K214188-NLP-Assignment2

September 25, 2022

[102]:
```
#importing libraries
import numpy as np
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn import metrics
import nltk
import re
import string
from nltk.stem import WordNetLemmatizer
```

[103]:
```
#reading the data
train_csv = pd.read_csv('train-authors.csv')
test_csv = pd.read_csv('test-authors.csv')
```

[104]:
```
#stopword removal and lemmatization
stopwords = nltk.download('stopwords')
#stopwords = nltk.corpus.stopwords.words('english')
#lemmatizer = WordNetLemmatizer()
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Faiz\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

[105]:
```
stopwords = nltk.corpus.stopwords.words('english')
```

[106]:
```
lemmatizer = WordNetLemmatizer()
```

[107]:
```
train_csv.head()
```

[107]:

	text	author
0	She wanted clothes to keep her warm, and food...	dickens
1	The question now was, who was the man,\nand w...	doyle
2	I therefore\n smoked a great number of t...	doyle
3	I am partial to the modern\nFrench school. \n...	doyle
4	” She stood smiling, holding up a little slip ...	doyle

[108]:
```
train_X_non = train_csv['text']   # '0' refers to the review text
train_y = train_csv['author']   # '1' corresponds to Label (1 - positive and 0 -␣
↪negative)
test_X_non = test_csv['text']
test_y = test_csv['author']
train_X=[]
test_X=[]
```

[109]:
```
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\Faiz\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

[109]: True

[110]:
```
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\Faiz\AppData\Roaming\nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
```

[110]: True

[111]:
```
#text pre processing
for i in range(0, len(train_X_non)):
    review = re.sub('[^a-zA-Z]', ' ', train_X_non[i])
    review = review.lower()
    review = review.split()
    review = [lemmatizer.lemmatize(word) for word in review if not word in␣
↪set(stopwords)]
    review = ' '.join(review)
    train_X.append(review)
```

[112]:
```
#text pre processing
for i in range(0, len(test_X_non)):
    review = re.sub('[^a-zA-Z]', ' ', test_X_non[i])
    review = review.lower()
    review = review.split()
    review = [lemmatizer.lemmatize(word) for word in review if not word in␣
↪set(stopwords)]
    review = ' '.join(review)
    test_X.append(review)
```

[113]:
```
train_X[10]
```

[113]: 'much fairly clear sure right mr holmes cried client mr toller cellar said husband lie snoring kitchen rug shall soon see managed'

[114]:
```
#tf idf
tf_idf = TfidfVectorizer()
#applying tf idf to training data
X_train_tf = tf_idf.fit_transform(train_X)
#applying tf idf to training data
X_train_tf = tf_idf.transform(train_X)
```

[115]:
```
print("n_samples: %d, n_features: %d" % X_train_tf.shape)
```

```
n_samples: 30000, n_features: 29000
```

[116]:
```
#transforming test data into tf-idf matrix
X_test_tf = tf_idf.transform(test_X)
```

[117]:
```
print("n_samples: %d, n_features: %d" % X_test_tf.shape)
```

```
n_samples: 10000, n_features: 29000
```

[118]:
```
#naive bayes classifier
naive_bayes_classifier = MultinomialNB()
naive_bayes_classifier.fit(X_train_tf, train_y)
```

[118]:
```
MultinomialNB()
```

[119]:
```
#predicted y
y_pred = naive_bayes_classifier.predict(X_test_tf)
y_pred
```

[119]:
```
array(['dickens', 'defoe', 'doyle', ..., 'dickens', 'twain', 'dickens'],
      dtype='<U7')
```

[120]:
```
print(metrics.classification_report(test_y, y_pred, target_names=['dickens',␣
↪'doyle','defoe','twain']))
```

```
              precision    recall  f1-score   support

     dickens       0.85      0.97      0.91      2431
       doyle       0.92      0.85      0.88      2507
       defoe       0.90      0.91      0.91      2540
       twain       0.92      0.85      0.88      2522

    accuracy                           0.90     10000
   macro avg       0.90      0.90      0.90     10000
weighted avg       0.90      0.90      0.90     10000
```

