LAPORAN TUGAS CRAWLING DAN PREPROCESSING DATA TEXT
MATA KULIAH TEXT MINING & NATURAL LANGUAGE PROCESSING

Tim Penyusun:

1. <5231811022> <Lathif Ramadhan>
2. <5231811029> <Andini Angel Meivita>
3. <5231811033> <Rama Panji Nararendra>
4. <5231811036> <Giffari Riyanda Pradithya>

PROGRAM STUDI SAINS DATA PROGRAM SARJANA
FAKULTAS SAINS & TEKNOLOGI
UNIVERSITAS TEKNOLOGI YOGYAKARTA
2025

1. **Hastag yang digunakan:** #ramadhan
2. **Tuliskan langkah-langkah crawling data text:**
   a. **Tuliskan/screenshotkan codingnya**

      Jawab:

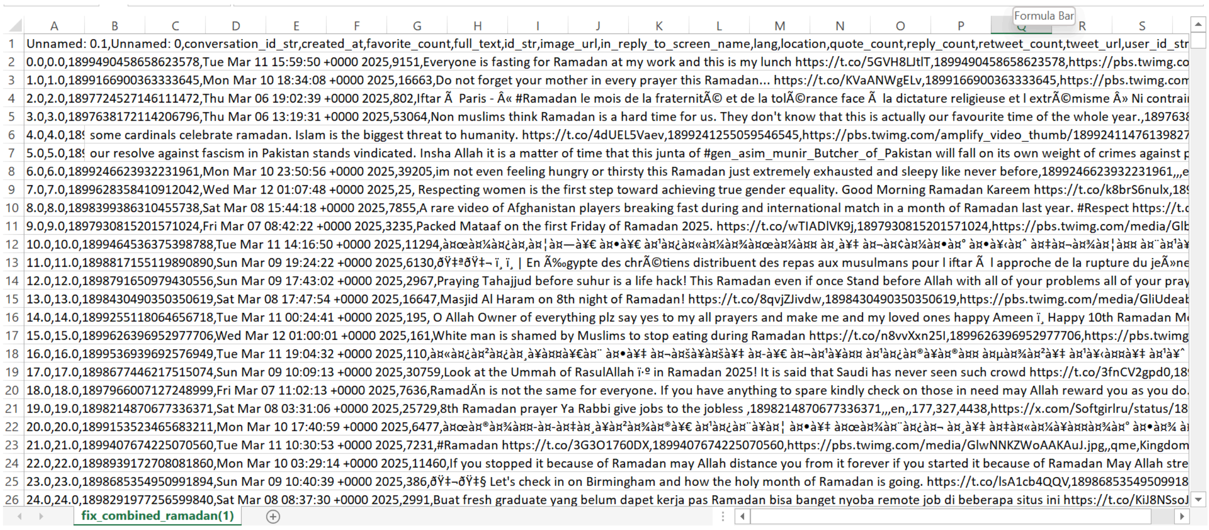
      Link file kodenya:

      https://colab.research.google.com/drive/1wxDKwRGiu_pd9CDx6kNHN6I8RfUdxLrS#scrollTo=o2wMKSTWhoNW

```
# instal Node.js dan npm (Node Package Manager) di sistem
!curl -sL https://deb.nodesource.com/setup_18.x | sudo -E
bash -
!sudo apt-get install -y nodejs

# nama file dari data yang berhasil dikumpulkan
data = "ramadan_1.csv"

# kata kunci dari data yang ingin dicari
search_keyword = "Ramadan"

# limit baris pencarian
limit = 200

# jalankan proses crawling tweet dengan bantuan tweet-harvest
menggunakan Twitter API token.
!npx --yes tweet-harvest@2.6.1 -o "ramadan_1.csv" -s
"{Ramadan}" -l {100} --token "-----------------"
```

