

12/21/21, 3:28 PM

DA4 jpyrb - Colaboratory

```
import pandas as pd
import re

train=pd.read_csv("train.csv")

train.head()
```

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation

```
# drop col 'id' (as it is of no use) and replace it in the same variable
train.drop("id",inplace=True,axis=1)

train.head()
```

	label	tweet
0	0	@user when a father is dysfunctional and is s...
1	0	@user @user thanks for #lyft credit i can't us...
2	0	bihday your majesty
3	0	#model i love u take with u all the time in ...
4	0	factsguide: society now #motivation

```
temp = train.groupby("label").size()
temp
```

```
label
0    29720
1     2242
dtype: int64
```

```
import nltk
#nltk.download()
```

```
from nltk.stem import PorterStemmer
stemmer = PorterStemmer()

def clean_sentences(text):
    text = text.lower() # convert text to lower case
    text = re.sub(r"[^a-z0-9^.\!\,\"]", "", text) # remove special char's
    text = " ".join(text.split())
    text = " ".join(stemmer.stem(word) for word in text.split()) # do stemming
    return text

x = train['tweet']
y = train['label']

x = x.map(lambda a: clean_sentences(a))

x.head()
```

```
0    user when a father is dysfunct and is so self...
1    user user thank for lyft credit i can't use ca...
2                                bihday your majesti
3    model i love u take with u all the time in ur !!!
4                factsguid societ now motiv
Name: tweet, dtype: object
```

```
pip install sklearn
```

```
Requirement already satisfied: sklearn in /usr/local/lib/python3.7/dist-packages (0.0)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (from sklearn) (1.0.1)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-packages (from scikit-learn->sklearn) (1.1.0)
Requirement already satisfied: scipy>=1.1.0 in /usr/local/lib/python3.7/dist-packages (from scikit-learn->sklearn) (1.4.1)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from scikit-learn->sklearn) (3.0.0)
Requirement already satisfied: numpy>=1.14.6 in /usr/local/lib/python3.7/dist-packages (from scikit-learn->sklearn) (1.19.5)
```

```
from sklearn.model_selection import train_test_split

# split the dataset into training set & testing set
# data is split in a stratified fashion
x_train, x_test, y_train, y_test = train_test_split(x,y,stratify=y,random_state=42)

x_train.head()
```

```
1036    user like the spread of peanut butter on white...
2380    watch made in america o.j. simpson... 30for3...
31605   franci underwood seen leav marseil nojok
23437   get up get get enjoy music today free app free...
2669    my 1st juic experience! notsobad healthyliv ea...
Name: tweet, dtype: object
```

```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer(stop_words='english')

x_train = vectorizer.fit_transform(x_train)

x_test = vectorizer.transform(x_test)

from sklearn.svm import LinearSVC

model = LinearSVC(C=1.05, tol=0.5)

model.fit(x_train,y_train)

LinearSVC(C=1.05, tol=0.5)

from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, f1_score, recall_score
confusion_matrix(y_test,model.predict(x_test))
```

```
array([[77370,   60],
       [ 227,  334]])
```

```
accuracy_score(y_test,model.predict(x_test))

0.9640845951695658

recall_score(y_test,model.predict(x_test))

0.5953654188948306

precision_score(y_test,model.predict(x_test))

0.8477157360406091

f1_score(y_test,model.predict(x_test))

0.6994764397905757

# sample_text = ['I hate you']
# sample_text = ['I dont hate you']
sample_text = ['you are a bad person.']
sample_text = list(map(lambda a: clean_sentences(a), sample_text))
sample_text

['you are a bad person.']

sample_text = vectorizer.transform(sample_text)

model.predict(sample_text)[0]

0
```

✓ 2s

completed at 15:28

1/1