scikit-learn_ersterKorpus

December 22, 2020

[1]:
```
import sklearn
import numpy as np
from glob import glob
from sklearn import datasets
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.linear_model import SGDClassifier
from sklearn import metrics
from sklearn.pipeline import Pipeline

import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import CountVectorizer
```

```
[nltk_data] Downloading package stopwords to /home/piah/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

[24]:
```
categories = ['Balladen', 'Lyrik']
```

[25]:
```
docs_to_train = sklearn.datasets.load_files("/home/piah/Dokumente/Uni/
↪Projektarbeit/Projektarbeit_LyrikGattungszuweisung/scikit-learn/ersterKorpus/
↪", description=None, categories=categories, load_content=True, shuffle=True,␣
↪encoding='utf-8', decode_error='strict', random_state=42)
```

[26]:
```
pprint(list(docs_to_train.target_names))
```

```
Pretty printing has been turned OFF
```

[27]:
```
X_train, X_test, y_train, y_test = train_test_split(docs_to_train.data,
    docs_to_train.target, test_size=0.4)
```

[28]:
```
count_vect = CountVectorizer(stopwords.words('german'))
```

```
/home/piah/.local/lib/python3.8/site-packages/sklearn/utils/validation.py:67:
FutureWarning: Pass input=['aber', 'alle', 'allem', 'allen', 'aller', 'alles',
'als', 'also', 'am', 'an', 'ander', 'andere', 'anderem', 'anderen', 'anderer',
```

```
'anderes', 'anderm', 'andern', 'anderr', 'anders', 'auch', 'auf', 'aus', 'bei',
'bin', 'bis', 'bist', 'da', 'damit', 'dann', 'der', 'den', 'des', 'dem', 'die',
'das', 'dass', 'daß', 'derselbe', 'derselben', 'denselben', 'desselben',
'demselben', 'dieselbe', 'dieselben', 'dasselbe', 'dazu', 'dein', 'deine',
'deinem', 'deinen', 'deiner', 'deines', 'denn', 'derer', 'dessen', 'dich',
'dir', 'du', 'dies', 'diese', 'diesem', 'diesen', 'dieser', 'dieses', 'doch',
'dort', 'durch', 'ein', 'eine', 'einem', 'einen', 'einer', 'eines', 'einig',
'einige', 'einigem', 'einigen', 'einiger', 'einiges', 'einmal', 'er', 'ihn',
'ihm', 'es', 'etwas', 'euer', 'eure', 'eurem', 'euren', 'eurer', 'eures', 'für',
'gegen', 'gewesen', 'hab', 'habe', 'haben', 'hat', 'hatte', 'hatten', 'hier',
'hin', 'hinter', 'ich', 'mich', 'mir', 'ihr', 'ihre', 'ihrem', 'ihren', 'ihrer',
'ihres', 'euch', 'im', 'in', 'indem', 'ins', 'ist', 'jede', 'jedem', 'jeden',
'jeder', 'jedes', 'jene', 'jenem', 'jenen', 'jener', 'jenes', 'jetzt', 'kann',
'kein', 'keine', 'keinem', 'keinen', 'keiner', 'keines', 'können', 'könnte',
'machen', 'man', 'manche', 'manchem', 'manchen', 'mancher', 'manches', 'mein',
'meine', 'meinem', 'meinen', 'meiner', 'meines', 'mit', 'muss', 'musste',
'nach', 'nicht', 'nichts', 'noch', 'nun', 'nur', 'ob', 'oder', 'ohne', 'sehr',
'sein', 'seine', 'seinem', 'seinen', 'seiner', 'seines', 'selbst', 'sich',
'sie', 'ihnen', 'sind', 'so', 'solche', 'solchem', 'solchen', 'solcher',
'solches', 'soll', 'sollte', 'sondern', 'sonst', 'über', 'um', 'und', 'uns',
'unsere', 'unserem', 'unseren', 'unser', 'unseres', 'unter', 'viel', 'vom',
'von', 'vor', 'während', 'war', 'waren', 'warst', 'was', 'weg', 'weil',
'weiter', 'welche', 'welchem', 'welchen', 'welcher', 'welches', 'wenn', 'werde',
'werden', 'wie', 'wieder', 'will', 'wir', 'wird', 'wirst', 'wo', 'wollen',
'wollte', 'würde', 'würden', 'zu', 'zum', 'zur', 'zwar', 'zwischen'] as keyword
args. From version 0.25 passing these as positional arguments will result in an
error
  warnings.warn("Pass {} as keyword args. From version 0.25 "
```

[29]:
```
X_train_counts = count_vect.fit_transform(raw_documents=X_train)
```

[30]:
```
tfidf_transformer = TfidfTransformer(use_idf=True)
```

[31]:
```
X_train_tfidf = tfidf_transformer.fit_transform(X_train_counts)
```

[32]:
```
count_vect = CountVectorizer(stopwords.words('german'))
X_test_counts = count_vect.fit_transform(raw_documents=X_test)
```

