

STUDI INDEPENDEN BERSERTIFIKAT
E-COMMERCE

Informasi Tugas dan Kertas Tugas Capstone Project
TM021-IMR-TI dan Data-Tugas02

Bagian A. Informasi Tugas

1. Untuk Kelompok Tugas
 - a. Isi data di kotak berwarna hijau
 - b. Setelah tugas dikerjakan, unggah dokumen secara utuh (Bagian A, B, C, D jangan dihapus) ke:
 - i. Draft Kertas Tugas ke GDrive folder "03 Kertas Tugas-draft"
 - ii. Kertas Tugas Final ke GDrive folder "04 Kertas Tugas-final"
 - c. Cara penamaan Kertas Tugas: No Urut Tugas – Draft/Final – Kode Kelompok – Kode Sub Kelompok (bila ini tugas bersama isi dengan "Z")-Nama Kelompok. Contoh:
 - i. Tugas01-Draft-BT02-01-A-Beauty01-Amethyst
 - ii. Tugas02-Final-BT02-05-Z-FnB01
2. Data Tugas

ID Tugas	TM021	Nama Mentor Pembimbing Tugas	Muhammad Imran
No Urut Tugas	Tugas02	Tipe Kelompok	Tipe B
Hari dan Tanggal Tugas	Sabtu. 20 November 2021	Bidang / Profesi	IT dan Data
Hari dan Tanggal Kertas Tugas diserahkan ke Kelompok Tugas	Sabtu. 20 November 2021	Catatan tambahan	Rubrik penilaian masih dalam proses penyusunan oleh mentor pembimbing dan mentor penilai
Nama Mentor Penilai Tugas (MNIT)	Valentinus Paramartha	Tanggal penyerahan draft Kertas Tugas dan PPT ke MNIT	(diisi oleh Kelompok Tugas, sesuai data di GSheet Sentra Tugas)

Target tanggal penyerahan draft	Sabtu, 27 November 2021	Tanggal Mentoring Tugas Sinkronus	(diisi oleh Kelompok Tugas, sesuai data di GSheet Sentra Tugas)
Tanggal penilaian Kertas Tugas dan PPT	(diisi oleh MNIT, diisi sesuai data di GSheet Sentra Tugas)	Tanggal penyerahan Kertas Tugas dan PPT final ke MNIT	(diisi oleh Kelompok Tugas, sesuai data di GSheet Sentra Tugas)
Data Kelompok Tugas			
Kode Kelompok Tugas dan Nama Kelompok Tugas	(diisi oleh Kelompok Tugas, sesuai data di GSheet Sentra Tugas)	Nama Sub-Kelompok Tugas	(diisi oleh Kelompok Tugas, sesuai data di GSheet Sentra Tugas)
Nama Anggota Sub-Kelompok yang berkontribusi	(diisi oleh Kelompok Tugas, sesuai data di GSheet Sentra Tugas) 1. Arif Widagdo	Nama Anggota Kelompok yang berkontribusi dalam pengerjaan tugas / review tugas	(diisi oleh Kelompok Tugas) 1.
	2. Eko Prasetyo	Nama Anggota Kelompok yang berkontribusi dalam pengerjaan tugas / review tugas	2.
	3. Harnum Gina Fortuna		3.
	4. Putra Surya Jaya Togatorp		4.
			5.
			6.
			7.
			8.
			9.
			10.

3. Uraian Tugas

Automatisasi data pipelines dengan menggunakan python atau proses data ingestion menggunakan python

4. Rubrik Tugas dan Nilai

Kriteria Penilaian: Pemahaman mengenai fondasi Data Pipelines & ETL, paham data engineering

Rubrik Tugas TM021-IMR-TI dan Data-Tugas02			
No Rubrik	1	2	3

Nama Rubrik	Kesesuaian output dengan metode penarikan data	Proses ETL	Visualisasi data
100%	30%	30%	40%
(diisi oleh MNIT nilai final)	(diisi oleh MNIT nilai final)	(diisi oleh MNIT nilai final)	(diisi oleh MNIT nilai final)
75-100	Dua metodenya memiliki output yang sesuai	Men-drop salah satu kolom dan mengintegrasikan beberapa tabel	Visualisasi data lengkap (real time dan batch processing)
50-<75	Salah satu metodenya memiliki metode yang sesuai	Men-drop salah satu kolom namun tidak mengintegrasikan tabel (atau salah satunya)	Visualisasi data yang ditampilkan hanya salah satu (antara real time atau batch processing)
25-<50	Souce codenya ada namun gagal tereksekusi	Souce codenya ada namun gagal mengintegrasikan dan men-drop tabel	Souce codenya ada namun gagal dalam memvisualisasikan data
<25	Tidak ada souce codenya	Tidak ada souce codenya	Tidak ada souce codenya

