Praktikum Big Data Analytics - **Pt (1) Data Preprocessing**

Topik : Word Counting from Tweets

Nama : Catherine V. Pangemanan

NRP : 2C2220008

Kelas : IV/A S1 Sains Data

Text Pre-Processing (0) Import Library & Dataset

Mengimport Library Pandas & NumPy

```
#!pip install pandas
import pandas as pd
import numpy as n
```

+ Code + Text

Mengunggah Dataset `Somethinc.csv`

```
from google.colab import files
uploads = files.upload()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving Somethinc.csv to Somethinc.csv
```

Membaca dan Menampilkan 10 Data Teratas dari Dataset `Somethinc`

```
somethinc = pd.read_csv('Somethinc.csv', delimiter=",")
somethinc.head(10)
```

	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location	quote_count	reply_c
0	1.786430e+18	Fri May 03 23:40:40 +0000 2024	0	@Somethinc4u my birthday wish for somethinc se...	1.786540e+18	NaN	Somethinc4u	in	NaN	0	
1	1.786430e+18	Fri May 03 23:26:16 +0000 2024	0	@Somethinc4u happy birthday somethinc aku sela...	1.786540e+18	NaN	Somethinc4u	in	Yogyakarta, Indonesia	0	
2	1.786430e+18	Fri May 03 23:24:24 +0000 2024	0	@Somethinc4u selamat ulang tahun somethinc! se...	1.786540e+18	NaN	Somethinc4u	in	NaN	0	
3	1.785570e+18	Fri May 03 23:22:07 +0000 2024	0	@Somethinc4u @afgan___ titip salam Min jangan ...	1.786540e+18	NaN	Somethinc4u	in	Indonesia	0	
4	1.786430e+18	Fri May 03 23:19:12 +0000 2024	0	@Somethinc4u Happy Birthday Somethinc ! Semoga...	1.786540e+18	NaN	Somethinc4u	in	Indonesia	0	
5	1.786430e+18	Fri May 03 23:18:07 +0000 2024	0	@Somethinc4u Wihhhh selamat ulang tahun @Somet...	1.786540e+18	NaN	Somethinc4u	in	NaN	0	
6	1.786430e+18	Fri May 03 23:14:06 +0000 2024	0	@Somethinc4u Selamat ulang tahun somethinc sem...	1.786530e+18	NaN	Somethinc4u	in	mulfand	0	
7	1.786530e+18	Fri May 03 23:03:43 +0000 2024	0	WTS / Want To Sell / Jual NEW bukan preloved...	1.786530e+18	https://pbs.twimg.com/media/GMsJZrJa0AA_z7o.jpg	NaN	in	she/her	0	
8	1.786430e+18	Fri May 03 22:48:32 +0000 2024	0	@Somethinc4u Birthday wish buat Somethinc semo...	1.786530e+18	NaN	Somethinc4u	in	NaN	0	
9	1.786430e+18	Fri May 03 22:43:49 +0000 2024	0	@Somethinc4u Happy birthday somethinc terima k...	1.786530e+18	https://pbs.twimg.com/tweet_video_thumb/GMsE7n...	Somethinc4u	in	Surakarta, Jawa Tengah	0	

Text Pre-Processing (1) Membuang Kolom tidak yang dibutuhkan

Index Python dimulai dari 0.

```
somethinc2 = somethinc[somethinc.columns[[3,14]]]
somethinc2
```

	full_text	username
0	@Somethinc4u my birthday wish for somethinc se...	bokhyeonie
1	@Somethinc4u happy birthday somethinc aku sela...	kasihajamilkita
2	@Somethinc4u selamat ulang tahun somethinc! se...	yourrellia
3	@Somethinc4u @afgan___ titip salam Min jangan ...	iyaiyao_
4	@Somethinc4u Happy Birthday Somethinc ! Semoga...	iyaiyao_
...	...	...
101	@ohmybeautybank Somethinc gentle jelly 350 ml:...	magnoliaodyssey
102	Wts preloved skincare Somethinc calm down calm...	illusixxn
103	Somethinc Hooman Breathable Cushion Cover ⭐ 8...	sleepinbeautea
104	mascara somethinc bagus buanget tp sayang keri...	cheerieamour
105	Wts skincare preloved Somethinc calm down pha ...	illusixxn

106 rows × 2 columns

Text Pre-Processing (2) Membuat Kolom `full_text` menjadi Lower Case Text

```
somethinc2['caseFolded_fullText'] = somethinc2['full_text'].str.lower()
somethinc2.head(10)
```

```
<ipython-input-7-e6c96badeaa0>:1: 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/user_guide/indexing.html#returning-a-view-versus-a-copy
  somethinc2['caseFolded_fullText'] = somethinc2['full_text'].str.lower()
```

