z7sljzezx

November 10, 2024

1 2602164660 - Dimas Aditya Susetyo

[ ]:
```
#!pip install google-api-python-client python-dotenv pandas
```

1.1 (1a) scraping komen yutup

[ ]:
```
from googleapiclient.discovery import build
import pandas as pd
from google.colab import files, drive
import getpass
import googleapiclient.discovery
import pandas as pd
```

[ ]:
```
api_key = "AIzaSyCcPwchCDpBPruxots_N0blXNCNf55Gkg4"

#playlist link: https://youtube.com/playlist?
 ↪list=PL45Zi-IFHBniYaW1x_KQ2zAPAPcg-hJGu&feature=shared;
playlist_ids = ['PL45Zi-IFHBniYaW1x_KQ2zAPAPcg-hJGu']
```

[ ]:
```
# Build the YouTube client
youtube = build('youtube', 'v3', developerKey=api_key)
```

[ ]:
```
#get ID's

def get_all_video_ids_from_playlists(youtube, playlist_ids):
    all_videos = []

    for playlist_id in playlist_ids:
        next_page_token = None

        # Fetch videos from the current playlist
        while True:
            playlist_request = youtube.playlistItems().list(
                part='contentDetails',
                playlistId=playlist_id,
                maxResults=50,
                pageToken=next_page_token)
```

```
            playlist_response = playlist_request.execute()

            all_videos += [item['contentDetails']['videoId'] for item in␣
 ↪playlist_response['items']]

            next_page_token = playlist_response.get('nextPageToken')

            if next_page_token is None:
                break

    return all_videos

# Fetch all video IDs from the specified playlists
video_ids = get_all_video_ids_from_playlists(youtube, playlist_ids)
```

[ ]:
```
# Function to get the video author
def get_video_author(youtube, video_id):
    video_request = youtube.videos().list(
        part="snippet",
        id=video_id
    )
    video_response = video_request.execute()
    video_author = video_response['items'][0]['snippet']['channelTitle']
    return video_author

# Function to get replies for a specific comment
def get_replies(youtube, parent_id, video_id, video_author):
    replies = []
    next_page_token = None

    while True:
        reply_request = youtube.comments().list(
            part="snippet",
            parentId=parent_id,
            textFormat="plainText",
            maxResults=100,
            pageToken=next_page_token
        )
        reply_response = reply_request.execute()

        for item in reply_response['items']:
            comment = item['snippet']
            replies.append({
                'Timestamp': comment['publishedAt'],
                'Username': comment['authorDisplayName'],
                'VideoID': video_id,
                'VideoAuthor': video_author,
```

```
                'Comment': comment['textDisplay'],
                'Likes': comment.get('likeCount', 0),
                'Date': comment['updatedAt'] if 'updatedAt' in comment else␣
 ↪comment['publishedAt']
            })

        next_page_token = reply_response.get('nextPageToken')
        if not next_page_token:
            break

    return replies

# Function to get all comments (including replies) for a single video
def get_comments_for_video(youtube, video_id):
    all_comments = []
    next_page_token = None

    # Get video author
    video_author = get_video_author(youtube, video_id)

    while True:
        comment_request = youtube.commentThreads().list(
            part="snippet",
            videoId=video_id,
            pageToken=next_page_token,
            textFormat="plainText",
            maxResults=100
        )
        comment_response = comment_request.execute()

        for item in comment_response['items']:
            top_comment = item['snippet']['topLevelComment']['snippet']
            all_comments.append({
                'Timestamp': top_comment['publishedAt'],
                'Username': top_comment['authorDisplayName'],
                'VideoID': video_id,
                'VideoAuthor': video_author,
                'Comment': top_comment['textDisplay'],
                'Likes': top_comment.get('likeCount', 0),
                'Date': top_comment['updatedAt'] if 'updatedAt' in top_comment␣
 ↪else top_comment['publishedAt']
            })

            if item['snippet']['totalReplyCount'] > 0:
                all_comments.extend(get_replies(youtube,␣
 ↪item['snippet']['topLevelComment']['id'], video_id, video_author))
```

```
        next_page_token = comment_response.get('nextPageToken')
        if not next_page_token:
            break

    return all_comments

# List to hold all comments from all videos
all_comments = []

for video_id in video_ids:
    video_comments = get_comments_for_video(youtube, video_id)
    all_comments.extend(video_comments)

# Create DataFrame
comments_df = pd.DataFrame(all_comments)
```

[ ]:
```
comments_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1592 entries, 0 to 1591
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Timestamp    1592 non-null   object
 1   Username     1592 non-null   object
 2   VideoID      1592 non-null   object
 3   VideoAuthor  1592 non-null   object
 4   Comment      1592 non-null   object
 5   Likes        1592 non-null   int64
 6   Date         1592 non-null   object
dtypes: int64(1), object(6)
memory usage: 87.2+ KB
```

[ ]:
```
comments_df
```

[ ]:

