In [2]:
```
import pandas as pd
#get data
json_data=pd.read_json("dm-2024-isa-5810-lab-2-homework/tweets_DM.json",lines=True)
ident_data=pd.read_csv("dm-2024-isa-5810-lab-2-homework/data_identification.csv")
emotion_data=pd.read_csv("dm-2024-isa-5810-lab-2-homework/emotion.csv")
```

In [3]:
```
#only use the "tweet" object in data
hashtags=[]
tweet_id=[]
text=[]
for i in range(len(json_data)):
    hashtags.append(json_data["_source"][i]["tweet"]["hashtags"])
    tweet_id.append(json_data["_source"][i]["tweet"]["tweet_id"])
    text.append(json_data["_source"][i]["tweet"]["text"])
data={"tweet_id":tweet_id,
      "hashtags":hashtags,
      "text":text}
df = pd.DataFrame(data)
```

In [4]:
```
#merge data , add "identification" column , separate to train and test dataframe
new = pd.merge(df,ident_data,on='tweet_id')
test=new[new["identification"]=="test"]
new = pd.merge(new,emotion_data,on='tweet_id')
train=new[new["identification"]=="train"]
```

In [ ]:
```
#combine column "hashtags" and "text" to one column "texts"
train["hashtags"]=train["hashtags"].apply(lambda x:','.join(x))
train["texts"]=train["hashtags"]+","+train["text"]
train
```

Out[ ]:

	tweet_id	hashtags	text	identification	emoti
0	0x376b20	S,n,a,p,c,h,a,t	People who post "add me on #Snapchat" must be ...	train	anticipati
1	0x2d5350	f,r,e,e,p,r,e,s,s,,,T,r,u,m,p,L,e,g,a,c,y,,,C,N,N	@brianklaas As we see, Trump is dangerous to #...	train	sadne
2	0x1cd5b0		Now ISSA is stalking Tasha 😂😂😂 <LH>	train	fe
3	0x1d755c	a,u,t,h,e,n,t,i,c,,,L,a,u,g,h,O,u,t,L,o,u,d	@RISKshow @TheKevinAllison Thx for the BEST TI...	train	j
4	0x2c91a8		Still waiting on those supplies Liscus. <LH>	train	anticipati
...	...	...	...	...	
1455558	0x321566	N,o,W,o,n,d,e,r,,,H,a,p,p,y	I'm SO HAPPY!!! #NoWonder the name of this sho...	train	j
1455559	0x38959e		In every circumtance I'd like to be thankful t...	train	j
1455560	0x2cbca6	b,l,e,s,s,y,o,u	there's currently two girls walking around the...	train	j
1455561	0x24faed		Ah, corporate life, where you can date <LH> us...	train	j
1455562	0x34be8c	S,u,n,d,a,y,v,i,b,e,s	Blessed to be living #Sundayvibes <LH>	train	j

1455563 rows × 8 columns

In [ ]:

```
#same to train
test["hashtags"]=test["hashtags"].apply(lambda x:','.join(x))
test["texts"]=test["hashtags"]+","+test["text"]
test
```

Out[ ]:

	tweet_id	hashtags	text	identification	
2	0x28b412	bibleverse	Confident of your obedience, I write to you, k...	test	bibleverse,Con obe
4	0x2de201		"Trust is not the same as faith. A friend is s...	test	,"Trust is not the sa
9	0x218443	materialism,money,possessions	When do you have enough ? When are you satisfi...	test	materialism,money,poss
30	0x2939d5	GodsPlan,GodsWork	God woke you up, now chase the day #GodsPlan #...	test	GodsPlan,GodsWork,(up,
33	0x26289a		In these tough times, who do YOU turn to as yo...	test	,In these tough times
...	...	...	...	...	
1867525	0x2913b4		"For this is the message that ye heard from th...	test	,"For this is the me l
1867529	0x2a980e		"There is a lad here, which hath five barley l...	test	,"There is a lad here, w
1867530	0x316b80	mixedfeeling,butimTHATperson	When you buy the last 2 tickets remaining for ...	test	mixedfeeling,butimTHA
1867531	0x29d0cb		I swear all this hard work gone pay off one da...	test	,I swear all this hard w
1867532	0x2a6a4f		@Parcel2Go no card left when I wasn't in so I ...	test	,@Parcel2Go no ca w

411972 rows × 5 columns

In [ ]:
```
#Let emotion string to 0~7 number in column "emotion_new"
#tweet_id Hexadecimal to Decimal number
emotions=["anger", "anticipation", "disgust", "fear", "sadness", "surprise", "trust
for i in range(8):
    train["emotion_new"]=train["emotion"].replace(emotions[i], i)
train["tweet_id_new"]=train["tweet_id"].apply(lambda x: int(x, 16))
```

In [7]:
```
train["texts"]
```

Out[7]:
```
0          Snapchat,People who post "add me on #Snapchat"...
1          freepress,TrumpLegacy,CNN,@brianklaas As we se...
2                      ,Now ISSA is stalking Tasha 😂😂😂 <LH>
3          authentic,LaughOutLoud,@RISKshow @TheKevinAlli...
4              ,Still waiting on those supplies Liscus. <LH>
                                ...
1455558    NoWonder,Happy,I'm SO HAPPY!!! #NoWonder the n...
1455559    ,In every circumtance I'd like to be thankful ...
1455560    blessyou,there's currently two girls walking a...
1455561    ,Ah, corporate life, where you can date <LH> u...
1455562    Sundayvibes,Blessed to be living #Sundayvibes ...
Name: texts, Length: 1455563, dtype: object
```

In [ ]:
```
#build transform
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split

# Convert text to numerical features using TF-IDF
tfidf = TfidfVectorizer(max_features=1000)
x = tfidf.fit_transform(train["texts"]).toarray()
```

In [55]:
```
x.toarray().shape
```

Out[55]:
```
(1455563, 1000)
```

In [ ]:
```
# Split into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(x, train["emotion"].tolist(), t

# Train a Random Forest classifier
clf = RandomForestClassifier(n_estimators=100, random_state=42)
clf.fit(X_train, y_train)

# Predict
predicted = clf.predict(X_test)
clf.score(X_test,y_test)
```

Out[ ]:
```
0.5204611267789484
```

In [ ]:
```
#predict test data
tfidf = TfidfVectorizer(max_features=1000)
xt = tfidf.fit_transform(test["texts"]).toarray()

predicted = clf.predict(xt)
```

In [ ]:
```
#save data to csv
data={
    "id":test["tweet_id"].tolist(),
    "emotion":predicted,
```

```
}

df = pd.DataFrame(data)
df.to_csv('gfg2.csv',index= False)
```