Evaluasi CPMK 1 dan

Soal 1

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.

```
# Import required Python package
!pip install pandas

# Install Node.js (because tweet-harvest built using Node.js)
!sudo apt-get update
!sudo apt-get install -y ca-certificates curl gnupg
!sudo mkdir -p /etc/apt/keyrings
!curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key | sudo gpg --

!NODE_MAJOR=20 && echo "deb [signed-by=/etc/apt/keyrings/nodesource.gpg] https://de

!sudo apt-get update
!sudo apt-get install nodejs -y

!node -v
```

```
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.10
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-pack
Requirement already satisfied: numpy>=1.21.0 in /usr/local/lib/python3.10/dist-pac
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages
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
Hit:3 https://deb.nodesource.com/node_20.x nodistro InRelease
Get:4 http://security.ubuntu.com/ubuntu jammy-security InRelease [110 kB]
Hit:5 http://archive.ubuntu.com/ubuntu jammy InRelease
Get:6 http://archive.ubuntu.com/ubuntu jammy-updates InRelease [119 kB]
Hit:7 https://ppa.launchpadcontent.net/c2d4u.team/c2d4u4.0+/ubuntu jammy InRelease
Get:8 http://archive.ubuntu.com/ubuntu jammy-backports InRelease [109 kB]
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
Fetched 338 kB in 3s (132 kB/s)
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 (7.81.0-1ubuntu1.14).
gnupg is already the newest version (2.2.27-3ubuntu2.1).
0 upgraded, 0 newly installed, 0 to remove and 20 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_2
Hit:1 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
Hit:2 https://deb.nodesource.com/node_20.x nodistro InRelease
Hit:3 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64
Hit:4 http://security.ubuntu.com/ubuntu jammy-security 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
nodejs is already the newest version (20.8.1-1nodesource1).
0 upgraded, 0 newly installed, 0 to remove and 20 not upgraded.
v20.8.1
```

```
# Crawl Data

```

stemmed_twitter_data.csv stopword_twitter

..
node_modules
sample_data
tweets-data
cleaned_result.csv
cleaned_twitter_data.csv
package-lock.json
package.json
parsed_result.csv
result.csv
slang_tokens.csv
slang_twitter_data.csv
slangword.csv
stemmed_twitter_data.csv
stopword_twitter_data.csv
tokenized_twitter_data.csv

```
filename = 'result.csv'
search_keyword = 'ganjar pranowo'
limit = 500

!npx --yes tweet-harvest@latest -o "{filename}" -s "{search_keyword}" -l {limit}
```

```
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 340

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 351

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 367

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 382

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 397

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 408

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 420

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 431

--Taking a break, waiting for 10 seconds...

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 440

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 454

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 465

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 480

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 497

Got some tweets, saving to file...
Your tweets saved to: /content/tweets-data/result.csv
Total tweets saved: 512
Already got 512 tweets, done scrolling...
```

```
import pandas as pd

# Specify the path to your CSV file
file_path = f"tweets-data/{filename}"

# Read the CSV file into a pandas DataFrame
df = pd.read_csv(file_path, delimiter=";")

# Display the DataFrame
display(df)
```

	created_at	id_str	full_text	quote_count	reply_count	re
0	Fri Oct 20 11:03:30 +0000 2023	1715322846833205374	@syahirularif Yang pasti @jokowi @gibran_tweet...	0	0	
1	Fri Oct 20 11:03:25 +0000 2023	1715322827472335187	@Nikmatul_Sg @ganjarpranowo @mohmahfudmd Visi ...	0	0	
2	Fri Oct 20 11:03:23 +0000 2023	1715322818022592572	@Anase1985 @Muhayya__ @ganjarpranowo @mohmahfu...	0	0	
3	Fri Oct 20 11:03:22 +0000 2023	1715322813689843824	@Melihat_Indo @ganjarpranowo @mohmahfudmd maki...	0	0	
4	Fri Oct 20 11:03:22 +0000 2023	1715322813408837699	@Prihati_utami Dimana pun berada pilihan tetap...	0	0	
...	...	...	...	...	...	
507	Fri Oct 20 10:34:07 +0000 2023	1715315452325642461	@ganjarpranowo Pak Ganjar memang pantas di duk...	0	0	
508	Fri Oct 20 10:34:07 +0000 2023	1715315450840551783	@Melihat_Indo @ganjarpranowo @mohmahfudmd Mere...	0	0	
509	Fri Oct 20 10:34:05 +0000 2023	1715315443194605775	@imadya @ganjarpranowo @basuki_btp Ditunggu re...	0	0	

```
# Cek jumlah data yang didapatkan

num_tweets = len(df)
print(f"Jumlah tweet dalam dataframe adalah {num_tweets}.")
```

```
Jumlah tweet dalam dataframe adalah 512.
```