[121]:
```
print("Confusion matrix:")
print(metrics.confusion_matrix(test_y, y_pred))
```

```
Confusion matrix:
[[2361   26   20   24]
 [ 134 2138  127  108]
 [  97   72 2315   56]
 [ 184   94  101 2143]]
```

[128]:
```
#test =
#test_csv['text']
#for i in range(len(test_csv)):
#    print(test_csv.loc[i, "text"])
#test
authors = []

for index, row in test_csv.iterrows():
  #  print(test_csv["text"][index])
    review = re.sub('[^a-zA-Z]', ' ', test_csv["text"][index])
    review = review.lower()
    review = review.split()
    review = [lemmatizer.lemmatize(word) for word in review if not word in␣
 ↪set(stopwords)]
    test_processed =[ ' '.join(review)]
  #  print(test_processed)
    test_input = tf_idf.transform(test_processed)
    test_input.shape
    res=naive_bayes_classifier.predict(test_input)[0]
  #  print(res)
    authors.append(res)
```

[129]:
```
cars
```

[129]:
```
['dickens',
 'defoe',
 'doyle',
 'defoe',
 'twain',
 'defoe',
 'doyle',
 'dickens',
 'dickens',
 'twain',
 'defoe',
 'twain',
 'dickens',
 'twain',
 'doyle',
 'twain',
```

```
'defoe',
'twain',
'defoe',
'defoe',
'dickens',
'doyle',
'dickens',
'dickens',
'defoe',
'defoe',
'defoe',
'twain',
'defoe',
'defoe',
'dickens',
'doyle',
'defoe',
'twain',
'doyle',
'defoe',
'twain',
'defoe',
'doyle',
'defoe',
'defoe',
'defoe',
'twain',
'dickens',
'doyle',
'defoe',
'doyle',
'dickens',
'dickens',
'dickens',
'doyle',
'dickens',
'dickens',
'dickens',
'doyle',
'defoe',
'doyle',
'doyle',
'dickens',
'doyle',
'defoe',
'defoe',
'defoe',
```

```
'twain',
'doyle',
'doyle',
'defoe',
'defoe',
'defoe',
'twain',
'dickens',
'twain',
'dickens',
'doyle',
'twain',
'defoe',
'dickens',
'doyle',
'doyle',
'defoe',
'dickens',
'doyle',
'dickens',
'defoe',
'twain',
'doyle',
'twain',
'defoe',
'dickens',
'twain',
'twain',
'twain',
'defoe',
'twain',
'twain',
'doyle',
'defoe',
'doyle',
'defoe',
'doyle',
'doyle',
'defoe',
'twain',
'defoe',
'twain',
'dickens',
'doyle',
'dickens',
'defoe',
'dickens',
```

```
'twain',
'twain',
'doyle',
'doyle',
'doyle',
'defoe',
'dickens',
'defoe',
'dickens',
'defoe',
'twain',
'doyle',
'twain',
'defoe',
'dickens',
'doyle',
'dickens',
'defoe',
'defoe',
'defoe',
'doyle',
'dickens',
'dickens',
'defoe',
'doyle',
'dickens',
'defoe',
'twain',
'dickens',
'dickens',
'dickens',
'dickens',
'doyle',
'dickens',
'twain',
'defoe',
'dickens',
'twain',
'dickens',
'twain',
'doyle',
'twain',
'dickens',
'doyle',
'dickens',
'twain',
'defoe',
```

```
'twain',
'dickens',
'twain',
'defoe',
'twain',
'doyle',
'doyle',
'dickens',
'doyle',
'defoe',
'dickens',
'doyle',
'dickens',
'defoe',
'dickens',
'twain',
'defoe',
'dickens',
'defoe',
'dickens',
'doyle',
'doyle',
'doyle',
'dickens',
'defoe',
'dickens',
'dickens',
'twain',
'defoe',
'twain',
'doyle',