	Timestamp	Username	VideoID	\
0	2024-10-02T09:05:30Z	@ParahitaBali	hGWMaKbP_Ss	
1	2024-10-02T09:16:23Z	@AhmadWildani	hGWMaKbP_Ss	
2	2024-10-02T09:19:02Z	@ParahitaBali	hGWMaKbP_Ss	
3	2024-10-02T09:33:04Z	@AnekaTransportasi	hGWMaKbP_Ss	
4	2024-10-03T01:14:51Z	@yunantoabdul5651	hGWMaKbP_Ss	
…	…	…	…	
1587	2021-01-28T05:02:04Z	@varsad7650	NVQ-58Z7UB4	
1588	2021-01-28T05:02:04Z	@axzelmarvellos2799	NVQ-58Z7UB4	
1589	2021-01-28T05:02:00Z	@yosualaban4033	NVQ-58Z7UB4	
1590	2021-01-28T05:01:40Z	@miftahululum4646	NVQ-58Z7UB4	
1591	2021-01-28T05:01:28Z	@miftahululum4646	NVQ-58Z7UB4	

```
           VideoAuthor                                                Comment  \
0        Ahmad Wildani  Sayangnya angkatan pagi sudah tidak beroperasi…
1        Ahmad Wildani                                     Wah iyakah mas ?
2        Ahmad Wildani  @@AhmadWildani line perdana nya 17 September. …
3        Ahmad Wildani  @@ParahitaBali okupansinya ga maksimal. Karena…
4        Ahmad Wildani  Tadi saya mau pesen ini ke agen, info agen sud…
…                  …                                                  …
1587  Bima Rahmatulloh                                              Hadir
1588  Bima Rahmatulloh                                        Mantap masss
1589  Bima Rahmatulloh                                                 Ke3
1590  Bima Rahmatulloh                                 assalamulaikum mas
1591  Bima Rahmatulloh                                   pertama komen

      Likes                 Date
0        34  2024-10-02T09:05:30Z
1         6  2024-10-02T09:16:23Z
2         5  2024-10-02T09:19:02Z
3         0  2024-10-02T09:33:04Z
4         0  2024-10-03T01:14:51Z
…     …                    …
1587      1  2021-01-28T05:02:04Z
1588      2  2021-01-28T05:02:04Z
1589      1  2021-01-28T05:02:00Z
1590      1  2021-01-28T05:01:40Z
1591      2  2021-01-28T05:01:28Z

[1592 rows x 7 columns]
```

[ ]:
```
#jumlah komentar

print("jumlah komentar: ", len(comments_df))
print()
comment_counts = comments_df.groupby(['VideoID', 'VideoAuthor']).size().
 ↪reset_index(name='CommentCount')
comment_counts
```

```
jumlah komentar:  1592
```

[ ]:
```
        VideoID         VideoAuthor  CommentCount
0  0zv_6F4reqQ  Andriawan Pratikto           359
1  NVQ-58Z7UB4    Bima Rahmatulloh           407
2  U1IlHiWLGmc       Ahmad Wildani           243
3  ZBXy3ktKMEg  Andriawan Pratikto           377
4  hGWMaKbP_Ss       Ahmad Wildani           206
```

[ ]:
```
csv_file = '27TransTripReport.csv'  # Name your file
comments_df.to_csv(csv_file, index=False)
```

1.2 (1b) text preprocessing

[233]:
```
!pip install num2words
!pip install emoji
!pip install Sastrawi
```

```
Requirement already satisfied: num2words in /usr/local/lib/python3.10/dist-
packages (0.5.13)
Requirement already satisfied: docopt>=0.6.2 in /usr/local/lib/python3.10/dist-
packages (from num2words) (0.6.2)
Requirement already satisfied: emoji in /usr/local/lib/python3.10/dist-packages
(2.14.0)
Requirement already satisfied: Sastrawi in /usr/local/lib/python3.10/dist-
packages (1.0.1)
```

[234]:
```
import pandas as pd
import numpy as np
import re
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer

from collections import Counter

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!
```

[234]:
```
True
```

[235]:
```
df = pd.read_csv('27TransTripReport.csv')
df.head(10)
```

[235]:

	Timestamp	Username	VideoID	VideoAuthor	\
0	2024-10-02T09:05:30Z	@ParahitaBali	hGWMaKbP_Ss	Ahmad Wildani	
1	2024-10-02T09:16:23Z	@AhmadWildani	hGWMaKbP_Ss	Ahmad Wildani	
2	2024-10-02T09:19:02Z	@ParahitaBali	hGWMaKbP_Ss	Ahmad Wildani	
3	2024-10-02T09:33:04Z	@AnekaTransportasi	hGWMaKbP_Ss	Ahmad Wildani	
4	2024-10-03T01:14:51Z	@yunantoabdul5651	hGWMaKbP_Ss	Ahmad Wildani	

```
5  2024-10-03T12:24:10Z    @jonathanwiyono1451  hGWMaKbP_Ss  Ahmad Wildani
6  2024-10-03T13:31:10Z  @afrizaldwirahman4061  hGWMaKbP_Ss  Ahmad Wildani
7  2024-10-07T10:38:51Z  @bagusyogapratama3948  hGWMaKbP_Ss  Ahmad Wildani
8  2024-10-16T10:49:51Z     @yohansetiawan5387  hGWMaKbP_Ss  Ahmad Wildani
9  2024-11-03T07:26:30Z      @dodisetiawan9063  hGWMaKbP_Ss  Ahmad Wildani

                                          Comment  Likes  \
0  Sayangnya angkatan pagi sudah tidak beroperasi…     34
1                                Wah iyakah mas ?      6
2  @@AhmadWildani line perdana nya 17 September. …      5
3  @@ParahitaBali okupansinya ga maksimal. Karena…      0
4  Tadi saya mau pesen ini ke agen, info agen sud…      0
5  Yah sia2 mas wildan bkin video ini, busnya mal…      0
6  Oalah pantes tak pantau di situbondo kota udah…      0
7               @@AhmadWildani mas pakai kamera apa      0
8  @@AhmadWildani skrg president class malah suda…      0
9  Di Jawa jam 6 sudah pagi, tapi di Bali jam 7 u…      0

                  Date
0  2024-10-02T09:05:30Z
1  2024-10-02T09:16:23Z
2  2024-10-02T09:19:02Z
3  2024-10-02T09:33:04Z
4  2024-10-03T01:14:51Z
5  2024-10-03T12:24:10Z
6  2024-10-03T13:32:17Z
7  2024-10-07T10:38:51Z
8  2024-10-16T10:49:51Z
9  2024-11-03T07:26:30Z
```

