasanya, bawa, mu... |
| 4 | [Giveaway, Skincare, Untuk, orang, pemenang, C... | [mutiaraaaf] | orang, pemenang, gampang, aku, domisili, kamu,... |
| ... | ... | ... | ... |
| 214 | [Ubel, ada, ide, urebranding, imej, pepengbiso... | [ubelbluebell] | ada, ide, kalo, personal, jadi, yang, kaya, ba... |
| 215 | [Start, from, now, Desember, Beli, skincare, h... | [AdaraID] | Desember, hemat, hingga, start, aja, juga, bak... |
| 216 | [Butuh, rekomendasi, face, moisturizer, yang, ... | [todayispat] | rekomendasi, yang, masuk, boikot, selain, dan,... |
| 217 | [Dr, semua, bisnis, kecilkecilan, yg, pernah, ... | [amysorayaa] | semua, bisnis, pernah, aku, coba, mulai, panas... |
| 218 | [What, do, you, want, in, Alis, tebal, dan, bu... | [msbbid] | do, tebal, dan, bulu, mata, panjang, lentik |

219 rows x 3 columns

```
1 # Mengambil kata-kata slang yang telah ditemukan
2 found_slang_list = data_parsing['found_slang_in_text'].str.split(', ').explode().tolist()
3
4 # Menyimpan seluruh slang ke dalam 'all slang'
5 all_slang = found_slang_list
```

```
1 # Mengambil 100 kata terakhir
2 slang_100 = found_slang_list[-100:]
3
4 # Menampilkan 100 kata slang terakhir
5 slang_100
```

```
['dong',
 'biar',
 'kulit',
 'wajah',
 'sehat',
 'jadi',
```



```
'nih',
'muka',
'lagi',
'sering',
'aku',
'banyak',
'yang',
'ini',
'ya',
'definisi',
'campuran',
'manis',
'dan',
'juga',
'mendeskripsikan',
'wangi',
'wanita',
'dewasa',
'seserahan',
'yang',
'contoh',
'',
'ada',
'ide',
'kalo',
'personal',
'jadi',
'yang',
'kaya',
'banget',
'dengan',
'penghasilan',
'per',
'hari',
'banyak',
'yang',
'suka',
'Desember',
'hemat',
'hingga',
'start',
'aja',
'juga',
'bakalan',
'hadiah',
'produk',
'favorit',
'kamu',
'yang',
'rekomendasi',
'yang',
'masuk',
```

```
1 # Membuat DataFrame dari list kata-kata slang
2 all_slang_df = pd.DataFrame({'all_slang': all_slang})
3 all_slang_df
```

| | all_slang |
|------|-----------|
| 0 | akhir |
| 1 | tahun |
| 2 | ini |
| 3 | aku |
| 4 | kulit |
| ... | ... |
| 2572 | dan |
| 2573 | bulu |
| 2574 | mata |
| 2575 | panjang |
| 2576 | lentik |

2577 rows × 1 columns

```
1 # Simpan DataFrame ke dalam file CSV
2 all_slang_df.to_csv('3B.Data_Slang.csv', index=False)
```

▼ c. Lakukan tokenizing berdasarkan hasil 3b

- simpan hasilnya dalam bentuk csv/excel
- tampilkan 100 token pertama. (bobot:10)

```
1 # Import library yang diperlukan
2 from nltk.tokenize import word_tokenize
3
4 # Tokenizing kolom 'all_slang'
5 all_slang_df['tokens'] = all_slang_df['all_slang'].apply(word_tokenize)
6
7 # Tampilkan DataFrame setelah tokenizing
8 display(all_slang_df)
```

| | all_slang | tokens | |
|------|-----------|-----------|--|
| 0 | akhir | [akhir] | |
| 1 | tahun | [tahun] | |
| 2 | ini | [ini] | |
| 3 | aku | [aku] | |
| 4 | kulit | [kulit] | |
| ... | ... | ... | |
| 2572 | dan | [dan] | |
| 2573 | bulu | [bulu] | |
| 2574 | mata | [mata] | |
| 2575 | panjang | [panjang] | |
| 2576 | lentik | [lentik] | |