	full_text	username	caseFolded_fullText
0	@Somethinc4u my birthday wish for somethinc se...	bokhyeonie	@somethinc4u my birthday wish for somethinc se...
1	@Somethinc4u happy birthday somethinc aku sela...	kasihajamilkita	@somethinc4u happy birthday somethinc aku sela...
2	@Somethinc4u selamat ulang tahun somethinc! se...	yourrellia	@somethinc4u selamat ulang tahun somethinc! se...
3	@Somethinc4u @afgan___ titip salam Min jangan ...	iyaiyao_	@somethinc4u @afgan___ titip salam min jangan ...
4	@Somethinc4u Happy Birthday Somethinc ! Semoga...	iyaiyao_	@somethinc4u happy birthday somethinc ! semoga...
5	@Somethinc4u Wihhhh selamat ulang tahun @Somet...	seperlunyaajaa	@somethinc4u wihhhh selamat ulang tahun @somet...
6	@Somethinc4u Selamat ulang tahun somethinc sem...	onyounjm	@somethinc4u selamat ulang tahun somethinc sem...
7	WTS / Want To Sell / Jual NEW bukan preloved...	jjkxzz	wts / want to sell / jual new bukan preloved...
8	@Somethinc4u Birthday wish buat Somethinc semo...	sandikala_	@somethinc4u birthday wish buat somethinc semo...
9	@Somethinc4u Happy birthday somethinc terima k...	khoirulaf_ifah	@somethinc4u happy birthday somethinc terima k...

Text Pre-Processing (3) Membersihkan text yang tidak diperlukan menggunakan Regular Expression (Regex)

```
#Cleaning
import re
import string

def cleaningText(text):
    text = re.sub(r'@[A-Za-z0-9]+', '', text) # remove mentions
    text = re.sub(r'#[A-Za-z0-9]+', '', text) # remove hashtag
    text = re.sub(r'RT[\s]', '', text) # remove RT(Retweet)
    text = re.sub(r"http\S+", '', text) # remove link
    text = re.sub(r'[0-9]+', '', text) # remove numbers
    text = re.sub(r'[^A-Za-z ]+', '', text) # remove all character non alfabet

    text = text.replace('\n', ' ') # remove new line into space
    text = text.translate(str.maketrans('', '', string.punctuation)) # remove all punctuations
    text = text.strip(' ') # remove characters space from both left and right text
    return text

somethinc2['cleaning'] = somethinc2['caseFolded_fullText'].apply(cleaningText)
somethinc2.head(10)
```

```
<ipython-input-8-c36f140352ea>:18: 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/user_guide/indexing.html#returning-a-view-versus-a-copy
  somethinc2['cleaning'] = somethinc2['caseFolded_fullText'].apply(cleaningText)
```

	full_text	username	caseFolded_fullText	cleaning
0	@Somethinc4u my birthday wish for somethinc se...	bokhyeonie	@somethinc4u my birthday wish for somethinc se...	my birthday wish for somethinc semoga produk s...
1	@Somethinc4u happy birthday somethinc aku sela...	kasihajamilkita	@somethinc4u happy birthday somethinc aku sela...	happy birthday somethinc aku selalu cocok pake...
2	@Somethinc4u selamat ulang tahun somethinc! se...	yourrellia	@somethinc4u selamat ulang tahun somethinc! se...	selamat ulang tahun somethinc semoga makin suk...
3	@Somethinc4u @afgan___ titip salam Min jangan ...	iyaiyao_	@somethinc4u @afgan___ titip salam min jangan ...	titip salam min jangan lupa pakai selalu somet...
4	@Somethinc4u Happy Birthday Somethinc ! Semoga...	iyaiyao_	@somethinc4u happy birthday somethinc ! semoga...	happy birthday somethinc semoga bisa ters men...
5	@Somethinc4u Wihhhh selamat ulang tahun @Somet...	seperlunyaajaa	@somethinc4u wihhhh selamat ulang tahun @somet...	wihhhh selamat ulang tahun semoga makin sukse...
6	@Somethinc4u Selamat ulang tahun somethinc sem...	onyounjm	@somethinc4u selamat ulang tahun somethinc sem...	selamat ulang tahun somethinc semoga produknya...
7	WTS / Want To Sell / Jual NEW bukan preloved...	jjkxzz	wts / want to sell / jual new bukan preloved...	wts want to sell jual new bukan preloved mas...
8	@Somethinc4u Birthday wish buat Somethinc semo...	sandikala_	@somethinc4u birthday wish buat somethinc semo...	birthday wish buat somethinc semoga bisa terus...
9	@Somethinc4u Happy birthday somethinc terima k...	khoirulaf_ifah	@somethinc4u happy birthday somethinc terima k...	happy birthday somethinc terima kasih somethin...

