PREPROCESSING DATA TEXT

Chapter 3

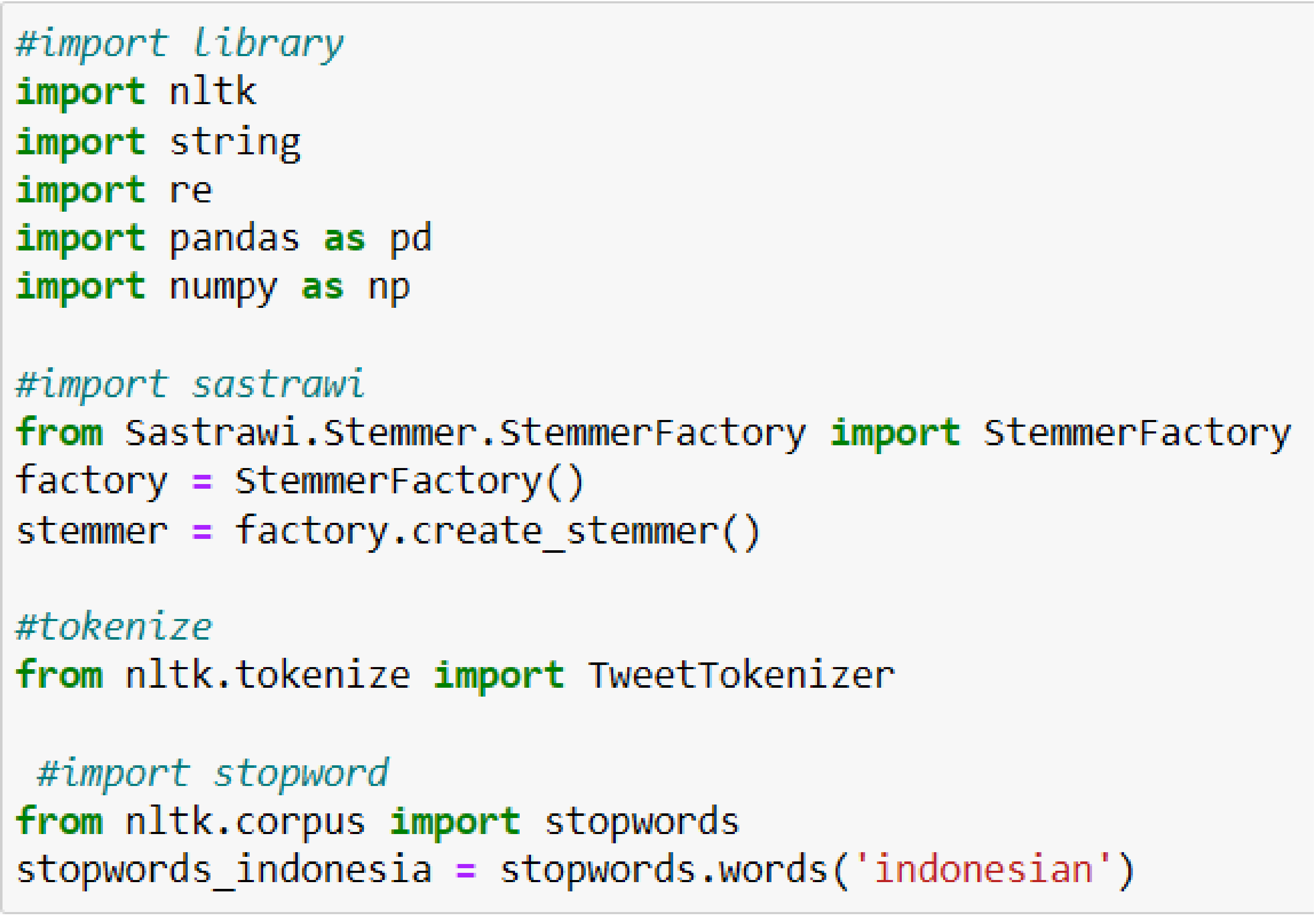
Sains Data

LANGKAH-LANGKAH PREPROCESSING TEXT

1. Cleaning : Pembersihan dari karakter atau elemen yang tidak diperlukan pada teks, seperti tanda baca, angka, dan karakter khusus.
2. Case folding : mengubah semua huruf menjadi huruf kecil semua.
3. Tokenizing : proses memecah teks menjadi unit-unit kecil seperti kata atau kalimat.
4. Filtering (Stopword Removal) : menghapus kata-kata yang tidak memiliki makna dalam teks, seperti "dan", "di", "ke", dll
5. Stemming : mengubah kata menjadi bentuk dasar