Bagian B. Kertas Tugas

Tugas ini adalah Data pipeline dari hashtag business, tempat kami mengambil data yaitu dari Twitter dengan menggunakan API Twitter yang disediakan. Data yang telah terintegrasi selanjutnya divisualisasikan menggunakan Power BI.

1). Metode yang digunakan

- **ETL** (Extract, Transform, Load)

Pertama dilakukan pemanggilan library dengan coding sebagai berikut:

```
import tweepy
import csv
import pandas as pd

from textblob import TextBlob
from nltk.corpus import stopwords
import re
```

Setelah library telah di import maka dilakukan definisi variabel untuk melakukan proses penarikan, variable tersebut merupakan authenticatin, dengan menggunakan coding berikut untuk menjalankan authentikaasi yang memanggil consumer_key, consumer_secret, access_token, dan access_secret

```
authentication = tweepy.OAuthHandler(consumer_key, consumer_secret)
authentication.set_access_token(access_token, access_secret)
api = tweepy.API(authentication, wait_on_rate_limit=True)
```

```
maxId = -1
tweetCount = 0
```

Sebelum proses penarikan di lakukan kami membuat function dimana dalam function tersebut kami akan membersihkan data text yang di ambil, proses tersebut menggunakan coding berikut:

```
#clean tweet text
def clean_text(text):
    ex_list = ['rt', 'http', 'RT']
    exc = '|'.join(ex_list)
    text = re.sub(exc, ' ', text)
    text = text.lower()
    words = text.split()
    stopword_list = stopwords.words('english')
    words = [word for word in words if not word in stopword_list]
    clean_text = ' '.join(words)
    return clean_text

def sentiment_score(text):
    analysis = TextBlob(text)
    if analysis.sentiment.polarity > 0:
        return 1
    elif analysis.sentiment.polarity == 0:
        return 0
    else:
        return -1
```

- Selain itu kami juga membuat array atau wadah yang nantinya akan menampung data yang kami tarik untuk di **Transform** kedalam CSV, berikut ada code dari array yang akan menampung baris dan kolom yang akan kami panggil, pada array I merupakan wadah yang menampung untuk id_tweet, ca untuk menampung create_at, tt untuk menampung text_tweet, rc untuk menampung retweet_count, fc untuk menampung favorite_count, dan rs untuk menampung result_score. Selain itu ada lagi dimana kita menampung Username, dan media source, berikut adalah kode yang digunakan untuk menampung data yang akan di panggil.

```
csvFile = open("tweet.csv", "a+", newline="", encoding="utf-8")
csvWriter = csv.writer(csvFile)
i = []
ca = []
tt = []
rc = []
fc = []
rs = []
```

Setelah semua nya sudah dilakukan, maka proses **Extratcing Data** akan dilakukan menggunakan sintax berikut :

```
while tweetCount < maxTweets:
    if(maxId <= 0):
        newTweets = api.search_tweets(q=hashtag, count=tweetsPerQry, lang="id",
result_type="recent", tweet_mode="extended")
    else:
        newTweets = api.search_tweets(q=hashtag, count=tweetsPerQry, lang="id",
max_id=str(maxId - 1),result_type="recent", tweet_mode="extended")

    if not newTweets:
        print("Done")
        break

    for tweet in newTweets:
        id = tweet.id
        created_at = str(tweet.created_at)
        tweet_text = tweet.full_text
        tweet_text_sent = tweet.full_text
        retweet_count = tweet.retweet_count
        fav_count = tweet.favorite_count
        tweet_text_sent = clean_text(tweet_text_sent)
        result_score = sentiment_score(tweet_text_sent)

        print(tweet.id, str(tweet.created_at), clean_text(tweet_text_sent),
tweet.retweet_count, tweet.favorite_count, sentiment_score(tweet_text_sent))
        i.append(id)
        ca.append(created_at)
        tt.append(tweet_text)
        rc.append(retweet_count)
        fc.append(fav_count)
        rs.append(result_score)

        tweets=[tweet.id, str(tweet.created_at),
clean_text(tweet_text_sent),tweet.retweet_count, tweet.favorite_count,
sentiment_score(tweet_text_sent)]
        csvWriter.writerow(tweets)
```

Parameter lang = "id" kami gunakan untuk memanggil data dengan keyword bahasa indonesia.

