

Praktikum 1 Word Count (Big Data Analysis)

Nama: Ryan Fadhilah Faizal Hakim
NRP: 2C2220007
Subject: Word Counting Garuda Indonesia word datasets


Mengimport Library

```
import pandas as pd
import numpy as np
```

Pada cell di atas, library yang diperlukan yaitu pandas dan numpy

Melakukan Upload file

```
from google.colab import files
upload = 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 garudaindonesia4mav2024.csv to garudaindonesia4mav2024.csv

Pada cell di atas, dilakukan upload file garudaindonesia4may2024.csv dengan fungsi upload()

Membuat Dataframe dari file (.csv)

```
df = pd.read_csv('garudaindonesia4may2024.csv', delimiter=';')
df.head()
```

	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_i
0	1.786550e+18	Sat May 04 00:00:00 +0000 2024	0	Nikmati potongan hingga Rp500.000 setiap hari ...	1.786550e+18	https://pbs.twimg.com/media/GMsUluuaoAACOl8.jpg	
1	1.786360e+18	Fri May 03 11:42:45 +0000 2024	0	Kami Polsek Metro Penjarangan Bangga Dengan Ti...	1.786360e+18	https://pbs.twimg.com/media/GMptkasaMAAp-6w.jpg	
2	1.786330e+18	Fri May 03 10:00:01 +0000 2024	1	Jelajahi kenyamanan eksklusif bersama GarudaMi...	1.786330e+18	https://pbs.twimg.com/media/GMnzghKasAAQU7P.jpg	
3	1.786200e+18	Fri May 03 00:51:21 +0000 2024	1	Saatnya terbang lebih hemat di hari Jumat! Nik...	1.786200e+18	https://pbs.twimg.com/media/GMnYhC5bAAE_SxI.jpg	
4	1.786040e+18	Thu May 02 14:17:36 +0000 2024	3	SUPER BIG MATCH perebutan peringkat ketiga Pia...	NaN		NaN

Pada cell di atas, dataframe dibuat dengan menggunakan fungsi pd.read_csv() dan menyimpannya ke variabel df. Setelah itu outputnya ditampilkan dengan fungsi head() untuk menampilkan 5 baris dataframe

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 67 entries, 0 to 66
Data columns (total 15 columns):
#   Column                Non-Null Count  Dtype
---  -
0   conversation_id_str    67 non-null    float64
1   created_at             67 non-null    object
2   favorite_count         67 non-null    int64
3   full_text              67 non-null    object
4   id_str                 60 non-null    float64
5   image_url              49 non-null    object
6   in_reply_to_screen_name 6 non-null     object
7   lang                   60 non-null    object
8   location               21 non-null    object
9   quote_count            60 non-null    float64
10  reply_count            60 non-null    float64
11  retweet_count           60 non-null    float64
12  tweet_url              60 non-null    object
13  user_id_str            60 non-null    float64
14  username                60 non-null    object
dtypes: float64(6), int64(1), object(8)
memory usage: 8.0+ KB
```

Pada cell di atas, ditambahkan fungsi `info()` pada variabel `df` untuk melihat banyaknya indeks data, kolom (nama kolom), jumlah hitungan value non-null pada setiap kolom, dan juga tipe data dari value yang terdapat dalam setiap kolom

▼ Select Kolom dari Dataframe

```
df2 = df[df.columns[[3, 14]]]
df2
```


	full_text	username
0	Nikmati potongan hingga Rp500.000 setiap hari ...	IndonesiaGaruda
1	Kami Polsek Metro Penjaringan Bangga Dengan Ti...	Polsekmetroopen2
2	Jelajahi kenyamanan eksklusif bersama GarudaMi...	IndonesiaGaruda
3	Saatnya terbang lebih hemat di hari Jumat! Nik...	IndonesiaGaruda
4	SUPER BIG MATCH perebutan peringkat ketiga Pia...	NaN
...
62	Nikmati Bonus 50% Miles dengan Tukar Poin BCA ...	IndonesiaGaruda
63	Hemat lebih banyak saat terbang dengan America...	IndonesiaGaruda
64	Kami tdk ada niat mau demo / apapun Ke jakarta...	KerabatDarwis
65	Nikmati potongan tambahan hingga Rp550.000 den...	IndonesiaGaruda
66	Target yang sebenarnya Shin Tae-Yong targetkan...	r66sports