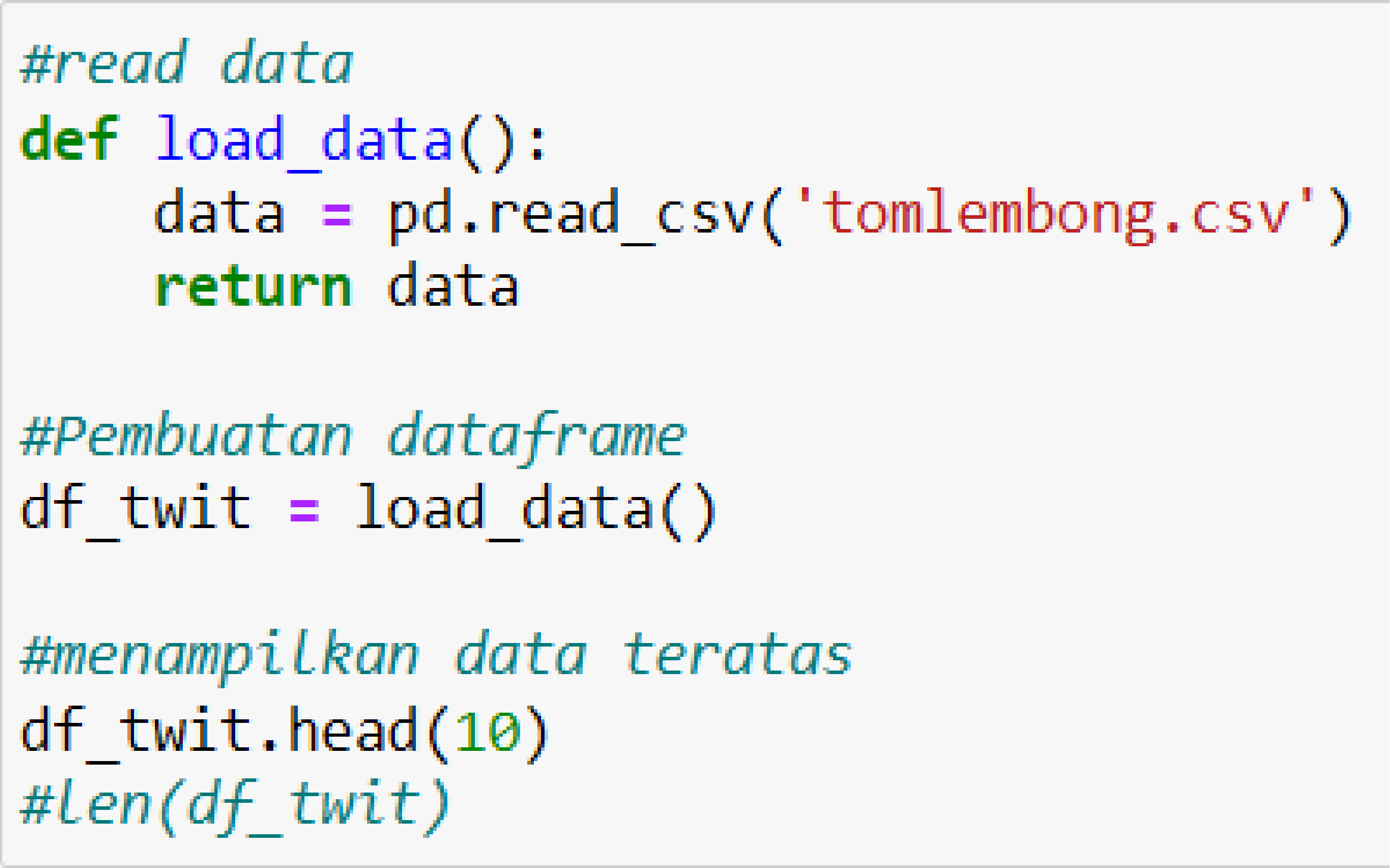
Contoh Codingnya

IMPORT LIBRARY



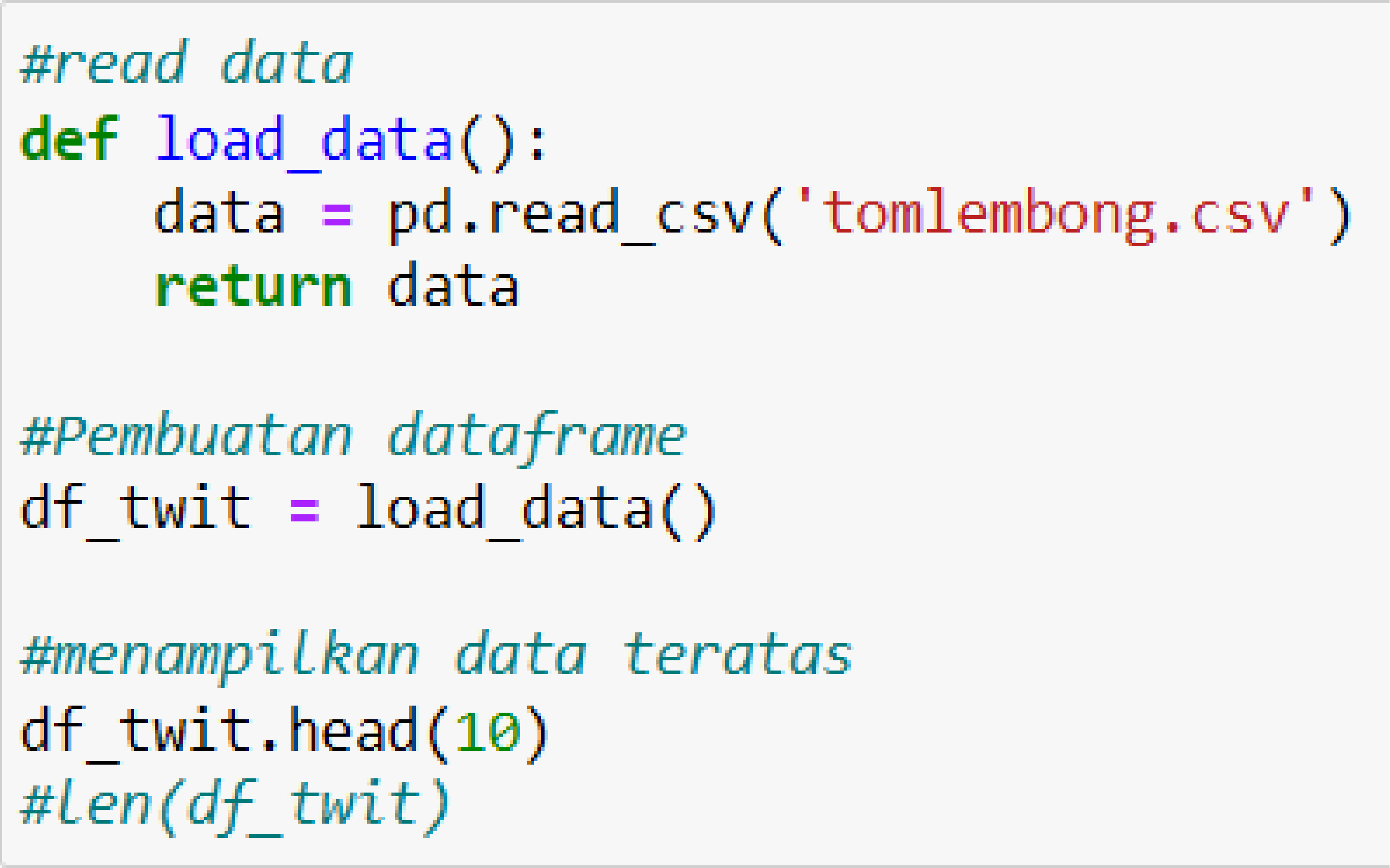
```python
#import library
import nltk
import string
import re
import pandas as pd
import numpy as np

#import sastrawi
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
factory = StemmerFactory()
stemmer = factory.create_stemmer()

#tokenize
from nltk.tokenize import TweetTokenizer

 #import stopword
from nltk.corpus import stopwords
stopwords_indonesia = stopwords.words('indonesian')
```