```
/home/piah/.local/lib/python3.8/site-packages/sklearn/utils/validation.py:67:
FutureWarning: Pass input=['aber', 'alle', 'allem', 'allen', 'aller', 'alles',
'als', 'also', 'am', 'an', 'ander', 'andere', 'anderem', 'anderen', 'anderer',
'anderes', 'anderm', 'andern', 'anderr', 'anders', 'auch', 'auf', 'aus', 'bei',
'bin', 'bis', 'bist', 'da', 'damit', 'dann', 'der', 'den', 'des', 'dem', 'die',
'das', 'dass', 'daß', 'derselbe', 'derselben', 'denselben', 'desselben',
'demselben', 'dieselbe', 'dieselben', 'dasselbe', 'dazu', 'dein', 'deine',
'deinem', 'deinen', 'deiner', 'deines', 'denn', 'derer', 'dessen', 'dich',
'dir', 'du', 'dies', 'diese', 'diesem', 'diesen', 'dieser', 'dieses', 'doch',
'dort', 'durch', 'ein', 'eine', 'einem', 'einen', 'einer', 'eines', 'einig',
```

```
'einige', 'einigem', 'einigen', 'einiger', 'einiges', 'einmal', 'er', 'ihn',
'ihm', 'es', 'etwas', 'euer', 'eure', 'eurem', 'euren', 'eurer', 'eures', 'für',
'gegen', 'gewesen', 'hab', 'habe', 'haben', 'hat', 'hatte', 'hatten', 'hier',
'hin', 'hinter', 'ich', 'mich', 'mir', 'ihr', 'ihre', 'ihrem', 'ihren', 'ihrer',
'ihres', 'euch', 'im', 'in', 'indem', 'ins', 'ist', 'jede', 'jedem', 'jeden',
'jeder', 'jedes', 'jene', 'jenem', 'jenen', 'jener', 'jenes', 'jetzt', 'kann',
'kein', 'keine', 'keinem', 'keinen', 'keiner', 'keines', 'können', 'könnte',
'machen', 'man', 'manche', 'manchem', 'manchen', 'mancher', 'manches', 'mein',
'meine', 'meinem', 'meinen', 'meiner', 'meines', 'mit', 'muss', 'musste',
'nach', 'nicht', 'nichts', 'noch', 'nun', 'nur', 'ob', 'oder', 'ohne', 'sehr',
'sein', 'seine', 'seinem', 'seinen', 'seiner', 'seines', 'selbst', 'sich',
'sie', 'ihnen', 'sind', 'so', 'solche', 'solchem', 'solchen', 'solcher',
'solches', 'soll', 'sollte', 'sondern', 'sonst', 'über', 'um', 'und', 'uns',
'unsere', 'unserem', 'unseren', 'unser', 'unseres', 'unter', 'viel', 'vom',
'von', 'vor', 'während', 'war', 'waren', 'warst', 'was', 'weg', 'weil',
'weiter', 'welche', 'welchem', 'welchen', 'welcher', 'welches', 'wenn', 'werde',
'werden', 'wie', 'wieder', 'will', 'wir', 'wird', 'wirst', 'wo', 'wollen',
'wollte', 'würde', 'würden', 'zu', 'zum', 'zur', 'zwar', 'zwischen'] as keyword
args. From version 0.25 passing these as positional arguments will result in an
error
  warnings.warn("Pass {} as keyword args. From version 0.25 "
```

[33]:
```
tfidf_transformer = TfidfTransformer(use_idf=True)
X_test_tfidf = tfidf_transformer.fit_transform(X_test_counts)
```

[34]:
```
text_clf = Pipeline([('vect', CountVectorizer(stopwords.words('german'))),
    ('tfidf', TfidfTransformer(use_idf=True)),
    ('clf', SGDClassifier(loss='hinge', penalty='l2', alpha=1e-3,␣
 ↪random_state=42,
    verbose=1)),])
```