+0000 2023 sudah terbukti

Soal 2

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

```
import regex as re
import pandas as pd
```

```
def clean_text(text):
    # Menghapus tanda baca dan karakter non-alfanumerik
    cleaned_text = re.sub(r'[^a-zA-Z\s]', '', text)
    return cleaned_text
```

```
input_file = "result.csv"
output_file = "cleaned_result.csv"
```

```
file_path = f"tweets-data/{filename}"
df = pd.read_csv(file_path, delimiter=";")
```

```
df['full_text'] = df['full_text'].apply(clean_text)
```

```
df.to_csv(output_file, index=False, encoding='utf-8')

print('Data sudah dibersihkan dengan nama file cleaned_result.csv')
```

```
Data sudah dibersihkan dengan nama file cleaned_result.csv
```

```
pd.read_csv('/content/cleaned_result.csv')
```

	created_at	id_str	full_text	quote_count	reply_count	retw
0	Fri Oct 20 11:03:30 +0000 2023	1715322846833205374	syahirularif Yang pasti jokowi gibrantweet duk...	0	0	
1	Fri Oct 20 11:03:25 +0000 2023	1715322827472335187	NikmatulSg ganjarpranowo mohmahfudmd Visi misi...	0	0	
2	Fri Oct 20 11:03:23 +0000 2023	1715322818022592572	Anase Muhayya ganjarpranowo mohmahfudmd chchot...	0	0	
3	Fri Oct 20 11:03:22 +0000 2023	1715322813689843824	MelihatIndo ganjarpranowo mohmahfudmd makin ma...	0	0	
4	Fri Oct 20 11:03:22 +0000 2023	1715322813408837699	Prihatiutami Dimana pun berada pilihan tetap g...	0	0	
...	...	...	...	...	...	
507	Fri Oct 20 10:34:07 +0000 2023	1715315452325642461	ganjarpranowo Pak Ganjar memang pantas di duku...	0	0	
508	Fri Oct 20 10:34:07 +0000 2023	1715315450840551783	MelihatIndo ganjarpranowo mohmahfudmd Mereka m...	0	0	
509	Fri Oct 20 10:34:05 +0000 2023	1715315443194605775	imadya ganjarpranowo basukibtp Ditunggu rekama...	0	0	
510	Fri Oct 20 10:33:58 +0000 2023	1715315416174842157	MelihatIndo ganjarpranowo mohmahfudmd Kita yak...	0	0	
511	Fri Oct 20 10:33:57 +0000 2023	1715315411368210936	ganjarpranowo Pak ganjar yang sudah terbukti k...	0	0	

512 rows × 12 columns

Soal 3a

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

```
# Buka file CSV asal dan buat file CSV baru untuk hasil parsing
with open('cleaned_result.csv', 'r', encoding='utf-8') as csvfile:
    reader = csv.DictReader(csvfile)

    with open('parsed_result.csv', 'w', newline='', encoding='utf-8') as outputfile
        fieldnames = ['Word']  # Kolom dalam file CSV baru
        writer = csv.DictWriter(outputfile, fieldnames=fieldnames)
        writer.writeheader()

        for row in reader:
            # Akses data dari kolom "full_text"
            full_text = row['full_text']

            # Memisahkan teks menjadi kata-kata
            words = full_text.split()

            # Menulis kata-kata ke dalam file CSV baru
```

```
            for word in words:
                writer.writerow({'Word': word})

print('Hasil parsing telah disimpan dalam "parsed_result.csv"')
```

```
Hasil parsing telah disimpan dalam "parsed_result.csv"
```