PEMBACAAN DATA

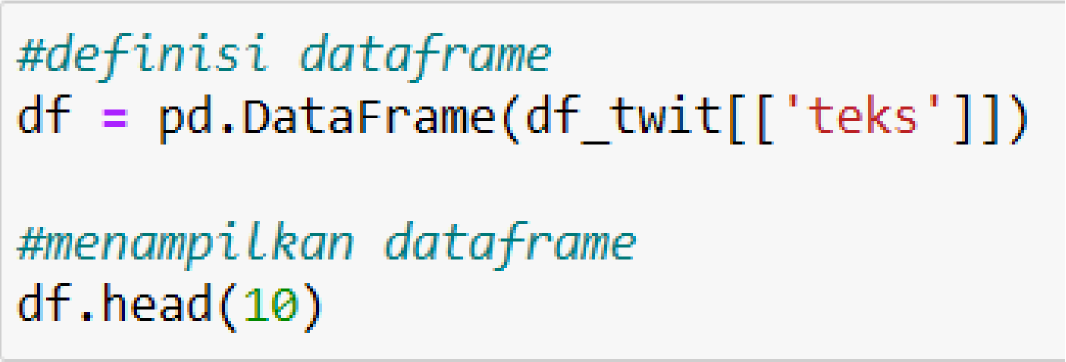


```
#read data
def load_data():
    data = pd.read_csv('tomlembong.csv')
    return data

#Pembuatan dataframe
df_twit = load_data()

#menampilkan data teratas
df_twit.head(10)
#len(df_twit)
```

PEMBACAAN DATA



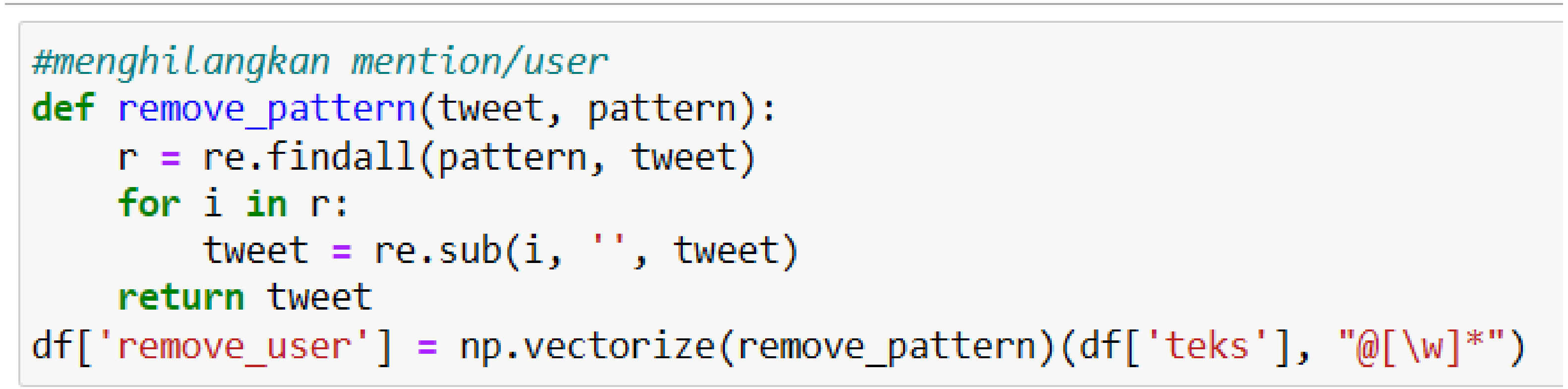
```python
#read data
def load_data():
    data = pd.read_csv('tomlembong.csv')
    return data

#Pembuatan dataframe
df_twit = load_data()

#menampilkan data teratas
df_twit.head(10)
#len(df_twit)
```



```
#definisi dataframe
df = pd.DataFrame(df_twit[['teks']])

#menampilkan dataframe
df.head(10)
```