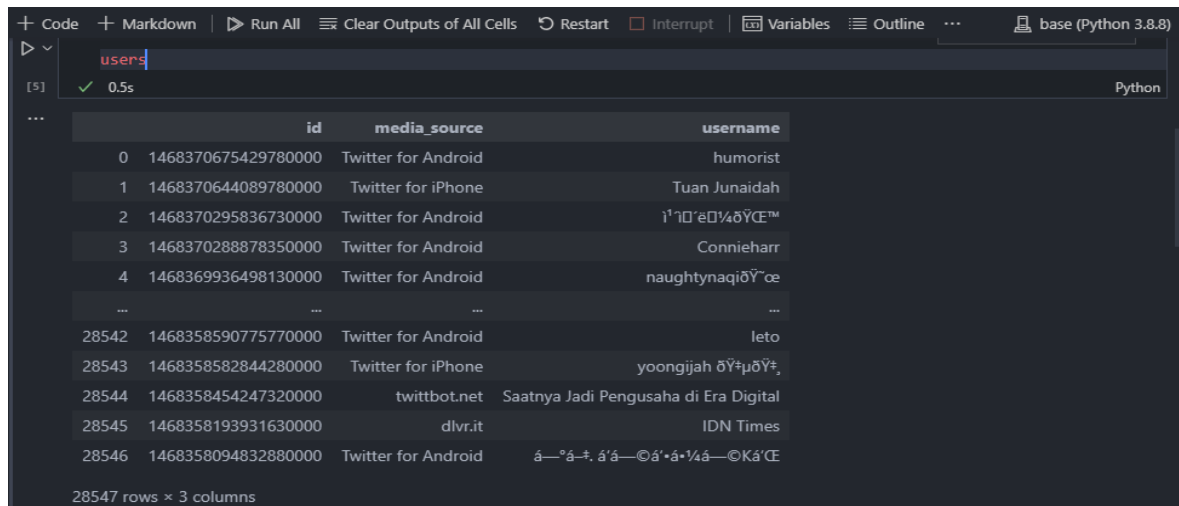
Proses selanjutnya data akan di **Load** kedalam CSV, yang nantinya data akan diolah kembali untuk proses integrasi dan cleaning data yang telah di panggil.

2). Integrasi Tabel

Setelah melalui proses ETL dan data pipeline, kelompok kami menjadikan data tabel yang akan diintegrasikan berdasarkan data yang akan ditampilkan pada visualisasi.

Tabel yang kami gunakan yaitu Tabel Users dan Tabel Tweet, berikut adalah data frame yang kami panggil :

DataFrame Users :