'doyle',
'doyle',
'doyle',
'twain',
'doyle',
'dickens',
'dickens',
'defoe',
'twain',
'doyle',
'twain',
'dickens',
'defoe',
'dickens',
'dickens',
'doyle',
```

```
'dickens',
'dickens',
'defoe',
'dickens',
'defoe',
'defoe',
'twain',
'doyle',
'doyle',
'twain',
'defoe',
'twain',
'dickens',
'defoe',
'dickens',
'twain',
'dickens',
'defoe',
'twain',
'twain',
'dickens',
'doyle',
'dickens',
'defoe',
'dickens',
'doyle',
'twain',
'twain',
'defoe',
'dickens',
'defoe',
'dickens',
'twain',
'twain',
'defoe',
'defoe',
'doyle',
'defoe',
'defoe',
'doyle',
'doyle',
'dickens',
'doyle',
'defoe',
'doyle',
'doyle',
'doyle',
```

```
'doyle',
'doyle',
'doyle',
'twain',
'defoe',
'defoe',
'twain',
'doyle',
'doyle',
'doyle',
'dickens',
'twain',
'dickens',
'twain',
'defoe',
'dickens',
'doyle',
'defoe',
'dickens',
'doyle',
'doyle',
'doyle',
'dickens',
'twain',
'twain',
'doyle',
'twain',
'doyle',
'dickens',
'defoe',
'dickens',
'dickens',
'doyle',
'twain',
'defoe',
'twain',
'doyle',
'defoe',
'dickens',
'dickens',
'twain',
'doyle',
'defoe',
'dickens',
'defoe',
'doyle',
'twain',
```

```
'dickens',
'defoe',
'dickens',
'doyle',
'doyle',
'defoe',
'doyle',
'doyle',
'defoe',
'defoe',
'doyle',
'defoe',
'defoe',
'doyle',
'dickens',
'defoe',
'doyle',
'doyle',
'defoe',
'dickens',
'twain',
'dickens',
'twain',
'dickens',
'doyle',
'dickens',
'dickens',
'dickens',
'dickens',
'doyle',
'defoe',
'dickens',
'defoe',
'doyle',
'dickens',
'dickens',
'defoe',
'doyle',
'twain',
'doyle',
'dickens',
'dickens',
'dickens',
'defoe',
'defoe',
'twain',
'dickens',
```

```
'doyle',
'doyle',
'twain',
'defoe',
'defoe',
'defoe',
'dickens',
'twain',
'twain',
'dickens',
'defoe',
'twain',
'doyle',
'defoe',
'twain',
'defoe',
'defoe',
'defoe',
'dickens',
'defoe',
'dickens',
'doyle',
'twain',
'defoe',
'twain',
'dickens',
'dickens',
'twain',
'defoe',
'defoe',
'doyle',
'doyle',
'twain',
'twain',
'dickens',
'doyle',
'defoe',
'defoe',
'doyle',
'defoe',
'twain',
'twain',
'doyle',
'twain',
'doyle',
'twain',
'twain',
```

```
'twain',
'defoe',
'doyle',
'defoe',
'doyle',
'doyle',
'doyle',
'defoe',
'defoe',
'defoe',
'defoe',
'dickens',
'doyle',
'doyle',
'dickens',
'twain',
'defoe',
'defoe',
'doyle',
'doyle',
'dickens',
'dickens',
'defoe',
'dickens',
'defoe',
'doyle',
'defoe',
'dickens',
'twain',
'defoe',
'twain',
'twain',
'defoe',
'twain',
'doyle',
'twain',
'twain',
'dickens',
'defoe',
'doyle',
'defoe',
'twain',
'doyle',
'defoe',
'defoe',
'dickens',
'dickens',
```

```
'doyle',
'defoe',
'doyle',
'dickens',
'twain',
'doyle',
'defoe',
'doyle',
'twain',
'doyle',
'doyle',
'defoe',
'defoe',
'defoe',
'dickens',
'dickens',
'defoe',
'defoe',
'doyle',
'doyle',
'doyle',
'doyle',
'dickens',
'defoe',
'doyle',
'doyle',
'doyle',
'twain',
'twain',
'dickens',
'twain',
'twain',