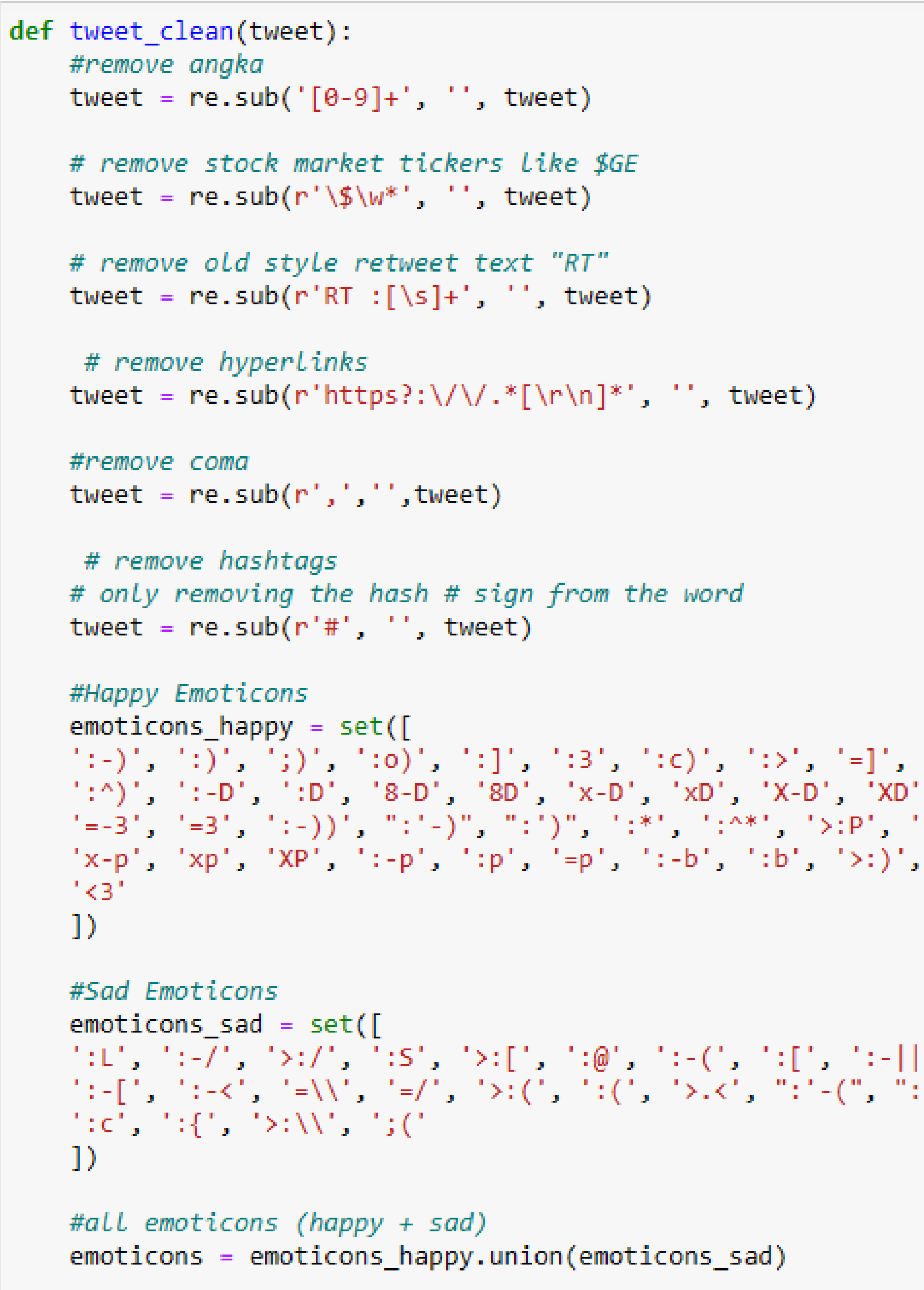
MENGAMBIL KOLOM TEKS SAJA

MENGHILANGKAN MENTION

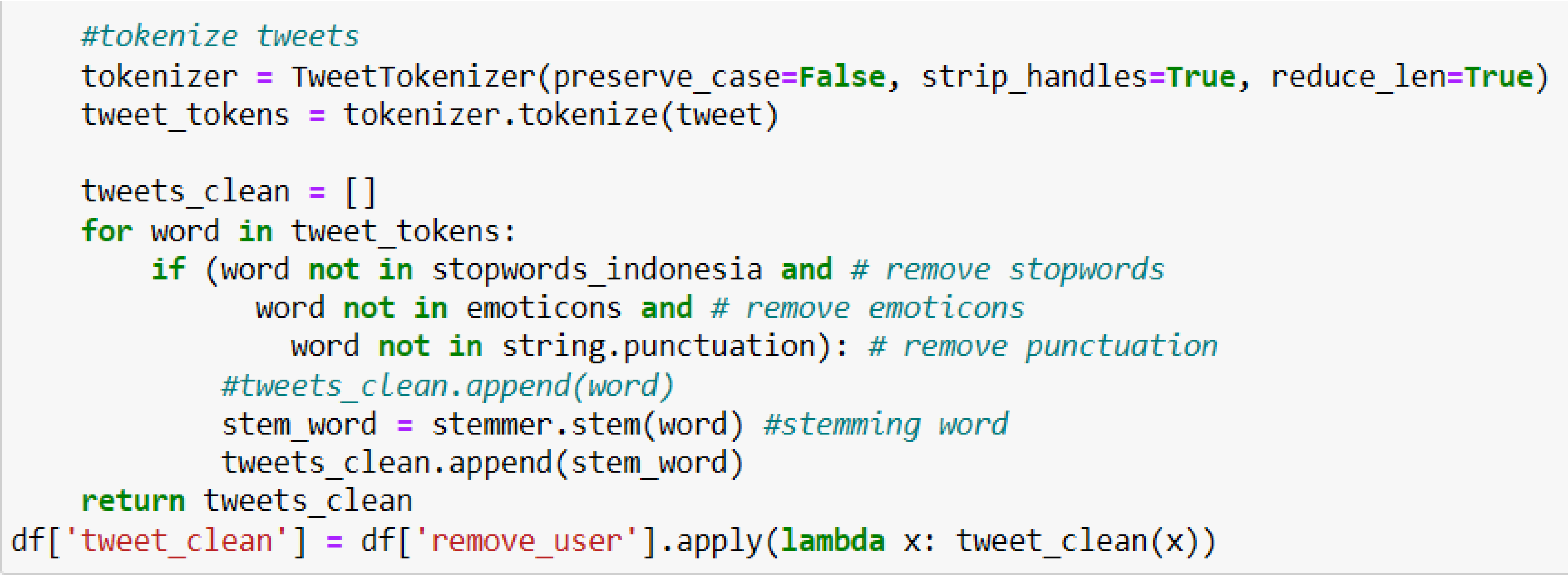


```python
#menghilangkan mention/user
def remove_pattern(tweet, pattern):
    r = re.findall(pattern, tweet)
    for i in r:
        tweet = re.sub(i, '', tweet)
    return tweet
df['remove_user'] = np.vectorize(remove_pattern)(df['teks'], "@[\w]*")
```

REGULER EXPRESSION

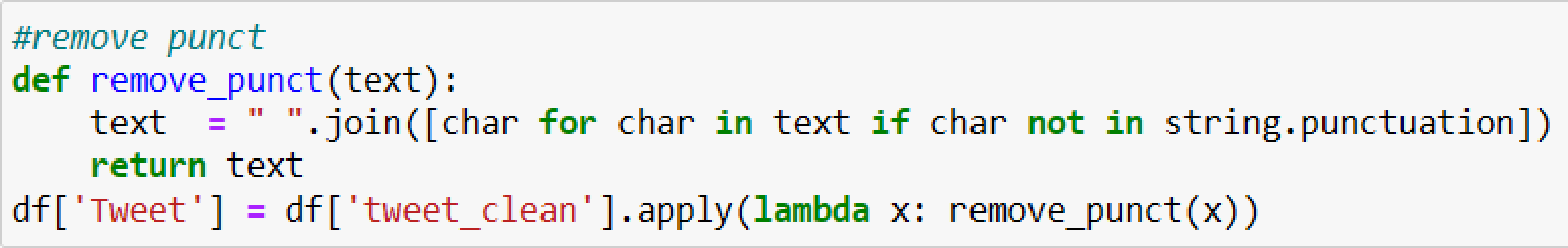


```
def tweet_clean(tweet):
    #remove angka
    tweet = re.sub('[0-9]+', '', tweet)

    # remove stock market tickers like $GE
    tweet = re.sub(r'\$\w*', '', tweet)

    # remove old style retweet text "RT"
    tweet = re.sub(r'RT :[\s]+', '', tweet)

     # remove hyperlinks
    tweet = re.sub(r'https?:\/\/.*[\r\n]*', '', tweet)

    #remove coma
    tweet = re.sub(r',','',tweet)

     # remove hashtags
    # only removing the hash # sign from the word
    tweet = re.sub(r'#', '', tweet)

    #Happy Emoticons
    emoticons_happy = set([
    ':-)', ':)', ';)', ':o)', ':]', ':3', ':c)', ':>', '=]',
    ':^)', ':-D', ':D', '8-D', '8D', 'x-D', 'xD', 'X-D', 'XD'
    '=-3', '=3', ':-))', ":'-)", ":')", ':*', ':^*', '>:P', '
    'x-p', 'xp', 'XP', ':-p', ':p', '=p', ':-b', ':b', '>:)',
    '<3'
    ])

    #Sad Emoticons
    emoticons_sad = set([
    ':L', ':-/', '>:/', ':S', '>:[', ':@', ':-(', ':[', ':-||
    ':-[', ':-<', '=\\', '=/', '>:(', ':(', '>.<', ":'-(", ":
    ':c', ':{', '>:\\', ';('
    ])

    #all emoticons (happy + sad)
    emoticons = emoticons_happy.union(emoticons_sad)
```