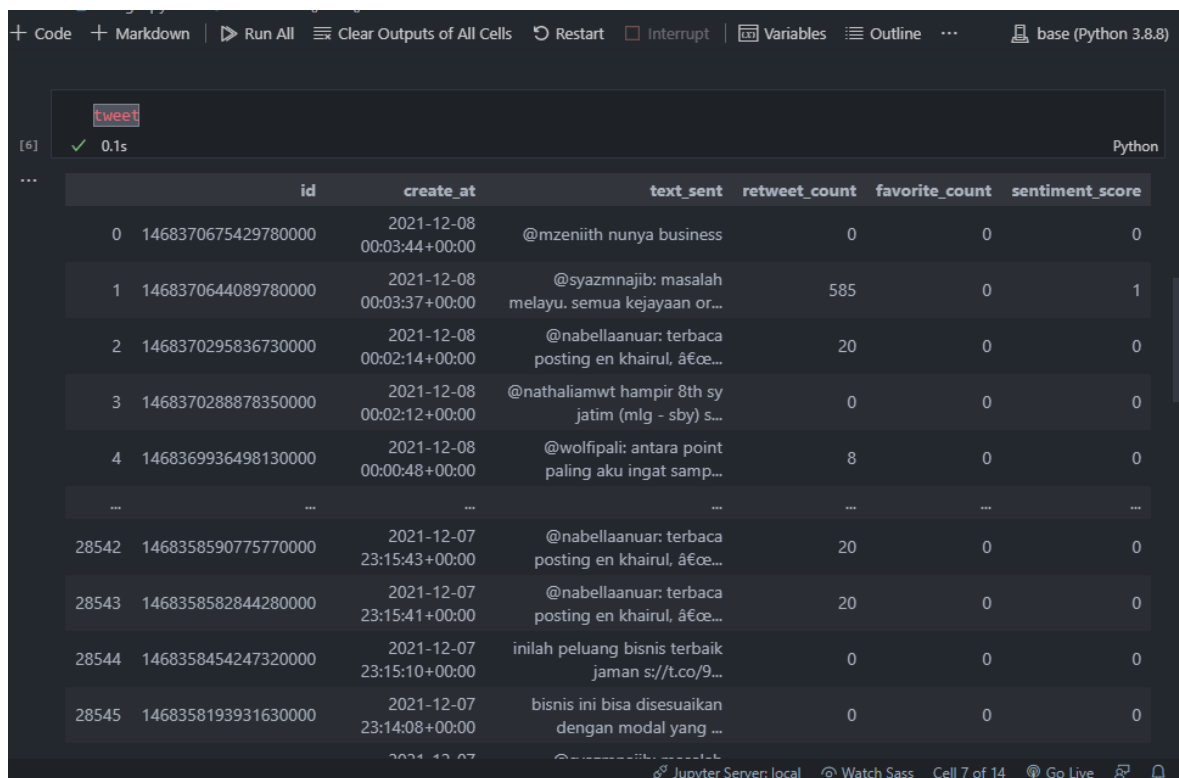


	id	media_source	username
0	1468370675429780000	Twitter for Android	humorist
1	1468370644089780000	Twitter for iPhone	Tuan Junaidah
2	1468370295836730000	Twitter for Android	1'10'e0'46YCE™
3	1468370288878350000	Twitter for Android	Connieharr
4	1468369936498130000	Twitter for Android	naughtynaqi0Y"œ
...
28542	1468358590775770000	Twitter for Android	leto
28543	1468358582844280000	Twitter for iPhone	yoongijah 0Y*µ0Y*,
28544	1468358454247320000	twittbot.net	Saatnya Jadi Pengusaha di Era Digital
28545	1468358193931630000	dlvr.it	IDN Times
28546	1468358094832880000	Twitter for Android	á—*á-†, á'á—©á'á•¼á—©Ká'CE

28547 rows × 3 columns

Gambar 1, Dataframe Users

DataFrame Tweet :