[236]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1592 entries, 0 to 1591
Data columns (total 7 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Timestamp    1592 non-null   object
 1   Username     1592 non-null   object
 2   VideoID      1592 non-null   object
 3   VideoAuthor  1592 non-null   object
 4   Comment      1592 non-null   object
 5   Likes        1592 non-null   int64
 6   Date         1592 non-null   object
dtypes: int64(1), object(6)
memory usage: 87.2+ KB
```

[237]:
```
df = df.drop(columns=['Timestamp', 'Username', 'Date'])
df
```

[237]:

	VideoID	VideoAuthor	Comment	Likes
0	hGWMaKbP_Ss	Ahmad Wildani	Sayangnya angkatan pagi sudah tidak beroperasi…	34
1	hGWMaKbP_Ss	Ahmad Wildani	Wah iyakah mas ?	6
2	hGWMaKbP_Ss	Ahmad Wildani	@@AhmadWildani line perdana nya 17 September. …	5
3	hGWMaKbP_Ss	Ahmad Wildani	@@ParahitaBali okupansinya ga maksimal. Karena…	0
4	hGWMaKbP_Ss	Ahmad Wildani	Tadi saya mau pesen ini ke agen, info agen sud…	0
…	…	…	…	…
1587	NVQ-58Z7UB4	Bima Rahmatulloh	Hadir	1
1588	NVQ-58Z7UB4	Bima Rahmatulloh	Mantap masss	2
1589	NVQ-58Z7UB4	Bima Rahmatulloh	Ke3	1
1590	NVQ-58Z7UB4	Bima Rahmatulloh	assalamulaikum mas	1
1591	NVQ-58Z7UB4	Bima Rahmatulloh	pertama komen	2

[1592 rows x 4 columns]

1.2.1 cleansing

prepare handling kata alay

[238]:
```
gakbaku = pd.read_csv('new_kamusalay.csv', header=None, encoding='latin-1')
gakbaku = gakbaku.rename(columns={0: 'original', 1: 'replacement'})
gakbaku.head(10)
```

[238]:

	original	replacement
0	anakjakartaasikasik	anak jakarta asyik asyik
1	pakcikdahtua	pak cik sudah tua
2	pakcikmudalagi	pak cik muda lagi
3	t3tapjokowi	tetap jokowi
4	3x	tiga kali
5	aamiin	amin
6	aamiinn	amin
7	aamin	amin

```
8          aammiin                    amin
9             abis                   habis
```

[239]:
```
mapping_gakbaku = dict(zip(gakbaku['original'], gakbaku['replacement']))
```

[240]:
```
mapping_gakbaku.update({
    'yg': 'yang',
    'yng': 'yang',
    'skrg': 'sekarang',
    'skrng': 'sekarang',
    'sy': 'saya',
    'sya': 'saya',
    'aq': 'saya',
    'aku': 'saya',
    'q': 'saya',
    'gk': 'tidak',
    'tdk': 'tidak',
    'kmrn': 'kemarin',
    'kmrin': 'kemarin',
    'kemaren': 'kemarin',
    'kmren': 'kemarin',
    'dr': 'dari',
    'dri': 'dari',
    'jln': 'jalan',
    'jl': 'jalan',
    'udh': 'sudah',
    'sdh': 'sudah',
    'dh': 'sudah',
    'ms': 'mas',
    'mz': 'mas',
    'dket': 'dekat',
    'dkt': 'dekat',
    'ad': 'ada',
    'bgks': 'bungkus',
    'bngks': 'bungkus',
    'brgkt': 'berangkat',
    'brngkt': 'berangkat',
    'dg': 'dengan',
    'dgn': 'dengan',
    'dlu': 'dulu',
    'dl': 'dulu',
    'org': 'orang',
    'orng': 'orang',
    'klu': 'kalau',
    'kalo': 'kalau',
    'klo': 'kalau',
    'kl': 'kalau',
```

```
    'nek': 'kalau',
    'nk': 'kalau',
    'ky': 'kayak',
    'kyk': 'kayak',
    'kya': 'kayak',
    'kek': 'kayak',
    'd': 'di',
    'g': 'tidak',
    'y': 'ya',
    'gak': 'tidak',
    'ga': 'tidak',
    'nggak': 'tidak',
    'ngga': 'tidak',
    'ngk': 'tidak',
    'ndk': 'tidak',
    'ndak': 'tidak',
    'ngak': 'tidak',
    'ng': 'tidak',
    'brp': 'berapa',
    'brpa': 'berapa',
    'brpha': 'berapa',
    'brpan': 'berapaan',
    'brpkah': 'berapakah',
    'lgsg': 'langsung',
    'lnsng': 'langsung',
    'nntn': 'tonton',
    'dpt': 'dapat',
    'dpn': 'depan',
    'bg': 'bang',
    'bng': 'bang',
    'abg': 'bang',
    'abgda': 'bang',
    'abangda': 'bang',
    'slm': 'salam',
    'smg': 'semoga',
    'sllu': 'selalu',
    'sll': 'selalu',
    'bwt': 'buat',
    'lg': 'lagi',
    'lgi': 'lagi',
    'mnh': 'lagi',
    'maneh': 'lagi',
    'th': 'tahun',
    'thn': 'tahun',
    'thun': 'tahun',
    'hri': 'hari',
    'km': 'kamu',
```

```
    'kmu': 'kamu',
    'prnh': 'pernah',
    'blm': 'belum',
    'blum': 'belum',
    'blom': 'belum',
    'lom': 'belum',
    'emng': 'memang',
    'emg': 'memang',
    'mmg': 'memang',
    'emang': 'memang',
    'hdr': 'hadir',
    'hdir': 'hadir',
    'jg': 'juga',
    'jga': 'juga',
    'tp': 'tapi',
    'tpi': 'tapi',
    'ap': 'apa',
    'mn': 'mana',
    'mna': 'mana',
    'jgn': 'jangan',
    'jan': 'jangan',
    'mntap': 'mantap',
    'mntp': 'mantap',
    '&': 'dan',
    'rb': 'ribu',
    'rbu': 'rabu',
    'byk': 'banyak',
    'bnyk': 'banyak',
    'bnyak': 'banyak',
    'msh': 'masih',
    'msih': 'masih',
    'krn': 'karena',
    'krna': 'karena',
    'jd': 'jadi',
    'jdi': 'jadi',
    'bgt': 'banget',
    'bgtu': 'begitu',
    'gitu': 'begitu',
    'bngt': 'banget',
    'bnget': 'banget',
    'lwt': 'lewat',
    'kpn': 'kapan',
    'yt': 'youtube',
    'hbs': 'habis',
    'hbis': 'habis',
    'smp': 'sampai',
    'ampe': 'sampai',
```