Soal 3b

b.Carilah kata-kata slangword yang ada dalam dataset Anda, dengan cara mencocokan dengan kamus KBBI (terlampir).
Simpan hasilnya dalam bentuk csv/excel.
Tampilkan 100 kata slang yang Anda dapatkan dan tampilkan dalam bentuk Gunakan (bobot:25)

```
# Baca file CSV dengan dataset Twitter (cleaned_result.csv)
twitter_data = pd.read_csv('cleaned_result.csv', encoding='utf-8')

# Baca file CSV dengan daftar kata-kata slang (slangword.csv)
slang_data = pd.read_csv('slangword.csv', encoding='utf-8', names=['Slang Word'])

# Konversi daftar kata-kata slang ke dalam bentuk set untuk pencocokan lebih efisie
slang_set = set(slang_data['Slang Word'])

# Buat kolom baru dalam dataset Twitter untuk kata-kata yang bukan slang
twitter_data['Cleaned Text'] = twitter_data['full_text'].apply(
    lambda text: ' '.join([word for word in text.split() if word not in slang_set])
)

# Simpan hasilnya dalam file CSV baru
twitter_data.to_csv('slang_twitter_data.csv', index=False, encoding='utf-8')

# Tampilkan 100 kata slang
print('100 Kata Slang:')
print(slang_data['Slang Word'][:100])

# Tampilkan hasilnya dalam bentuk data frame
print('Hasil Data Frame Setelah Menghapus Slang:')
print(twitter_data.head())
```

```
100 Kata Slang:
0       Slangwords
1             Adam
2              Aga
3          Agustus
4             Ahad
          ...
95            Jogi
96           Johar
97            Juja
98            Juli
99    Jumadilakhir
Name: Slang Word, Length: 100, dtype: object
Hasil Data Frame Setelah Menghapus Slang:
                       created_at              id_str  \
0  Fri Oct 20 11:03:30 +0000 2023  1715322846833205374
1  Fri Oct 20 11:03:25 +0000 2023  1715322827472335187
2  Fri Oct 20 11:03:23 +0000 2023  1715322818022592572
3  Fri Oct 20 11:03:22 +0000 2023  1715322813689843824
4  Fri Oct 20 11:03:22 +0000 2023  1715322813408837699

                                        full_text  quote_count  \
0  syahirularif Yang pasti jokowi gibrantweet duk...            0
1  NikmatulSg ganjarpranowo mohmahfudmd Visi misi...            0
2  Anase Muhayya ganjarpranowo mohmahfudmd chchot...            0
3  MelihatIndo ganjarpranowo mohmahfudmd makin ma...            0
4  Prihatiutami Dimana pun berada pilihan tetap g...            0

   reply_count  retweet_count  favorite_count lang          user_id_str  \
0            0              0               0   in           3244905644
1            0              0               0   in  1707053631005212672
2            0              0               0   in  1260061504767836162
3            0              0               0   in  1659432246464622594
4            0              0               0   in  1641586661233692678

   conversation_id_str        username  \
0  1715320765036495140         Mi73Hel
1  1715286319017103792        Anita33_
2  1715305293226348847       Dharma_tc
3  1715293032529203504     bougenvill4
4  1715026015330914478  fadlian_syah29
```

```
                                           tweet_url  \
0  https://twitter.com/Mi73Hel/status/17153228468...
1  https://twitter.com/Anita33_/status/1715322827...
2  https://twitter.com/Dharma_tc/status/171532281...
3  https://twitter.com/bougenvill4/status/1715322...
4  https://twitter.com/fadlian_syah29/status/1715...

                                        Cleaned Text
0  syahirularif Yang jokowi gibrantweet ganjarpra...
1  NikmatulSg ganjarpranowo mohmahfudmd Visi dila...
2  Anase Muhayya ganjarpranowo mohmahfudmd chchot...
3  MelihatIndo ganjarpranowo mohmahfudmd VisiMisi...
4        Prihatiutami Dimana pranowo gaspol menangkan
```

```
# # Membaca file "slangword.csv" untuk mengumpulkan daftar slangword
# slangwords = set()
# with open('slangword.csv', 'r', encoding='utf-8') as slangfile:
#     reader = csv.reader(slangfile)
#     for row in reader:
#         slangwords.add(row[0].strip().lower())  # Memastikan semua slangword dal

# # Membaca file "cleaned_result.csv" untuk mencari kata-kata slang
# slang_found = set()
# with open('cleaned_result.csv', 'r', encoding='utf-8') as csvfile:
#     reader = csv.DictReader(csvfile)

#     for row in reader:
#         full_text = row['full_text']

#         # Memisahkan teks menjadi kata-kata
#         words = full_text.split()

#         # Mencari kata-kata slang
#         for word in words:
#             if word.lower() in slangwords:
#                 slang_found.add(word)

# # Menyimpan hasil pencarian dalam file CSV
# with open('slang_found.csv', 'w', newline='', encoding='utf-8') as outputfile:
#     fieldnames = ['Slang Word']
#     writer = csv.DictWriter(outputfile, fieldnames=fieldnames)
#     writer.writeheader()
#     for word in slang_found:
#         writer.writerow({'Slang Word': word})

# # Menampilkan 100 kata slang pertama
# slang_found_list = list(slang_found)
# print('kata slang yang ditemukan:')
# for i in range(400):
#     print(slang_found_list[i])
```