67 rows x 2 columns

Pada cell di atas, membuat variabel baru bernama `df2` dengan nilai berisi kolom dari indeks 3 dan indeks 14 dikarenakan hanya mengambil kolom bernama `full_text` dan kolom `username` serta menampilkan variabel `df2`

▼ Merubah Nilai Kolom `full_text` Menjadi Lowercase

```
df2['case_folded_full_text'] = df2['full_text'].str.lower()
df2.head()
```

 <ipython-input-13-067d47264602>: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

df2['case_folded_full_text'] = df2['full_text'].str.lower()

	full_text	username	case_folded_full_text
0	Nikmati potongan hingga Rp500.000 setiap hari ...	IndonesiaGaruda	nikmati potongan hingga rp500.000 setiap hari ...
1	Kami Polsek Metro Penjaringan Bangga Dengan Ti...	Polsekmetropen2	kami polsek metro penjaringan bangga dengan ti...
2	Jelajahi kenyamanan eksklusif bersama GarudaMi...	IndonesiaGaruda	jelajahi kenyamanan eksklusif bersama garudami...
3	Saatnya terbang lebih hemat di hari Jumat! Nik...	IndonesiaGaruda	saatnya terbang lebih hemat di hari jumat! nik...
4	SUPER BIG MATCH perebutan peringkat ketiga Pia...	NaN	super big match perebutan peringkat ketiga pia...

Pada cell di atas, ditambahkan kolom baru dengan nama `case_folded_full_text` yang memiliki nilai string lowercase dari kolom `full_text` menggunakan fungsi `str.lower()` serta menampilkan 5 baris dataframe

▼ Melakukan Text Cleaning

```
import re
import string

def textCleaning(text):
    # Remove mentions, hashtags, RT (Retweet), and links
    text = re.sub(r'@[A-Za-z0-9]+|#[A-Za-z0-9]+|RT[\s]+|http\S+', '', text)

    # Remove numbers
    text = re.sub(r'[0-9]+', '', text)


    # Remove all non-alphabetic characters
    text = re.sub(r'^A-Za-z ]+', '', text)

    # Remove new line characters and punctuations
    text = text.replace('\n', ' ').translate(str.maketrans('', '', string.punctuation))

    # Remove leading and trailing spaces
    text = text.strip()

    return text

df2['clean_text'] = df2['case_folded_full_text'].apply(textCleaning)
df2.head()
```

 <ipython-input-14-6a86bec13170>:22: 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

df2['clean_text'] = df2['case_folded_full_text'].apply(textCleaning)

	full_text	username	case_folded_full_text	clean_text
0	Nikmati potongan hingga Rp500.000 setiap hari ...	IndonesiaGaruda	nikmati potongan hingga rp500.000 setiap hari ...	nikmati potongan hingga rp setiap hari sabtu d...
1	Kami Polsek Metro Penjaringan Bangga Dengan Ti...	Polsekmetropen2	kami polsek metro penjaringan bangga dengan ti...	kami polsek metro penjaringan bangga dengan ti...
2	Jelajahi kenyamanan eksklusif bersama GarudaMi...	IndonesiaGaruda	jelajahi kenyamanan eksklusif bersama garudami...	jelajahi kenyamanan eksklusif bersama garudami...

Pada cell di atas, dilakukan pemanggilan library `re` dan `string` serta membuat suatu fungsi (User define function) bernama `textCleaning` yang berisi metode seperti menghilangkan mentions, hashtags, retweet, link, angka, simbol (karakter non alphabet), garis baru, dan space depan dan belakang.

▼ Mengimpor library `nltk`

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

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

Pada cell di atas, dilakukan import library `nltk` serta downloading nltk data bernama `punkt` untuk melakukan proses tokenisasi

▼ Melakukan Tokenization (Tokenisasi)

```
from nltk.tokenize import word_tokenize

df2['tokenize_word'] = df2['clean_text'].apply(word_tokenize)
df2
```