```
    'sampe': 'sampai',
    'psr': 'pasar',
    'ps': 'pasar',
    'o500r': 'bus mercedes',
    '0500r': 'bus mercedes',
    'o 500 r': 'bus mercedes',
    'osoor': 'bus mercedes',
    'osor': 'bus mercedes',
    'o500rs': 'bus mercedes',
    '0500rs': 'bus mercedes',
    'o 500 rs': 'bus mercedes',
    '1836': 'bus mercedes',
    '1526': 'bus mercedes',
    'mb 1836': 'bus mercedes',
    'rk8': 'bus hino',
    'rk 8': 'bus hino',
    'rm280': 'bus hino',
    'rm 280': 'bus hino',
    'rn285': 'bus hino',
    'rn 285': 'bus hino',
    'hr': 'haryanto',
    'jkt': 'jakarta',
    'mlg': 'malang',
    'dps': 'denpasar',
    'juragan99': 'juragan 99',
    'j99': 'juragan 99',
    'donk': 'dong',
    'adi putro': 'adiputro',
    'pntng': 'penting',
    'pntg': 'penting',
    'ptg': 'penting',
    'koq': 'kok',
    'lbh': 'lebih',
    'lbih': 'lebih',
    'hrsnya': 'harusnya',
    'br': 'baru',
    'bru': 'baru',
    'tlg': 'tolong',
    'gt': 'begitu',
    'gtu': 'begitu',
    'kls': 'kelas',
    'klas': 'kelas',
    'clas': 'kelas',
    'class': 'kelas',
    'mo': 'mau',
    'mw': 'mau',
    'syg': 'sayang',
```

```
    'syang': 'sayang',
    'hrs': 'harus',
    'ni': 'ini',
    'sinjy': 'sinar jaya',
    'sinjay': 'sinar jaya',
    'bs': 'bisa',
    'laen': 'lain',
    'brangjat': 'berangkat',
    'brangkat': 'berangkat',
    'bis': 'bus'
})
```

cek data pada kata

[241]:
```
df['Comment'].isna().sum()
```

[241]:
```
0
```

[242]:
```
#cek label berdasarkan kreatornya -> kreator 3 orang tpi jumlah videonya tetep␣
↪5 vid
prep = df['VideoAuthor'].value_counts()
print(prep)
```

```
VideoAuthor
Andriawan Pratikto    736
Ahmad Wildani         449
Bima Rahmatulloh      407
Name: count, dtype: int64
```

[243]:
```
#cek frequent word

cnt = Counter()
for text in df["Comment"].values:
    for word in text.split():
        cnt[word] += 1

cnt.most_common(20)
```

[243]:
```
[('mas', 389),
 ('di', 239),
 ('yg', 203),
 ('nya', 192),
 ('naik', 174),
 ('dan', 155),
 ('27', 153),
 ('bus', 148),
 ('yang', 144),
```

```
 ('ada', 138),
 ('ke', 114),
 ('Mas', 110),
 ('ya', 102),
 ('dari', 100),
 ('ini', 96),
 ('trans', 94),
 ('itu', 92),
 ('bisa', 88),
 ('bis', 86),
 ('sama', 85)]
```

[244]:
```
#cek rare word

cnt = Counter()
for text in df["Comment"].values:
    for word in text.split():
        cnt[word] += 1

n_rare_words = 20
cnt.most_common()[:-n_rare_words-1:-1]
```

[244]:
```
[('assalamulaikum', 1),
 ('Ke3', 1),
 ('hadirrr', 1),
 ('bandungan ', 1),
 ('Bimm', 1),
 ('selalu ', 1),
 ('tunggu2 ', 1),
 ('kedua', 1),
 ('Komentar', 1),
 ('coba ', 1),
 ('Menarik', 1),
 ('iguazu', 1),
 ('copyraid', 1),
 ('Gimana', 1),
 (' ', 1),
 ('Gass', 1),
 ('@@syafaatarya9074', 1),
 ('ini ', 1),
 ('lama..', 1),
 ('@@budipangestu.', 1)]
```

proses

[245]:
```
import re
import string
```

```
from num2words import num2words
import emoji
from Sastrawi.StopWordRemover.StopWordRemoverFactory import␣
 ↪StopWordRemoverFactory, StopWordRemover, ArrayDictionary
from nltk.tokenize import word_tokenize

factory = StopWordRemoverFactory()
stopword = factory.create_stop_word_remover()
#text preprocessing

def lowercasing(text):
  return text.lower()

def remove_mentions(text):
    return re.sub(r'@\w+', '', text)

def remove_timestamp_format(text):
    return re.sub(r'\b\d{1,2}:\d{2}\b', '', text)

def remove_punctuation(text):
    text = re.sub(r'(?<!\w)-{1}(?!\w)', ' ', text)
    translator = str.maketrans('', '', string.punctuation.replace('-', ''))
    return text.translate(translator)

def combine_27_trans(text):
    return re.sub(r'\b(?:27 trans|trans 27)\b', '27trans', text)

def normalize_gakbaku(text):
    return ' '.join([mapping_gakbaku[word] if word in mapping_gakbaku else word␣
 ↪for word in text.split()])

def combine_ribu(text):
    return re.sub(r'\b(\d+)\s+ribu\b', r'\1ribu', text)

def convert_numbers_to_words(text):
    numbers = re.findall(r'\b\d+\b', text)
    for number in numbers:
        word = num2words(int(number), lang='id')
        text = re.sub(rf'\b{number}\b', word, text, count=1)
    return text

def remove_emoji(string):
    emoji_pattern = re.compile("["
                               u"\U0001F600-\U0001F64F"  # emoticons
                               u"\U0001F300-\U0001F5FF"  # symbols & pictographs
                               u"\U0001F680-\U0001F6FF"  # transport & map␣
 ↪symbols
```