   b. **Tuliskan/screenshotkan hasil crawlingnya**

      Jawab:



```
Unnamed: 0.1,Unnamed: 0,conversation_id_str,created_at,favorite_count,full_text,id_str,image_url,in_reply_to_screen_name,lang,location,quote_count,reply_count,retweet_count,tweet_url,user_id_str
0.0,0.0,1899490458658623578,Tue Mar 11 15:59:50 +0000 2025,9151,Everyone is fasting for Ramadan at my work and this is my lunch https://t.co/5GVH8LJtlT,1899490458658623578,https://pbs.twimg.co
1.0,1.0,1899166900363333645,Mon Mar 10 18:34:08 +0000 2025,16663,Do not forget your mother in every prayer this Ramadan... https://t.co/KVaANWgELv,1899166900363333645,https://pbs.twimg.com
2.0,2.0,1897724527146111472,Thu Mar 06 19:02:39 +0000 2025,802,Iftar Ã  Paris - Â« #Ramadan le mois de la fraternitÃ© et de la tolÃ©rance face Ã  la dictature religieuse et l extrÃ©misme Â» Ni contrai
3.0,3.0,1897638172114206796,Thu Mar 06 13:19:31 +0000 2025,53064,Non muslims think Ramadan is a hard time for us. They don't know that this is actually our favourite time of the whole year.,1897638
4.0,4.0,18 some cardinals celebrate ramadan. Islam is the biggest threat to humanity. https://t.co/4dUEL5Vaev,1899241255059546545,https://pbs.twimg.com/amplify_video_thumb/18992411476139827
5.0,5.0,18 our resolve against fascism in Pakistan stands vindicated. Insha Allah it is a matter of time that this junta of #gen_asim_munir_Butcher_of_Pakistan will fall on its own weight of crimes against p
6.0,6.0,1899246623932231961,Mon Mar 10 23:50:56 +0000 2025,39205,im not even feeling hungry or thirsty this Ramadan just extremely exhausted and sleepy like never before,1899246623932231961,,,e
7.0,7.0,1899628358410912042,Wed Mar 12 01:07:48 +0000 2025,25, Respecting women is the first step toward achieving true gender equality. Good Morning Ramadan Kareem https://t.co/k8brS6nulx,189
8.0,8.0,1898399386310455738,Sat Mar 08 15:44:18 +0000 2025,7855,A rare video of Afghanistan players breaking fast during and international match in a month of Ramadan last year. #Respect https://t.c
9.0,9.0,1897930815201571024,Fri Mar 07 08:42:22 +0000 2025,3235,Packed Mataaf on the first Friday of Ramadan 2025. https://t.co/wTIADlVK9j,1897930815201571024,https://pbs.twimg.com/media/Glb
10.0,10.0,1899464536375398788,Tue Mar 11 14:16:50 +0000 2025,11294,à¤œà¤¼à¤¿à¤‚à¤¦à¤—à¥€ à¤•à¥€ à¤¹à¤¿à¤«à¤¼à¤¾à¤œà¤¼à¤¤ à¤¸à¥‡ à¤¬à¤¢à¤¼à¤•à¤° à¤•à¥‹à¤ˆ à¤‡à¤¬à¤¾à¤¦à¤¤ à¤¨à¤¹à¥
11.0,11.0,1898817155119890890,Sun Mar 09 19:24:22 +0000 2025,6130,ðŸ‡ªðŸ‡¬ ï¸ ï¸ | En Ã‰gypte des chrÃ©tiens distribuent des repas aux musulmans pour l iftar Ã  l approche de la rupture du jeÃ»ne
12.0,12.0,1898791650979430556,Sun Mar 09 17:43:02 +0000 2025,2967,Praying Tahajjud before suhur is a life hack! This Ramadan even if once Stand before Allah with all of your problems all of your pray
13.0,13.0,1898430490350350619,Sat Mar 08 17:47:54 +0000 2025,16647,Masjid Al Haram on 8th night of Ramadan! https://t.co/8qvjZJivdw,1898430490350350619,https://pbs.twimg.com/media/GliUdeab
14.0,14.0,1899255118064656718,Tue Mar 11 00:24:41 +0000 2025,195, O Allah Owner of everything plz say yes to my all prayers and make me and my loved ones happy Ameen ï¸ Happy 10th Ramadan M
15.0,15.0,1899626396952977706,Wed Mar 12 01:00:01 +0000 2025,161,White man is shamed by Muslims to stop eating during Ramadan https://t.co/n8vvXxn25I,1899626396952977706,https://pbs.twimg
16.0,16.0,1899536939692576949,Tue Mar 11 19:04:32 +0000 2025,110,à¤«à¤¿à¤²à¤¿à¤¸à¥à¤¤à¥€à¤¨ à¤•à¥‡ à¤¬à¤šà¥à¤šà¥‡ à¤-à¥€ à¤¬à¤¹à¥à¤¤ à¤¹à¤¿à¤®à¥à¤®à¤¤ à¤µà¤¾à¤²à¥‡ à¤¹à¥‹à¤¤à¥‡ à¤¹à¥ˆ
17.0,17.0,1898677446217515074,Sun Mar 09 10:09:13 +0000 2025,30759,Look at the Ummah of RasulAllah ï·º in Ramadan 2025! It is said that Saudi has never seen such crowd https://t.co/3fnCV2gpd0,18
18.0,18.0,1897966007127248999,Fri Mar 07 11:02:13 +0000 2025,7636,RamadÄn is not the same for everyone. If you have anything to spare kindly check on those in need may Allah reward you as you do.
19.0,19.0,1898214870677336371,Sat Mar 08 03:31:06 +0000 2025,25729,8th Ramadan prayer Ya Rabbi give jobs to the jobless ,1898214870677336371,,,en,,177,327,4438,https://x.com/Softgirlru/status/18
20.0,20.0,1899153523465683211,Mon Mar 10 17:40:59 +0000 2025,6477,à¤œà¤®à¤¾à¤¤-à¤-à¤‡à¤¸à¥à¤²à¤¾à¤®à¥€ à¤¹à¤¿à¤¨à¥à¤¦ à¤•à¥‡ à¤œà¤¾à¤¨à¤¿à¤¬ à¤¸à¥‡ à¤‡à¤«à¤¼à¥à¤¤à¤¾à¤° à¤•à¤¾ à¤
21.0,21.0,1899407674225070560,Tue Mar 11 10:30:53 +0000 2025,7231,#Ramadan https://t.co/3G3O1760DX,1899407674225070560,https://pbs.twimg.com/media/GlwNNKZWoAAKAuJ.jpg,,qme,Kingdom
22.0,22.0,1898939172708081860,Mon Mar 10 03:29:14 +0000 2025,11460,If you stopped it because of Ramadan may Allah distance you from it forever if you started it because of Ramadan May Allah stre
23.0,23.0,1898685354950991894,Sun Mar 09 10:40:39 +0000 2025,386,ðŸ‡¬ðŸ‡§ Let's check in on Birmingham and how the holy month of Ramadan is going. https://t.co/lsA1cb4QQV,18986853549509918
24.0,24.0,1898291977256599840,Sat Mar 08 08:37:30 +0000 2025,2991,Buat fresh graduate yang belum dapet kerja pas Ramadan bisa banget nyoba remote job di beberapa situs ini https://t.co/KiJ8NSsoJ
```