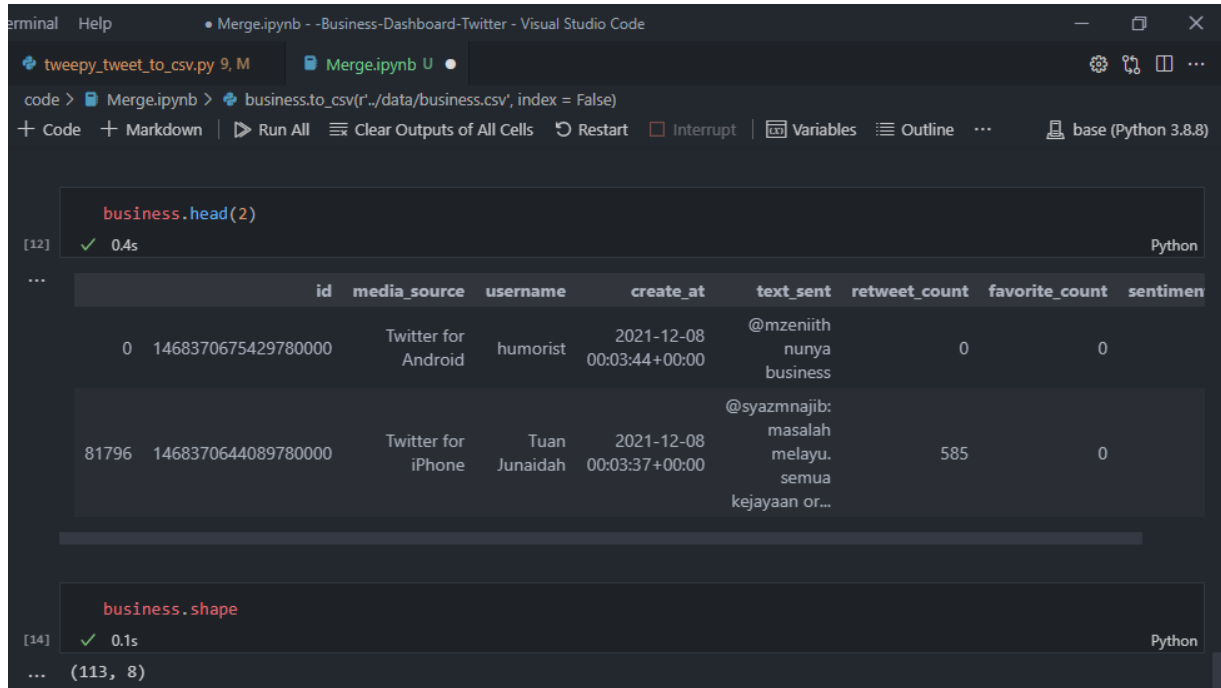


	id	create_at	text_sent	retweet_count	favorite_count	sentiment_score
0	1468370675429780000	2021-12-08 00:03:44+00:00	@mzeniith nunya business	0	0	0
1	1468370644089780000	2021-12-08 00:03:37+00:00	@syazmnajib: masalah melayu. semua kejayaan or...	585	0	1
2	1468370295836730000	2021-12-08 00:02:14+00:00	@nabellaanuar: terbaca posting en khairul, áœœ...	20	0	0
3	1468370288878350000	2021-12-08 00:02:12+00:00	@nathaliawmt hampir 8th sy jatim (mlg - sby) s...	0	0	0
4	1468369936498130000	2021-12-08 00:00:48+00:00	@wolfipali: antara point paling aku ingat samp...	8	0	0
...
28542	1468358590775770000	2021-12-07 23:15:43+00:00	@nabellaanuar: terbaca posting en khairul, áœœ...	20	0	0
28543	1468358582844280000	2021-12-07 23:15:41+00:00	@nabellaanuar: terbaca posting en khairul, áœœ...	20	0	0
28544	1468358454247320000	2021-12-07 23:15:10+00:00	inilah peluang bisnis terbaik jaman s//t.co/9...	0	0	0
28545	1468358193931630000	2021-12-07 23:14:08+00:00	bisnis ini bisa disesuaikan dengan modal yang ...	0	0	0

Jupyter Server: local Watch Sass Cell 7 of 14 Go Live

Gambar 2, Dataframe Tweet

Dari kedua Dataframe diatas memiliki data dengan banyak baris yang sama, yaitu sebanyak 28547 baris. Setelah itu maka kami lakukan proses penggabungan dataframe dengan menggunakan join keys id_tweet, selain itu kami mendrop kolom yang mempunyai banyak nilai null, yaitu kolom favorite_count dan juga baris yang redundan atau duplikasi kami drop, sehingga dari penggabungan kedua dataframe tersebut kami menghasilkan dataframe baru yaitu business, dan dataframe business yang kami peroleh benar benar bersih dari duplikasi data, berikut adalah dataframe business



```
business.head(2)
```

	id	media_source	username	create_at	text_sent	retweet_count	favorite_count	sentimen
0	1468370675429780000	Twitter for Android	humorist	2021-12-08 00:03:44+00:00	@mzeniith nunya business	0	0	
81796	1468370644089780000	Twitter for iPhone	Tuan Junaidah	2021-12-08 00:03:37+00:00	@syazmnajib: masalah melayu. semua kejayaan or...	585	0	

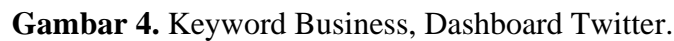
```
business.shape
```

```
(113, 8)
```

Gambar 3, Dataframe Business

Dari banyaknya baris yang di drop maka, kami ber asumsi banyaknya data yang redundan pada saat penarikan data dilakukan, sehingga data yang benar benar bersih kami peroleh sebanyak 113 baris.

Pada proses visualisasi data kelompok kami menggunakan power BI. Kelompok kami memiliki beberapa pilihan tampilan seperti, berdasarkan device yang digunakan, sentimenttext sent dan data secara umum. Berikut hasil power BI kami:



Assets File : <https://github.com/Arif-Widagdo/-Business-Dashboard-Twitter.git>