```
                                u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
                                u"\U00002500-\U00002BEF"  # chinese char
                                u"\U00002702-\U000027B0"
                                u"\U00002702-\U000027B0"
                                u"\U000024C2-\U0001F251"
                                u"\U0001f926-\U0001f937"
                                u"\U00010000-\U0010ffff"
                                u"\u2640-\u2642"
                                u"\u2600-\u2B55"
                                u"\u200d"
                                u"\u23cf"
                                u"\u23e9"
                                u"\u231a"
                                u"\ufe0f"  # dingbats
                                u"\u3030"
                                "]+", flags=re.UNICODE)
    return emoji_pattern.sub(r'', string)

def remove_html_tags(text):
    # Regex untuk mencocokkan tag HTML
    html_tag_pattern = r'<[^>]+>'
    return re.sub(html_tag_pattern, '', text)

def remove_urls(text):
    # Regex untuk mencocokkan URL
    url_pattern = r'http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\\(\\),]|(?:
 ↪%[0-9a-fA-F][0-9a-fA-F]))+'
    text = re.sub(url_pattern, '', text)
    # Hapus tag HTML setelah menghapus URL
    text = remove_html_tags(text)
    return text

def remove_double_whitespace(text):
    return re.sub(r'\s+', ' ', text).strip()

nltk.download('stopwords')
stop_words = set(stopwords.words('indonesian'))

def remove_stopwords(text):
    words = text.split()
    filtered_words = [word for word in words if word.lower() not in stop_words]
    return ' '.join(filtered_words)

factory = StopWordRemoverFactory()
stopword = factory.create_stop_word_remover()
stop_factory = StopWordRemoverFactory().get_stop_words()
more_stopword = ['mas', 'nya', 'ya', 'banget', 'bang', 'sih']
```

```
factory = StopWordRemoverFactory()
default_stopwords = factory.get_stop_words()
all_stopwords = default_stopwords + more_stopword

dictionary = ArrayDictionary(all_stopwords)
stopword_remover = StopWordRemover(dictionary)

def custom_stopwords(text):
    text = stopword_remover.remove(text)
    return text

#define preprocessing

def cleansing(text):
    text = lowercasing(text)
    text = remove_mentions(text)
    text = remove_timestamp_format(text)
    text = remove_punctuation(text)
    text = combine_27_trans(text)
    text = normalize_gakbaku(text)
    text = combine_ribu(text)
    text = convert_numbers_to_words(text)
    text = remove_emoji(text)
    text = remove_html_tags(text)
    text = remove_urls(text)
    text = remove_double_whitespace(text)
    text = remove_stopwords(text)
    text = custom_stopwords(text)
    return text
```

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

[246]:
```
df['clean_komen'] = df['Comment'].apply(lambda text: cleansing(text))
```

cross-check

[247]:
```
df['clean_komen']
```

[247]:
```
0       sayangnya angkatan pagi beroperasi trip angkat…
1                                                iyakah
2       line perdana tujuh belas september enam belas …
3           okupansinya maksimal pagi bali beda pagi jawa
4                       pesan agen info agen beroperasi
                              …
1587                                             hadir
1588                                       mantap masss
1589                                               ke3
```

```
1590                                          assalamulaikum
1591                                                   komen
Name: clean_komen, Length: 1592, dtype: object
```

[248]:
```
#cek frequent word

cnt = Counter()
for text in df["clean_komen"].values:
    for word in text.split():
        cnt[word] += 1

cnt.most_common(20)
```

[248]:
```
[('bus', 326),
 ('27trans', 193),
 ('kelas', 101),
 ('puluh', 95),
 ('mantap', 89),
 ('malang', 83),
 ('hadir', 83),
 ('trip', 81),
 ('coba', 74),
 ('jakarta', 69),
 ('mas', 67),
 ('po', 59),
 ('keren', 58),
 ('ratus', 55),
 ('presiden', 54),
 ('harga', 51),
 ('pakai', 50),
 ('andriawan', 48),
 ('bima', 46),
 ('hino', 40)]
```

[249]:
```
#cek rare word

cnt = Counter()
for text in df["clean_komen"].values:
    for word in text.split():
        cnt[word] += 1

n_rare_words = 20
cnt.most_common()[:-n_rare_words-1:-1]
```

[249]:
```
[('assalamulaikum', 1),
 ('ke3', 1),
 ('bandungan', 1),
```

```
 ('bimm', 1),
 ('tunggu2', 1),
 ('iguazu', 1),
 ('copyraid', 1),
 ('kaliberlomba', 1),
 ('youterbs', 1),
 ('keduluan', 1),
 ('report', 1),
 ('tek', 1),
 ('bener', 1),
 ('menemukan', 1),
 ('post', 1),
 ('sekian', 1),
 ('curhat', 1),
 ('okok', 1),
 ('setipis', 1),
 ('menipis', 1)]
```

1.2.2 stemming

[250]:
```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

factory = StemmerFactory()
stemmer = factory.create_stemmer()

def stem_text(text):
    return stemmer.stem(text)

df['clean_lemmaStem'] = df['clean_komen'].apply(stem_text)
```

1.2.3 tokenize data

[251]:
```
nltk.download('punkt')  # Mendownload resource punkt untuk tokenisasi

df['tokenizer_komen'] = df['clean_lemmaStem'].apply(lambda x: word_tokenize(x))
```

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

[252]:
```
df['tokenizer_komen']
```

[252]:
```
0       [sayang, angkat, pagi, operasi, trip, angkat, …
1                                                 [iya]
2       [line, perdana, tujuh, belas, september, enam,…
3       [okupansinya, maksimal, pagi, bal, beda, pagi,…
4                    [pesan, agen, info, agen, operasi]
                              …
```