2577 rows × 2 columns

```
1 # Menampilkan 100 token pertama
2 display(all_slang_df['tokens'].head(100))
```

```
0      [akhir]
1      [tahun]
2      [ini]
3      [aku]
4      [kulit]
...
95     [lalu]
96     [dan]
97     [terlalu]
98     [mempan]
99     [aku]
Name: tokens, Length: 100, dtype: object
```

```
1 # Simpan hasil tokenizing ke dalam file CSV
2 all_slang_df.to_csv('3C.Data_Slang_Tokenized.csv', index=False)
```

✓ d. Lakukan stopwords removing berdasarkan hasil 3c

- simpan hasilnya dalam bentuk csv/excel. (bobot:10)

```
1 # Mengimpor modul nltk dan mengunduh data stopwords
2 import nltk
3 nltk.download('stopwords')

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

```
1 # Daftar kata-kata yang telah Anda berikan
2 daftar_kata = all_slang
3
4 # Menggunakan NLTK untuk mendapatkan daftar stopwords dalam bahasa Indonesia
5 stopwords_list = set(stopwords.words('indonesian'))
6
7 # Menghapus stopwords dari daftar kata-kata
8 filter_words = [word for word in daftar_kata if word not in stopwords_list]
9
10 # Menampilkan hasil setelah penghapusan stopwords
11 filter_words
```

```
selesai',
'pakai',
'rehat',
'rombongan',
'nomor',
'pemenang',
'ya',
'nak',
'jerawat',
'bawa',
'kawan',
'kawan',
'nak',
'pakai',
'edisi',
'kalo',
'liburan',
'bawa',
'habis',
'orang',
'pemenang',
'gampang',
'domisili',
'tanggal',
'Januari',
'tanggung',
'berkembang',
'jualan',
'jujur',
'deh',
'rutin',
'maksimal',
'cerita',
'mempan',
'tau',
'ampuh',
'pas',
'iklan',
'si',
'pura',
'telpon',
'kabur',
'pas',
'selesai',
'ba',
'produk',
'banget',
'beli',
'aja',
'konsultasi',
'dokter',
'gratis',
'jam',
'salah',
'premium',
'formulasi',
'dokter',
'super',
'plug',
...
```

```
1 # Membuat DataFrame dari hasil penghapusan stopwords
2 result_stopword = pd.DataFrame({'hasil_stopwords': filter_words})
3
4 # Tampilkan
5 result_stopword
```

| hasil_stopwords | |
|-----------------|--------|
| 0 | kulit |
| 1 | wajah |
| 2 | bagus |
| 3 | kulit |
| 4 | wajah |
| ... | ... |
| 1303 | do |
| 1304 | tebal |
| 1305 | bulu |
| 1306 | mata |
| 1307 | lentik |

1308 rows × 1 columns

```
1 # Menyimpan hasilnya ke dalam file CSV
2 result_stopword.to_csv('3D.Hasil_Stopwords.csv', index=False)
```

✓ e. Lakukan stemming berdasarkan hasil 3d dan tampilkan 100 stem pertama. (bobot:10)

```
1 from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
2
3 # Inisialisasi stemmer
4 stemmer_factory = StemmerFactory()
5 stemmer = stemmer_factory.create_stemmer()
6
7 # Melakukan stemming pada hasil penghapusan stopwords
8 result_stopword['stem'] = result_stopword['hasil_stopwords'].apply(stemmer.stem)
9
10 # Menampilkan 100 stem pertama
11 display(result_stopword['stem'].head(100).to_frame(name='100_stem_pertama'))
```

| | 100_stem_pertama | |
|-----|------------------|--|
| 0 | kulit | |
| 1 | wajah | |
| 2 | bagus | |
| 3 | kulit | |
| 4 | wajah | |
| ... | ... | |
| 95 | person | |
| 96 | malas | |
| 97 | ... | |