Text Pre-Processing (4) Tokenisasi Data Untuk Memisahkan Kalimat Menjadi Kata Per Kata

```
import nltk
nltk.download('punkt')
```

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

```
# Tokenisasi
from nltk.tokenize import word_tokenize

def tokenizingText(text): # Tokeniseing or splitting a string, text into a list of tokens
    text = word_tokenize(text)
    return text

somethinc2['tokenizing'] = somethinc2['cleaning'].apply(tokenizingText)
# tweets = tweets[['text_clean','text_preprocessed']]
somethinc2.head(10)
```

```
<ipython-input-10-aab49e9bbab6>: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/user_guide/indexing.html#returning-a-view-versus-a-copy
  somethinc2['tokenizing'] = somethinc2['cleaning'].apply(tokenizingText)
```

	full_text	username	caseFolded_fullText	cleaning	tokenizing
0	@Somethinc4u my birthday wish for somethinc se...	bokhyeonie	@somethinc4u my birthday wish for somethinc se...	my birthday wish for somethinc semoga produk s...	[my, birthday, wish, for, somethinc, semoga, p...
1	@Somethinc4u happy birthday somethinc aku sela...	kasihajamilkita	@somethinc4u happy birthday somethinc aku sela...	happy birthday somethinc aku selalu cocok pake...	[happy, birthday, somethinc, aku, selalu, coco...
2	@Somethinc4u selamat ulang tahun somethinc! se...	yourrellia	@somethinc4u selamat ulang tahun somethinc! se...	selamat ulang tahun somethinc semoga makin suk...	[selamat, ulang, tahun, somethinc, semoga, mak...
3	@Somethinc4u @afgan___ titip salam Min jangan ...	iyaiyao_	@somethinc4u @afgan___ titip salam min jangan ...	titip salam min jangan lupa pakai selalu somet...	[titip, salam, min, jangan, lupa, pakai, selal...
4	@Somethinc4u Happy Birthday Somethinc ! Semoga...	iyaiyao_	@somethinc4u happy birthday somethinc ! semoga...	happy birthday somethinc semoga bisa ters men...	[happy, birthday, somethinc, semoga, bisa, ter...
5	@Somethinc4u Wihhhh selamat ulang tahun @Somet...	seperlunyaajaa	@somethinc4u wihhhh selamat ulang tahun @somet...	wihhhh selamat ulang tahun semoga makin sukse...	[wihhhh, selamat, ulang, tahun, semoga, makin,...
6	@Somethinc4u Selamat ulang tahun somethinc sem...	onyounjm	@somethinc4u selamat ulang tahun somethinc sem...	selamat ulang tahun somethinc semoga produknya...	[selamat, ulang, tahun, somethinc, semoga, pro...
7	WTS / Want To Sell / Jual NEW bukan preloved...	jjkxzz	wts / want to sell / jual new bukan preloved...	wts want to sell jual new bukan preloved mas...	[wts, want, to, sell, jual, new, bukan, prelov...
8	@Somethinc4u Birthday wish buat Somethinc semo...	sandikala_	@somethinc4u birthday wish buat somethinc semo...	birthday wish buat somethinc semoga bisa terus...	[birthday, wish, buat, somethinc, semoga, bisa...
9	@Somethinc4u Happy birthday somethinc terima k...	khoirulaf_ifah	@somethinc4u happy birthday somethinc terima k...	happy birthday somethinc terima kasih somethin...	[happy, birthday, somethinc, terima, kasih, so...

Text Pre-Processing (4.1) Download Data Hasil Text Pre-Processing

```
somethinc2.to_csv("DataSomethincPreprocessing.csv")
```

Text Pre-Processing (5) Mencari & Menghitung Kata yang Paling Sering Muncul

```
from nltk import FreqDist

# Gabungkan semua kalimat dalam kolom 'cleaning'  menjadi satu string
kalimat = ''.join(somethinc2['cleaning'])

tokens=nltk.tokenize.word_tokenize(kalimat)
frek = nltk.FreqDist(tokens)
frek.most_common(20)
```

```
[('somethinc', 89),
 ('dan', 40),
 ('yg', 38),
 ('aku', 35),
 ('semoga', 32),
 ('bisa', 27),
 ('produk', 26),
 ('pake', 25),
 ('makin', 25),
 ('di', 22),
 ('yang', 22),
 ('semakin', 16),
 ('buat', 15),
 ('birthday', 13),
 ('bagus', 13),
 ('beli', 13),
 ('nya', 13),
 ('skincare', 12),
 ('ada', 12),
 ('jadi', 11)]
```

Hasil dari penggunaan fungsi `frek.most_common(20)` adalah program akan menampilkan 20 kata teratas dengan frekuensi paling banyak disebut pada unggahan Twitter bertopik produk Somethinc.

Kata "somethinc" muncul sebanyak 89 kali, menunjukkan bahwa ini adalah topik utama dalam pembahasan tersebut. Kata "dan" muncul 40 kali, diikuti oleh "yg" sebanyak 38 kali, serta "aku" sebanyak 35 kali. Kata "semoga" muncul 32 kali, menandakan harapan atau doa yang sering disampaikan. Kata "bisa" muncul 27 kali, dan "produk" muncul 26 kali, mengindikasikan bahwa ada pembicaraan mengenai kemampuan dan produk tertentu. Kata "pake" dan "makin" masing-masing muncul 25 kali, diikuti oleh "di" dan "yang" masing-masing muncul 22 kali. Kata "semakin" muncul 16 kali, "buat" 15 kali, serta "birthday" dan "bagus" masing-masing muncul 13 kali. Selain itu, kata "beli", "nya", dan "skincare" muncul 13 dan 12 kali, menandakan ada topik tentang pembelian produk skincare. Kata "ada" muncul 12 kali, dan "jadi" muncul 11 kali.