```
1587                                                [hadir]
1588                                          [mantap, masss]
1589                                                  [ke3]
1590                                         [assalamulaikum]
1591                                                [komen]
Name: tokenizer_komen, Length: 1592, dtype: object
```

1.2.4 preprocessing result

[253]:
```
df[['VideoAuthor', 'Likes', 'Comment', 'clean_komen', 'clean_lemmaStem',␣
 ↪'tokenizer_komen']].sample(30)
```

[253]:

	VideoAuthor	Likes	Comment	
117	Ahmad Wildani	3	Ditemani oleh ibu negara	
1438	Bima Rahmatulloh	0	@@jakajakariya4890 \nUmume wong Indonesia mang…	
797	Andriawan Pratikto	0		
1026	Andriawan Pratikto	0		
1060	Andriawan Pratikto	1		
455	Andriawan Pratikto	0		
1493	Bima Rahmatulloh	1		
1389	Bima Rahmatulloh	1		
537	Andriawan Pratikto	1		
496	Andriawan Pratikto	1		
1153	Andriawan Pratikto	1		
969	Andriawan Pratikto	26		
477	Andriawan Pratikto	0		
286	Ahmad Wildani	1		
155	Ahmad Wildani	2		
1356	Bima Rahmatulloh	0		
743	Andriawan Pratikto	0		
810	Andriawan Pratikto	0		
1238	Bima Rahmatulloh	0		
472	Andriawan Pratikto	0		
132	Ahmad Wildani	3		
136	Ahmad Wildani	3		
781	Andriawan Pratikto	1		
1415	Bima Rahmatulloh	7		
681	Andriawan Pratikto	5		
1259	Bima Rahmatulloh	0		
959	Andriawan Pratikto	1		
44	Ahmad Wildani	0		
54	Ahmad Wildani	4		
65	Ahmad Wildani	3		

```
797
1026  Sehat2 terus mas tito,,semoga cepet sembuh dar…
1060  Keren… Sangat profesional nyonten… Gw …
455
1493  Bang Bima kapan trip bareng rosin double decke…
1389  Mas \nCoba bus garuda mas patas , chasis Mb 0 …
537   Fix saya simpulkan owner 27 TRANS waktu masih …
496   bedanya 27trans yg urban sama executive apa ya…
1153          Tambah lemu… Wes ikut meramaikan
969   Style modif nya yang saya pribadi suka. Modif …
477   Keinget dulu waktu SMA sering banget beli dist…
286   Aku malah salfok sama bis kids panda pada deti…
155                                              Solo
1356                          Kota Sejuta Bunga Hadir
743   @@dwi.prasetyono soalnya kontennya jadi monoto…
810                                    Ak sangat suka
1238                                   Servisnya biasa
472   Trip Bubulak-Surabaya tgl 3 Jan 2022  jam 14.3…
132              Dari Batu Jomba\nHadir bos ku..\n
136                            Sidrap nderek nonton lurr
781     Mantapp mas ttp semangat dan sukses selalu mas
1415                        4:57 ora oren ora keren
681   Kang tito, saya setuju explore makanan khas da…
1259  Nomer kode A2 yg di naiki youtuber dpc prodakc…
959   Wah Itu ada si hijau legenda melintas kalo bol…
44                                   Wah presiden clas
54    Single glass  \nNon "selendang"  \nNon sekat  …
65    Harus muter dulu kalo lewat tengah arusnya bis…

                                            clean_komen  \
117                                   ditemani negara
1438  umume orang indonesia mangan ngganggo tangan t…
797
1026   sehat sehat titosemoga cepat sembuh batuk pileky
1060  keren profesional nyonten gue serasa ikutan me…
455
1493  bima trip bareng rosin double decker adiputro …
1389  coba bus garuda patas chasis mb nol ratus r bu…
537   fix simpulkan owner 27trans kecilnya rental pa…
496                bedanya 27trans urban executive mas
1153                                        meramaikan
969   gaya modif pribadi suka modif fungsi peningkat…
477   beli distro inspired puluh tujuh genre metal s…
286   salah fokus bus kids panda detik ribu ratus puluh
155                                                solo
1356                          kota sejuta bunga hadir
743   prasetyono kontennya monoton bus po berulang u…
```

```
810                                                    suka
1238                                                servisnya
472   trip bubulak-surabaya tanggal ribu puluh jam s…
132                              batu jomba hadir bos ku
136                           sidrap nderek menonton lurr
781                                mantap semangat sukses
1415                                           oren keren
681   kang tito setuju explore makanan khas daerah f…
1259  nomer kode a2 naiki youtuber dewan pimpinan ca…
959                   hijau legenda melintas tau beli po
44                                        presiden kelas
54       single glass non selendang non sekat dispenser
65            memuter arusnya memakan bahan bakar kapal

