DOKUMENTASI TUGAS AKHIR DATA MINING A11.4408

Nama : Alfebrian Ivo Kurnia Adi
NIM : A11.2021.13920

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

```
df = pd.read_csv('Tweets.csv')
df.head()
```

Output :

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_location	user_timezone
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	cairdin	NaN	0	@VirginAmerica What @dhepburn said.	NaN	2015-02-24 11:35:52 -0800	NaN	Eastern Time (US & Canada)
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica plus you've added commercials t...	NaN	2015-02-24 11:15:59 -0800	NaN	Pacific Time (US & Canada)
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yvonnalynn	NaN	0	@VirginAmerica I didn't today... Must mean I n...	NaN	2015-02-24 11:15:48 -0800	Lets Play	Central Time (US & Canada)
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica it's really aggressive to blast...	NaN	2015-02-24 11:15:36 -0800	NaN	Pacific Time (US & Canada)
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica and it's a really big bad thing...	NaN	2015-02-24 11:14:45 -0800	NaN	Pacific Time (US & Canada)

```
import sqlite3
conn = sqlite3.connect('database.sqlite')
```

```
cur = conn.cursor()
cur.execute("Select * From Tweets Limit 5")
```

Output :

```
<sqlite3.Cursor at 0x7b276869c2c0>
```

```
rows = cur.fetchall()
for row in rows:
    print(row)
```

Output :

```
df.describe()
```

Output :

	tweet_id	airline_sentiment_confidence	negativereason_confidence	retweet_count	
count	1.464000e+04	14640.000000	10522.000000	14640.000000	
mean	5.692184e+17	0.900169	0.638298	0.082650	
std	7.791112e+14	0.162830	0.330440	0.745778	
min	5.675883e+17	0.335000	0.000000	0.000000	
25%	5.685592e+17	0.692300	0.360600	0.000000	
50%	5.694779e+17	1.000000	0.670600	0.000000	
75%	5.698905e+17	1.000000	1.000000	0.000000	
max	5.703106e+17	1.000000	1.000000	44.000000	

```
df.info()
```

Output :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   tweet_id                      14640 non-null  int64  
 1   airline_sentiment             14640 non-null  object 
 2   airline_sentiment_confidence  14640 non-null  float64
 3   negativereason                9178 non-null   object 
 4   negativereason_confidence     10522 non-null  float64
 5   airline                       14640 non-null  object 
 6   airline_sentiment_gold        40 non-null     object 
 7   name                          14640 non-null  object 
 8   negativereason_gold           32 non-null     object 
 9   retweet_count                 14640 non-null  int64  
 10  text                          14640 non-null  object 
 11  tweet_coord                   1019 non-null   object 
 12  tweet_created                 14640 non-null  object 
 13  tweet_location                9907 non-null   object 
 14  user_timezone                 9820 non-null   object 
dtypes: float64(2), int64(2), object(11)
memory usage: 1.7+ MB
```

```
df.select_dtypes('object').columns.tolist()
```

Output :

```
['airline_sentiment',
 'negativereason',
 'airline',
 'airline_sentiment_gold',
 'name',
 'negativereason_gold',
 'text',
 'tweet_coord',
 'tweet_created',
 'tweet_location',
 'user_timezone']
```

```
df['airline_sentiment'].unique()
```

Output :

```
array(['neutral', 'positive', 'negative'], dtype=object)
```

```
df['airline_sentiment'].value_counts()
```

Output :

```
negative 9178
neutral 3099
positive 2363
Name: airline_sentiment, dtype: int64
```

```
def execute_sql(command):
#     Function that execute command from the database and print the
results
    cur = conn.cursor()
    cur.execute(command)
    rows = cur.fetchall()
    for row in rows:
        print(row)
```

```
execute_sql("Select DISTINCT(airline_sentiment) From Tweets")
('neutral',)
('negative',)
('positive',)
```

Output :

```
('neutral',)
('negative',)
('positive',)
('positive',)
```

```
execute_sql("Select airline_sentiment, COUNT(airline_sentiment) From
Tweets GROUP BY airline_sentiment")
('negative', 9082)
('neutral', 3069)
('positive', 2334)
```

Output :

```
('negative', 9082)
('neutral', 3069)
('positive', 2334)
('positive', 2334)
```

```
pd.read_sql("Select DISTINCT(airline_sentiment) From Tweets", conn)
```

Output :

	airline_sentiment
0	neutral
1	negative
2	positive

```
df_neg = df.loc[(df.airline_sentiment == 'negative') &
(df.negativereason != "Can't Tell")].reset_index(drop=True)
df_neg.shape
```

Output :

```
(7988, 15)
```

```
df_neg_sql = pd.read_sql('''Select * From Tweets WHERE
airline_sentiment = 'negative' \
                        AND negativereason != "Can't Tell"''', conn)
df_neg_sql.shape
```

Output :

```
(7906, 15)
```

```
df_neg.groupby('negativereason').negativereason_confidence.mean()
```

Output :

```
negativereason
Bad Flight 0.631731
Cancelled Flight 0.783096
Customer Service Issue 0.780054
Damaged Luggage 0.733432
Flight Attendant Complaints 0.659639
Flight Booking Problems 0.606797
Late Flight 0.768907
Lost Luggage 0.813019
longlines 0.594076
Name: negativereason_confidence, dtype: float64
```

```
pd.read_sql('''Select negativereason, AVG(negativereason_confidence) AS
average_confidence From Tweets \
            WHERE airline_sentiment = 'negative' \
            AND negativereason != "Can't Tell" GROUP BY
negativereason''', conn)
```

Output :

	negativereason	average_confidence
0	Bad Flight	0.630785
1	Cancelled Flight	0.783200
2	Customer Service Issue	0.779946
3	Damaged Luggage	0.734204
4	Flight Attendant Complaints	0.658255
5	Flight Booking Problems	0.607153
6	Late Flight	0.768978
7	Lost Luggage	0.812209

negativereason	average_confidence	
8	longlines	0.593856