c. Berapa jumlah record yang didapatkan

Jawab: 6543 baris

3. Tuliskan langkah-langkah preprocessing data text:

Link Colab Coding Cleaning:

https://colab.research.google.com/drive/15PBqlsHWNYC5eV-kj9fXwbNN9nC9whRJ?usp=sharing

Link Colab Case Folding, Tokenizing, Filtering, Stemming, dan Simpan Data Bersih:

https://colab.research.google.com/drive/1pBv_cBklF6W0eB2FAZdkWwRCXI3_dIbM?usp=sharing#scrollTo=eW2NnAO87pD8

a. Tuliskan/screenshotkan coding cleaning

Jawab:

```
import pandas as pd
import glob
import os

path = r'/content/' # use your path
from os import listdir
from os.path import isfile, join
onlyfiles = [f for f in listdir(path) if isfile(join(path,
f))]
```

```
li = []
# print(all_files)S
for filename in onlyfiles:
           df    =    pd.read_csv(   r'/content/'   +   filename,
index_col=None, header=0)
    li.append(df)

frame = pd.concat(li, axis=0, ignore_index=True)

frame.to_csv("fix_combined_ramadan_real.csv", index=False)

df = pd.read_csv('/content/fix_combined_ramadan_real.csv')
df

df.drop_duplicates()
df.to_csv("/content/fix_combined_ramadan.csv", index=False)
```

Cleaning di preprocessing:

```
df.drop_duplicates(subset = 'Tweet', keep = 'first', inplace
= True)

# Memfilter datanya cuma bahasanya yang English
filt = (df_twit['lang'] == 'en')
df_twit = df_twit.loc[filt, :]
df_twit.reset_index(inplace=True)

#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]*")
df['remove_user']
```

b. **Tuliskan/screenshotkan Case folding**

Jawab:

Case folding adalah proses mengubah semua teks menjadi huruf kecil agar lebih konsisten dalam analisis.