                                          clean_lemmaStem  \
117                                           tani negara
1438  umume orang indonesia mangan ngganggo tangan t…
797
1026   sehat sehat titosemoga cepat sembuh batuk pileky
1060  keren profesional nyonten gue serasa ikut lancong
455
1493  bima trip bareng rosin double decker adiputro …
1389  coba bus garuda patas chasis mb nol ratus r bu…
537   fix simpul owner 27trans kecil rental pasar ma…
496                     beda 27trans urban executive mas
1153                                                ramai
969   gaya modif pribadi suka modif fungsi tingkat l…
477   beli distro inspired puluh tujuh genre metal s…
286   salah fokus bus kids panda detik ribu ratus puluh
155                                                  solo
1356                              kota juta bunga hadir
743   prasetyono konten monoton bus po ulang ulang u…
810                                                    suka
1238                                                servis
472   trip bubulak-surabaya tanggal ribu puluh jam r…
132                              batu jomba hadir bos ku
136                            sidrap nderek tonton lurr
781                               mantap semangat sukses
1415                                           oren keren
681   kang tito tuju explore makan khas daerah favor…
1259  nomer kode a2 naik youtuber dewan pimpin caban…
959                     hijau legenda lintas tau beli po
44                                        presiden kelas
54       single glass non selendang non sekat dispenser
65                    puter arus makan bahan bakar kapal

                                          tokenizer_komen
```

```
117                                      [tani, negara]
1438  [umume, orang, indonesia, mangan, ngganggo, ta…
797                                                  []
1026  [sehat, sehat, titosemoga, cepat, sembuh, batu…
1060  [keren, profesional, nyonten, gue, serasa, iku…
455                                                  []
1493  [bima, trip, bareng, rosin, double, decker, ad…
1389  [coba, bus, garuda, patas, chasis, mb, nol, ra…
537   [fix, simpul, owner, 27trans, kecil, rental, p…
496              [beda, 27trans, urban, executive, mas]
1153                                           [ramai]
969   [gaya, modif, pribadi, suka, modif, fungsi, ti…
477   [beli, distro, inspired, puluh, tujuh, genre, …
286   [salah, fokus, bus, kids, panda, detik, ribu, …
155                                            [solo]
1356                        [kota, juta, bunga, hadir]
743   [prasetyono, konten, monoton, bus, po, ulang, …
810                                            [suka]
1238                                         [servis]
472   [trip, bubulak-surabaya, tanggal, ribu, puluh,…
132                       [batu, jomba, hadir, bos, ku]
136                      [sidrap, nderek, tonton, lurr]
781                          [mantap, semangat, sukses]
1415                                    [oren, keren]
681   [kang, tito, tuju, explore, makan, khas, daera…
1259  [nomer, kode, a2, naik, youtuber, dewan, pimpi…
959             [hijau, legenda, lintas, tau, beli, po]
44                                   [presiden, kelas]
54    [single, glass, non, selendang, non, sekat, di…
65             [puter, arus, makan, bahan, bakar, kapal]
```

1.3 (1c) frequent word analysis

1.3.1 frekuensi

[254]:
```
all_words = [word for tokens in df['tokenizer_komen'] for word in tokens]
common_words = pd.Series(all_words).value_counts().head(20)

# Menampilkan 20 kata yang paling umum
print(common_words)
```

```
bus         341
27trans     193
puluh       109
kelas       109
coba         92
mantap       89
trip         89
```

```
malang       86
makan        85
hadir        83
jakarta      69
mas          68
ratus        60
keren        59
harga        59
po           59
presiden     55
nyaman       55
pakai        52
tunggu       49
Name: count, dtype: int64
```

ini adalah kumpulan top 20 kata yang sering muncul pada komentar

1.3.2 wordcloud

[255]:
```
!pip install wordcloud
```

```
Requirement already satisfied: wordcloud in /usr/local/lib/python3.10/dist-
packages (1.9.3)
Requirement already satisfied: numpy>=1.6.1 in /usr/local/lib/python3.10/dist-
packages (from wordcloud) (1.26.4)
Requirement already satisfied: pillow in /usr/local/lib/python3.10/dist-packages
(from wordcloud) (10.4.0)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.10/dist-
packages (from wordcloud) (3.8.0)
Requirement already satisfied: contourpy>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->wordcloud) (1.3.0)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-
packages (from matplotlib->wordcloud) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->wordcloud) (4.54.1)
Requirement already satisfied: kiwisolver>=1.0.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->wordcloud) (1.4.7)
Requirement already satisfied: packaging>=20.0 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->wordcloud) (24.1)
Requirement already satisfied: pyparsing>=2.3.1 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->wordcloud) (3.2.0)
Requirement already satisfied: python-dateutil>=2.7 in
/usr/local/lib/python3.10/dist-packages (from matplotlib->wordcloud) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.7->matplotlib->wordcloud) (1.16.0)
```

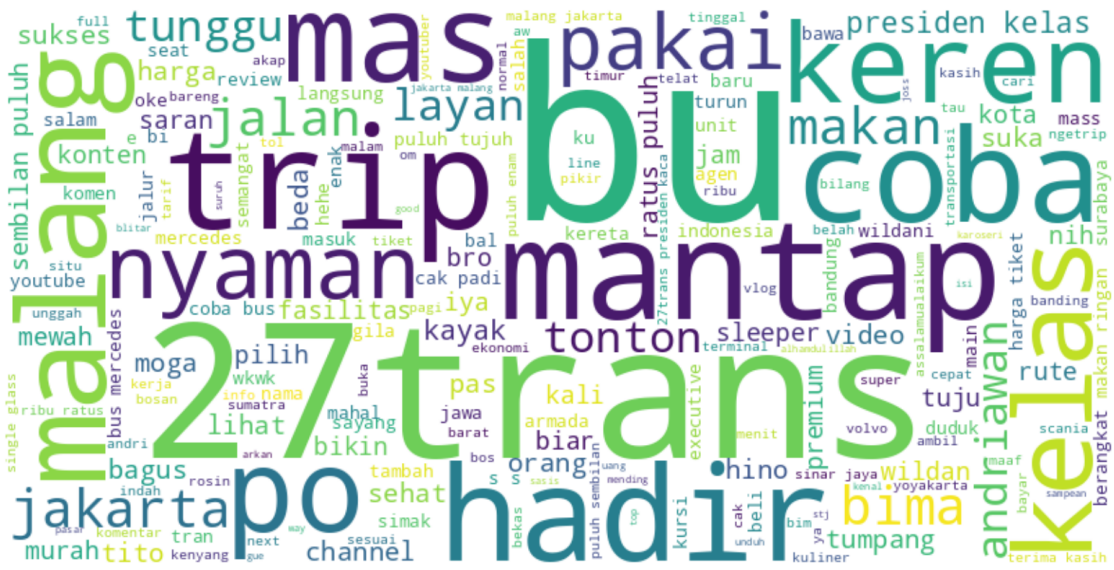
[256]:
```
from wordcloud import WordCloud
```

[257]:
```
import matplotlib.pyplot as plt

all_text = ' '.join(all_words)