'twain',
'doyle',
'twain',
'twain',
'twain',
'doyle',
'doyle',
'dickens',
'doyle',
'dickens',
'dickens',
'doyle',
'twain',
'doyle',
'twain',
```

```
'doyle',
'twain',
'doyle',
'doyle',
'twain',
'doyle',
'doyle',
'doyle',
'doyle',
'twain',
'doyle',
'twain',
'doyle',
'doyle',
'dickens',
'doyle',
'dickens',
'doyle',
'dickens',
'defoe',
'twain',
'doyle',
'twain',
'twain',
'dickens',
'doyle',
'defoe',
'defoe',
'doyle',
'defoe',
'dickens',
'twain',
'defoe',
'twain',
'twain',
'twain',
'defoe',
'defoe',
'dickens',
'defoe',
'doyle',
'twain',
'doyle',
'twain',
'doyle',
'defoe',
'doyle',
```

```
'dickens',
'dickens',
'defoe',
'defoe',
'defoe',
'doyle',
'doyle',
'defoe',
'twain',
'defoe',
'twain',
'doyle',
'doyle',
'doyle',
'dickens',
'defoe',
'twain',
'twain',
'defoe',
'doyle',
'twain',
'twain',
'dickens',
'defoe',
'doyle',
'defoe',
'dickens',
'dickens',
'doyle',
'dickens',
'doyle',
'defoe',
'doyle',
'defoe',
'dickens',
'dickens',
'dickens',
'twain',
'defoe',
'defoe',
'defoe',
'doyle',
'doyle',
'doyle',
'twain',
'defoe',
'defoe',
```

```
'twain',
'dickens',
'defoe',
'defoe',
'defoe',
'doyle',
'doyle',
'doyle',
'twain',
'dickens',
'twain',
'twain',
'defoe',
'doyle',
'dickens',
'twain',
'doyle',
'dickens',
'doyle',
'twain',
'dickens',
'twain',
'defoe',
'doyle',
'defoe',
'defoe',
'twain',
'defoe',
'defoe',
'dickens',
'defoe',
'twain',
'doyle',
'dickens',
'doyle',
'twain',
'doyle',
'twain',
'doyle',
'dickens',
'twain',
'doyle',
'defoe',
'twain',
'defoe',
'twain',
'doyle',
```

```
'defoe',
'twain',
'doyle',
'defoe',
'twain',
'doyle',
'defoe',
'defoe',
'defoe',
'dickens',
'doyle',
'twain',
'doyle',
'doyle',
'defoe',
'twain',
'doyle',
'twain',
'doyle',
'dickens',
'twain',
'defoe',
'twain',
'twain',
'doyle',
'defoe',
'doyle',
'defoe',
'twain',
'defoe',
'defoe',
'dickens',
'dickens',
'doyle',
'defoe',
'dickens',
'twain',
'twain',
'doyle',
'defoe',
'twain',
'dickens',
'twain',
'defoe',
'twain',
'doyle',
'defoe',
```

```
'dickens',
'doyle',
'defoe',
'doyle',
'defoe',
'doyle',
'twain',
'twain',
'twain',
'doyle',
'defoe',
'twain',
'defoe',
'twain',
'twain',
'doyle',
'twain',
'twain',
'doyle',
'defoe',
'twain',
'twain',
'defoe',
'dickens',
'dickens',
'doyle',
'twain',
'dickens',
'defoe',
'defoe',
'doyle',
'dickens',
'twain',
'doyle',
'defoe',
'defoe',
'twain',
'dickens',
'doyle',
'twain',
'defoe',
'defoe',
'twain',
'dickens',
'defoe',
'doyle',
'dickens',
```

```
'defoe',
'defoe',
'twain',
'dickens',
'doyle',
'defoe',
'dickens',
'defoe',
'defoe',
'defoe',
'doyle',
'defoe',
'doyle',
'twain',
'twain',
'defoe',
'doyle',
'doyle',
'doyle',
'twain',
'defoe',
'twain',
'twain',
'twain',
'twain',
'defoe',
'dickens',
'defoe',
'doyle',
'defoe',
'twain',
'dickens',
'doyle',
'twain',
'defoe',
'twain',
'defoe',
'defoe',