<ipython-input-22-28de4b81e48c>:3: 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
df2['tokenize_word'] = df2['clean_text'].apply(word_tokenize)

	full_text	username	case_folded_full_text	clean_text	tokenize_word
0	Nikmati potongan hingga Rp500.000 setiap hari ...	IndonesiaGaruda	nikmati potongan hingga rp500.000 setiap hari ...	nikmati potongan hingga rp setiap hari sabtu d...	[nikmati, potongan, hingga, rp, setiap, hari, ...
1	Kami Polsek Metro Penjaringan Bangga Dengan Ti...	Polsekmetropen2	kami polsek metro penjaringan bangga dengan ti...	kami polsek metro penjaringan bangga dengan ti...	[kami, polsek, metro, penjaringan, bangga, den...
2	Jelajahi kenyamanan eksklusif bersama GarudaMi...	IndonesiaGaruda	jelajahi kenyamanan eksklusif bersama garudami...	jelajahi kenyamanan eksklusif bersama garudami...	[jelajahi, kenyamanan, eksklusif, bersama, gar...
3	Saatnya terbang lebih hemat di hari Jumat! Nik...	IndonesiaGaruda	saatnya terbang lebih hemat di hari jumat! nik...	saatnya terbang lebih hemat di hari jumat nikm...	[saatnya, terbang, lebih, hemat, di, hari, jum...
4	SUPER BIG MATCH perebutan peringkat ketiga Pia...	NaN	super big match perebutan peringkat ketiga pia...	super big match perebutan peringkat ketiga pia...	[super, big, match, perebutan, peringkat, keti...
...
62	Nikmati Bonus 50% Miles dengan Tukar Poin BCA ...	IndonesiaGaruda	nikmati bonus 50% miles dengan tukar poin bca ...	nikmati bonus miles dengan tukar poin bca ke ...	[nikmati, bonus, miles, dengan, tukar, poin, b...
63	Hemat lebih banyak saat terbang dengan America...	IndonesiaGaruda	hemat lebih banyak saat terbang dengan america...	hemat lebih banyak saat terbang dengan america...	[hemat, lebih, banyak, saat, terbang, dengan, ...
64	Kami tdk ada niat mau demo / apapun Ke jakarta...	KerabatDarwis	kami tdk ada niat mau demo / apapun ke jakarta...	kami tdk ada niat mau demo apapun ke jakarta ...	[kami, tdk, ada, niat, mau, demo, apapun, ke, ...
65	Nikmati potongan tambahan hingga Rp550.000 den...	IndonesiaGaruda	nikmati potongan tambahan hingga rp550.000 den...	nikmati potongan tambahan hingga rp dengan mel...	[nikmati, potongan, tambahan, hingga, rp, deng...
66	Target yang sebenarnya	target yang sebenarnya shin tae-	target yang sebenarnya shin	[target, yang, sebenarnya, ...

Pada cell di atas, mengambil function `word_tokenize()` dari library `nltk` serta membuat kolom baru pada `df2` dengan nama `tokenize_word` yang memiliki nilai tokenisasi dari nilai kolom `clean_text` dengan menggunakan function `word_tokenize`

▼ Melakukan Word Count

```
tokens = [token for sublist in df2['tokenize_word'] for token in sublist]
wordFreq = nltk.probability.FreqDist(tokens)
wordFreq.most_common(20)
```

```
[('indonesia', 41),
 ('dan', 41),
 ('hingga', 37),
 ('garuda', 35),
 ('info', 33),
 ('dengan', 32),
 ('lengkap', 29),
 ('rp', 27),
 ('aplikasi', 22),
 ('flygaruda', 22),
 ('tiket', 22),
 ('promo', 20),
 ('gunakan', 19),
```

```
('saat', 18),  
( 'untuk', 18),  
( 'kode', 17),  
( 'melalui', 17),  
( 'anda', 16),  
( 'nikmati', 15),  
( 'potongan', 15)]
```

Pada cell di atas, dibuat list comprehension untuk merubah nilai pada kolom `tokenize_word` menjadi satu list yang disimpan ke dalam variabel `tokens`. Kemudian memanggil method `FreqDist()` dari library `nltk` pada variabel `tokens` untuk menghitung banyaknya kata dalam `tokens`. Lalu memanggil method `most_common()` dengan parameter `20` untuk menampilkan 20 word count teratas.

▼ Analisis Hasil

Berdasarkan hasil word count data `garudaindonesia4may2024.csv` didapatkan hasil bahwa kata yang paling banyak muncul adalah kata `indonesia` yang mencapai `41`, kata `dan` sebanyak `41`, kata `hingga` sebanyak `37`, kata `garuda` sebanyak `35`, dan seterusnya.