# Membuat Word Cloud
wordcloud = WordCloud(width=800, height=400, background_color='white',␣
 ↪colormap='viridis').generate(all_text)

# Tampilkan Word Cloud
plt.figure(figsize=(20, 10))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```



[258]:
```
all_words = [word for tokens in df['tokenizer_komen'] for word in tokens]
common_words = pd.Series(all_words).value_counts().head(10)
common_words
```

[258]:
```
bus        341
27trans    193
puluh      109
kelas      109
coba        92
mantap      89
trip        89
malang      86
makan       85
hadir       83
```

```
Name: count, dtype: int64
```

kalau kita lihat, kata-kata yang ada pada wordcloud dan frequensi ini menjelaskan bahwa top 10 common frequent words menjelaskan deskripsi topik dari kelima video yang ada. 10 'Kata Kunci' yang ada itu menggambarkan kalau konten yang terkandung adalah:

konten bus, busnya bus 27trans, kata *puluh* cenderung mengacu pada harga kelas, pelayanan yang mantap, kontennya tentang trip, rute busnya yang ke malang, kata *makan* yang merujuk pada servis makan, kata *hadir* merujuk pada asal viewers & subscriber

1.4 (1d) vektorisasi

[259]:
```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
```

[260]:
```
# Metode 1: CountVectorizer
count_vectorizer = CountVectorizer()
count_vectors = count_vectorizer.fit_transform(df['clean_lemmaStem'])
```

[261]:
```
# Menampilkan hasil vektorisasi CountVectorizer
print("CountVectorizer Feature Names:")
print(count_vectorizer.get_feature_names_out())

print("\nCountVectorizer Matrix:")
print(count_vectors.toarray())
```

```
CountVectorizer Feature Names:
['100rb' '104rb' '10hari' … 'yuotup' 'yus' 'zaman']

CountVectorizer Matrix:
[[0 0 0 … 0 0 0]
 [0 0 0 … 0 0 0]
 [0 0 0 … 0 0 0]
 …
 [0 0 0 … 0 0 0]
 [0 0 0 … 0 0 0]
 [0 0 0 … 0 0 0]]
```

[262]:
```
# Metode 2: TfidfVectorizer
tfidf_vectorizer = TfidfVectorizer()
tfidf_vectors = tfidf_vectorizer.fit_transform(df['clean_lemmaStem'])
```

[263]:
```
# Menampilkan hasil vektorisasi TfidfVectorizer
print("\nTfidfVectorizer Feature Names:")
print(tfidf_vectorizer.get_feature_names_out())
print("\nTfidfVectorizer Matrix:")
print(tfidf_vectors.toarray())
```

```
TfidfVectorizer Feature Names:
['100rb' '104rb' '10hari' … 'yuotup' 'yus' 'zaman']

TfidfVectorizer Matrix:
[[0. 0. 0. … 0. 0. 0.]
 [0. 0. 0. … 0. 0. 0.]
 [0. 0. 0. … 0. 0. 0.]
 …
 [0. 0. 0. … 0. 0. 0.]
 [0. 0. 0. … 0. 0. 0.]
 [0. 0. 0. … 0. 0. 0.]]
```

[270]:
```python
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score,␣
 ↪f1_score

# Membuat label contoh untuk data teks
df['label'] = df['VideoID'].apply(lambda x: 1 if 'k' in x.lower() else 0)  #␣
 ↪Label biner untuk contoh ini

# Fungsi untuk melatih dan mengevaluasi model
def evaluate_model(X, y):
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,␣
 ↪random_state=42)

    # Menggunakan Random Forest sebagai model
    model = RandomForestClassifier(n_estimators=200, random_state=42)
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    # Menghitung metrik evaluasi
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred, average='binary')
    recall = recall_score(y_test, y_pred, average='binary')
    f1 = f1_score(y_test, y_pred, average='binary')

    return accuracy, precision, recall, f1

# Evaluasi dengan CountVectorizer
X_count = count_vectors
y = df['label']
acc_count, prec_count, rec_count, f1_count = evaluate_model(X_count, y)

# Evaluasi dengan TfidfVectorizer
X_tfidf = tfidf_vectors
```

```
acc_tfidf, prec_tfidf, rec_tfidf, f1_tfidf = evaluate_model(X_tfidf, y)

# Menampilkan hasil evaluasi
print("Performa Model dengan CountVectorizer:")
print(f"Accuracy: {acc_count:.2f}, Precision: {prec_count:.2f}, Recall:␣
↪{rec_count:.2f}, F1 Score: {f1_count:.2f}\n")

print("Performa Model dengan TfidfVectorizer:")
print(f"Accuracy: {acc_tfidf:.2f}, Precision: {prec_tfidf:.2f}, Recall:␣
↪{rec_tfidf:.2f}, F1 Score: {f1_tfidf:.2f}")
```

```
Performa Model dengan CountVectorizer:
Accuracy: 0.69, Precision: 0.59, Recall: 0.41, F1 Score: 0.48

Performa Model dengan TfidfVectorizer:
Accuracy: 0.69, Precision: 0.60, Recall: 0.33, F1 Score: 0.43
```

1.5 video

[ ]:
```
#https://drive.google.com/file/d/1YXZaTULMKQrzdJopZkeyrTPv_UByEXBR/view?
↪usp=sharing;
```