Soal 3c

c.Lakukan tokenizing berdasarkan hasil 3b, simpan hasilnya dalam bentuk csv/excel dan tampilkan 100 token pertama. (bobot:10)

```
data = []
for slang in slang_found:
    tokens = re.findall(r'\b\w+\b', slang)  # Tokenisasi dengan menghilangkan karak
    data.append([slang, ' '.join(tokens)])

df = pd.DataFrame(data, columns=['Slang Word', 'Tokens'])

# Menyimpan hasil dalam file CSV
df.to_csv('slang_tokens.csv', index=False, encoding='utf-8')

# Menampilkan 100 token pertama dalam bentuk data frame
print('100 token pertama:')
print(df.head(100))
```

```
100 token pertama:
    Slang Word      Tokens
0         akal        akal
1        Tuhan       Tuhan
2         dulu        dulu
3        peran       peran
4        Ketua       Ketua
..         ...         ...
```

```
95      bangsa      bangsa
96  pernyataan  pernyataan
97     diskusi     diskusi
98   Kabupaten   Kabupaten
99  Kebangsaan  Kebangsaan

[100 rows x 2 columns]
```

```
import nltk
nltk.download('punkt')
from nltk.tokenize import word_tokenize
```

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

```
from nltk.tokenize import word_tokenize

# Baca file CSV dengan data Twitter yang telah dibersihkan (cleaned_twitter_data.cs
twitter_data = pd.read_csv('cleaned_twitter_data.csv', encoding='utf-8')

# Tokenisasi teks dalam kolom 'Cleaned Text'
twitter_data['Tokens'] = twitter_data['Cleaned Text'].apply(lambda text: word_toker

# Simpan hasil tokenisasi dalam file CSV
twitter_data.to_csv('tokenized_twitter_data.csv', index=False, encoding='utf-8')

# Tampilkan 100 token pertama dalam bentuk data frame
print('100 Token Pertama:')
print(twitter_data['Tokens'].apply(lambda tokens: tokens[:100]))

# Tampilkan hasilnya dalam bentuk data frame
print('Hasil Data Frame Setelah Tokenisasi:')
print(twitter_data.head())
```

```
3      [MelihatIndo, ganjarpranowo, mohmahfudmd, Visi...
4      [Prihatiutami, Dimana, pranowo, gaspol, menang...
                           ...
507                    [ganjarpranowo, Pak, Ganjar, RI]
508    [MelihatIndo, ganjarpranowo, mohmahfudmd, Mere...
509    [imadya, ganjarpranowo, basukibtp, Ditunggu, p...
510    [MelihatIndo, ganjarpranowo, mohmahfudmd, Kita...
511                    [ganjarpranowo, Pak, kinerjanya]
Name: Tokens, Length: 512, dtype: object
Hasil Data Frame Setelah Tokenisasi:
                      created_at              id_str  \
0  Fri Oct 20 11:03:30 +0000 2023  1715322846833205374
1  Fri Oct 20 11:03:25 +0000 2023  1715322827472335187
2  Fri Oct 20 11:03:23 +0000 2023  1715322818022592572
3  Fri Oct 20 11:03:22 +0000 2023  1715322813689843824
4  Fri Oct 20 11:03:22 +0000 2023  1715322813408837699

                                        full_text  quote_count  \
0  syahirularif Yang pasti jokowi gibrantweet duk...            0
1  NikmatulSg ganjarpranowo mohmahfudmd Visi misi...            0
2  Anase Muhayya ganjarpranowo mohmahfudmd chchot...            0
3  MelihatIndo ganjarpranowo mohmahfudmd makin ma...            0
4  Prihatiutami Dimana pun berada pilihan tetap g...            0

   reply_count  retweet_count  favorite_count lang          user_id_str  \
0            0              0               0   in           3244905644
1            0              0               0   in  1707053631005212672
2            0              0               0   in  1260061504767836162
3            0              0               0   in  1659432246464622594
4            0              0               0   in  1641586661233692678

   conversation_id_str        username  \
0  1715320765036495140         Mi73Hel
1  1715286319017103792        Anita33_
2  1715305293226348847       Dharma_tc
3  1715293032529203504     bougenvill4
4  1715026015330914478  fadlian_syah29

                                        tweet_url  \
0  https://twitter.com/Mi73Hel/status/17153228468...
1  https://twitter.com/Anita33_/status/1715322827...
2  https://twitter.com/Dharma_tc/status/171532281...
3  https://twitter.com/bougenvill4/status/1715322...
4  https://twitter.com/fadlian_syah29/status/1715...

                                      Cleaned Text  \
0  syahirularif Yang jokowi gibrantweet ganjarpra...
1  NikmatulSg ganjarpranowo mohmahfudmd Visi dila...
```

```
                                                Tokens
0  [syahirularif, Yang, jokowi, gibrantweet, ganj...
1  [NikmatulSg, ganjarpranowo, mohmahfudmd, Visi,...
2  [Anase, Muhayya, ganjarpranowo, mohmahfudmd, c...
3  [MelihatIndo, ganjarpranowo, mohmahfudmd, Visi...
4  [Prihatiutami, Dimana, pranowo, gaspol, menang...
```