→ Di kode ini, case folding terjadi secara otomatis saat tokenizing

```
#tokenize tweets
        tokenizer = TweetTokenizer(preserve_case=True,
strip_handles=True, reduce_len=True)

  tweet_tokens = tokenizer.tokenize(tweet)
  # print(f"Word after tokenizer : {tweet_tokens}")
  tweets_clean = []

  for word in tweet_tokens:
      if (word not in stopwords_english and word not in
emoticons and word not in string.punctuation): # remove
punctuation
      # print(f"Word before stemming: {word}")
      stem_word = stemmer.stem(word) #stemming word
      # print(f"Word after stem: {stem_word}")
      tweets_clean.append(stem_word)
  return tweets_clean
```

Parameter preserve_case=False dalam TweetTokenizer akan secara otomatis mengubah teks menjadi huruf kecil.

Kedua kode ini walaupun fungsinya untuk tokenizing dan stemming tetapi secara tidak langsung mereka juga melakukan case folding, dimana untuk tokenizer preserve_case nya di set ke False untuk di case folding.

Untuk bagian kode yang melakukan case folding di tokenizing adalah

```
tokenizer = TweetTokenizer(preserve_case=True,
strip_handles=True, reduce_len=True)
```

```
tweet_tokens = tokenizer.tokenize(tweet)
```

Dan untuk bagian kode yang melakukan case folding di stemming adalah

```
stem_word = stemmer.stem(word)
```

c. Tuliskan/screenshotkan Tokenizing

Tokenizing adalah proses memecah teks menjadi kata-kata atau token.

Jawab:

→ Bagian kode yang melakukan tokenizing:

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

- `tokenizer.tokenize(tweet)` akan memecah teks menjadi daftar kata-kata individu (token).
- `preserve_case=False` akan membuat kata-kata menjadi huruf kecil (bagian dari case folding juga).

d. Tuliskan/screenshotkan Filtering

Filtering adalah proses membersihkan teks dari karakter atau kata yang tidak diperlukan, seperti angka, URL, tanda baca, stopwords, dan emotikon.

Jawab:

→ Bagian kode yang melakukan filtering:

```
def tweet_clean(tweet):
  #remove angka
  tweet = re.sub('[0-9]+', '', tweet)
  # print(f"ss")
  # 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([
   ':-)', ':)', ';)', ':0)', ':]', '3', ':c)', ':>', '=]',
':^)', ':-D', ':D' '8-D', '8D', '-3', '-3', ':-))', ":'-)"
'x-D', 'xD', 'X-D', 'XD' '>: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=True,
strip_handles=True, reduce_len=True)

  tweet_tokens = tokenizer.tokenize(tweet)
  # print(f"Word after tokenizer : {tweet_tokens}")
  tweets_clean = []

  for word in tweet_tokens:
      if (word not in stopwords_english and word not in
emoticons and word not in string.punctuation): # remove
punctuation
      # print(f"Word before stemming: {word}")
      stem_word = stemmer.stem(word) #stemming word
      # print(f"Word after stem: {stem_word}")
      tweets_clean.append(stem_word)
  return tweets_clean
df['tweet_clean'] = df['remove_user'].apply(lambda x:
tweet_clean(x))

#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))
```

e. Tuliskan/screenshotkan Stemming

Stemming adalah proses mengubah kata menjadi bentuk dasarnya.

Jawab:

→ Bagian kode yang melakukan stemming menggunakan Sastrawi:

```
stem_word = stemmer.stem(word) #stemming word
```

f. Tuliskan/screenshotkan Simpan data text bersih

Jawab:

```
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('ramadan_clean_tweet.csv', encoding='utf8',
index=False)
df.head(10)
```

4. Lampirkan file RAW hasil Crawling dan file bersih hasil preprocessing