'dickens',
'defoe',
'defoe',
'twain',
'doyle',
'defoe',
'defoe',
'defoe',
'dickens',
```

```
'doyle',
'twain',
'dickens',
'twain',
'defoe',
'twain',
'dickens',
'dickens',
'defoe',
'defoe',
'doyle',
'doyle',
'defoe',
'defoe',
'twain',
'defoe',
'defoe',
'defoe',
'doyle',
'doyle',
'dickens',
'doyle',
'twain',
'twain',
'defoe',
'twain',
'defoe',
'doyle',
'twain',
'doyle',
'twain',
'defoe',
'dickens',
'twain',
'twain',
'dickens',
'dickens',
'defoe',
'dickens',
'dickens',
'dickens',
'doyle',
'twain',
'doyle',
'defoe',
'doyle',
'twain',
```

```
'dickens',
'doyle',
'twain',
'defoe',
'twain',
'dickens',
'defoe',
'doyle',
'defoe',
'defoe',
'defoe',
'defoe',
'defoe',
'dickens',
'defoe',
'defoe',
'doyle',
'twain',
'defoe',
'defoe',
'twain',
'dickens',
'twain',
'dickens',
'dickens',
'doyle',
'doyle',
'twain',
'twain',
'twain',
'twain',
'defoe',
'dickens',
'defoe',
'doyle',
'defoe',
'dickens',
'defoe',
'dickens',
'doyle',
'twain',
'dickens',
'dickens',
'doyle',
'dickens',
'dickens',
'twain',
```

```
'dickens',
'defoe',
'twain',
'defoe',
'doyle',
'dickens',
'twain',
'dickens',
'twain',
'defoe',
'dickens',
'defoe',
'doyle',
'doyle',
'defoe',
'defoe',
'doyle',
'defoe',
'defoe',
'doyle',
'defoe',
'doyle',
'dickens',
'twain',
'twain',
'defoe',
'doyle',
'defoe',
'doyle',
'dickens',
'twain',
'twain',
'doyle',
'doyle',
'defoe',
'doyle',
'dickens',
'doyle',
'dickens',
'defoe',
'doyle',
'defoe',
'doyle',
'dickens',
'twain',
'doyle',
'defoe',
```

```
'doyle',
'defoe',
'defoe',
'doyle',
'doyle',
'doyle',
'twain',
'defoe',
'dickens',
'dickens',
'doyle',
'twain',
'defoe',
'dickens',
'defoe',
'twain',
'twain',
'twain',
'defoe',
'doyle',
'defoe',
'defoe',
'defoe',
'defoe',
'dickens',
'defoe',
'doyle',
'dickens',
'defoe',
'twain',
'dickens',
'twain',
'defoe',
'doyle',
'dickens',
'twain',
'doyle',
'defoe',
'dickens',
'defoe',
'twain',
'defoe',
'doyle',
'doyle',
'twain',
'dickens',
'defoe',
```

```
'dickens',
'defoe',
'defoe',
'dickens',
'doyle',
'twain',
'defoe',
'defoe',
'dickens',
'defoe',
'twain',
'defoe',
'defoe',
'twain',
'dickens',
'doyle',
'twain',
'defoe',
'defoe',
'dickens',
'doyle',
'dickens',
'twain',
'dickens',
'doyle',
'dickens',
'dickens',
'doyle',
'defoe',
'defoe',
'twain',
'defoe',
'defoe',
'defoe',
'defoe',
'dickens',
'doyle',
'dickens',
'defoe',
'twain',
'defoe',
'doyle',
'twain',
'dickens',
...]
```

[130]:
```
#result.insert(0, 'id', test_df['id'])



#result = pd.DataFrame(cars, columns=['defoe','dickens','twain','doyle'])
#result.head()
result = pd.DataFrame(cars, columns=['author'])
#result.insert(0, 'id', test_df['id'])
result.head()
```

[130]:

	author
0	dickens
1	defoe
2	doyle
3	defoe
4	twain

[131]:
```
result.to_csv('results.csv', index=False, float_format='%.20f')
```

[ ]:
```

```