Soal 3d

d.Lakukan stopword removing berdasarkan hasil 3c, simpan hasilnya dalam bentuk csv/excel. (bobot:10)!

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

# Download dataset stopwords untuk bahasa Indonesia
nltk.download('stopwords')
nltk.download('punkt')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
True
```

```
# Baca file CSV dengan dataset Twitter (tokenized_twitter_data.csv)
twitter_data = pd.read_csv('tokenized_twitter_data.csv', encoding='utf-8')

# Fungsi untuk menghapus stopwords
def remove_stopwords(text):
    stop_words = set(stopwords.words('indonesian'))
    words = word_tokenize(text)
    return [word for word in words if word.lower() not in stop_words]

# Melakukan penghapusan stopwords pada kolom "Tokens"
twitter_data['Tokens'] = twitter_data['Tokens'].apply(remove_stopwords)

# Simpan hasilnya dalam file CSV baru
twitter_data.to_csv('stopword_twitter_data.csv', index=False, encoding='utf-8')

print('Stopwords telah dihapus dan hasilnya disimpan dalam "stopword_twitter_data.c
```

```
Stopwords telah dihapus dan hasilnya disimpan dalam "stopword_twitter_data.csv".
```

Soal 3e

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

```
!pip install Sastrawi
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
```

```
Requirement already satisfied: Sastrawi in /usr/local/lib/python3.10/dist-packages
```

```
# Baca file CSV dengan dataset Twitter yang sudah dibersihkan dari stopwords
twitter_data = pd.read_csv('stopword_twitter_data.csv', encoding='utf-8')

# Inisialisasi Stemmer Bahasa Indonesia
factory = StemmerFactory()
stemmer = factory.create_stemmer()

# Fungsi untuk melakukan stemming
def stem_text(text):
    return stemmer.stem(text)

# Melakukan stemming pada kolom "Cleaned Text"
twitter_data['Stemmed Text'] = twitter_data['Cleaned Text'].apply(stem_text)

# Simpan hasil stemming dalam file CSV baru
twitter_data.to_csv('stemmed_twitter_data.csv', index=False, encoding='utf-8')

# Tampilkan 100 kata hasil stemming pertama
print('100 Kata Hasil Stemming Pertama:')
stemmed_tokens = twitter_data['Stemmed Text'].apply(lambda text: ' '.join(text.spl
print(stemmed_tokens)
```

```
100 Kata Hasil Stemming Pertama:
0      syahirularif yang jokowi gibrantweet ganjarpra...
1      nikmatulsg ganjarpranowo mohmahfudmd visi laks...
2      anase muhayya ganjarpranowo mohmahfudmd chchot...
3      melihatindo ganjarpranowo mohmahfudmd visimisi...
4                 prihatiutami mana pranowo gaspol menang
                              ...
507                           ganjarpranowo pak ganjar ri
508    melihatindo ganjarpranowo mohmahfudmd mereka m...
509    imadya ganjarpranowo basukibtp tunggu podcast ...
510    melihatindo ganjarpranowo mohmahfudmd kita vis...
511                              ganjarpranowo pak kerja
Name: Stemmed Text, Length: 512, dtype: object
```

Disk 79.78 GB available