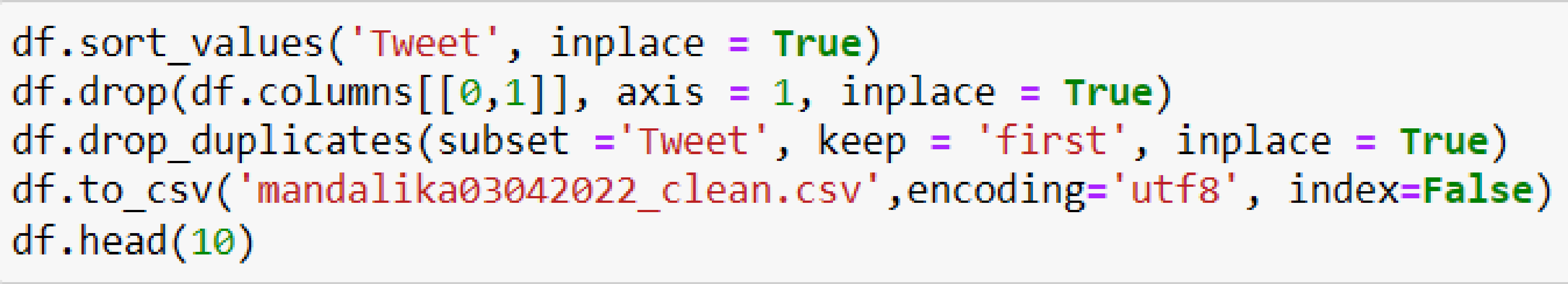
```
    #tokenize tweets
    tokenizer = TweetTokenizer(preserve_case=False, strip_handles=True, reduce_len=True)
    tweet_tokens = tokenizer.tokenize(tweet)

    tweets_clean = []
    for word in tweet_tokens:
        if (word not in stopwords_indonesia and # remove stopwords
              word not in emoticons and # remove emoticons
                word not in string.punctuation): # remove punctuation
            #tweets_clean.append(word)
            stem_word = stemmer.stem(word) #stemming word
            tweets_clean.append(stem_word)
    return tweets_clean
df['tweet_clean'] = df['remove_user'].apply(lambda x: tweet_clean(x))
```

TOKENISASI, STOPWORD, DAN STEMMING



```python
#remove punct
def remove_punct(text):
    text  = " ".join([char for char in text if char not in string.punctuation])
    return text
df['Tweet'] = df['tweet_clean'].apply(lambda x: remove_punct(x))
```

REMOVE PUNCTUATION



```
df.sort_values('Tweet', inplace = True)
df.drop(df.columns[[0,1]], axis = 1, inplace = True)
df.drop_duplicates(subset ='Tweet', keep = 'first', inplace = True)
df.to_csv('mandalika03042022_clean.csv',encoding='utf8', index=False)
df.head(10)
```

MENYIMPAN HASIL TEKS BERSIH

ADA YANG DITANYAKAN?