```
/home/piah/.local/lib/python3.8/site-packages/sklearn/utils/validation.py:67:
FutureWarning: Pass input=['aber', 'alle', 'allem', 'allen', 'aller', 'alles',
'als', 'also', 'am', 'an', 'ander', 'andere', 'anderem', 'anderen', 'anderer',
'anderes', 'anderm', 'andern', 'anderr', 'anders', 'auch', 'auf', 'aus', 'bei',
'bin', 'bis', 'bist', 'da', 'damit', 'dann', 'der', 'den', 'des', 'dem', 'die',
'das', 'dass', 'daß', 'derselbe', 'derselben', 'denselben', 'desselben',
'demselben', 'dieselbe', 'dieselben', 'dasselbe', 'dazu', 'dein', 'deine',
'deinem', 'deinen', 'deiner', 'deines', 'denn', 'derer', 'dessen', 'dich',
'dir', 'du', 'dies', 'diese', 'diesem', 'diesen', 'dieser', 'dieses', 'doch',
'dort', 'durch', 'ein', 'eine', 'einem', 'einen', 'einer', 'eines', 'einig',
'einige', 'einigem', 'einigen', 'einiger', 'einiges', 'einmal', 'er', 'ihn',
'ihm', 'es', 'etwas', 'euer', 'eure', 'eurem', 'euren', 'eurer', 'eures', 'für',
'gegen', 'gewesen', 'hab', 'habe', 'haben', 'hat', 'hatte', 'hatten', 'hier',
'hin', 'hinter', 'ich', 'mich', 'mir', 'ihr', 'ihre', 'ihrem', 'ihren', 'ihrer',
'ihres', 'euch', 'im', 'in', 'indem', 'ins', 'ist', 'jede', 'jedem', 'jeden',
'jeder', 'jedes', 'jene', 'jenem', 'jenen', 'jener', 'jenes', 'jetzt', 'kann',
'kein', 'keine', 'keinem', 'keinen', 'keiner', 'keines', 'können', 'könnte',
```

```
'machen', 'man', 'manche', 'manchem', 'manchen', 'mancher', 'manches', 'mein',
'meine', 'meinem', 'meinen', 'meiner', 'meines', 'mit', 'muss', 'musste',
'nach', 'nicht', 'nichts', 'noch', 'nun', 'nur', 'ob', 'oder', 'ohne', 'sehr',
'sein', 'seine', 'seinem', 'seinen', 'seiner', 'seines', 'selbst', 'sich',
'sie', 'ihnen', 'sind', 'so', 'solche', 'solchem', 'solchen', 'solcher',
'solches', 'soll', 'sollte', 'sondern', 'sonst', 'über', 'um', 'und', 'uns',
'unsere', 'unserem', 'unseren', 'unser', 'unseres', 'unter', 'viel', 'vom',
'von', 'vor', 'während', 'war', 'waren', 'warst', 'was', 'weg', 'weil',
'weiter', 'welche', 'welchem', 'welchen', 'welcher', 'welches', 'wenn', 'werde',
'werden', 'wie', 'wieder', 'will', 'wir', 'wird', 'wirst', 'wo', 'wollen',
'wollte', 'würde', 'würden', 'zu', 'zum', 'zur', 'zwar', 'zwischen'] as keyword
args. From version 0.25 passing these as positional arguments will result in an
error
  warnings.warn("Pass {} as keyword args. From version 0.25 "
```

[35]:
```
text_clf.fit(X_train, y_train)
```

```
-- Epoch 1
Norm: 20.93, NNZs: 2455, Bias: 0.033949, T: 134, Avg. loss: 0.513908
Total training time: 0.00 seconds.
-- Epoch 2
Norm: 18.37, NNZs: 3103, Bias: 0.037882, T: 268, Avg. loss: 0.076309
Total training time: 0.00 seconds.
-- Epoch 3
Norm: 17.25, NNZs: 3570, Bias: 0.056765, T: 402, Avg. loss: 0.044714
Total training time: 0.00 seconds.
-- Epoch 4
Norm: 15.78, NNZs: 3750, Bias: 0.101801, T: 536, Avg. loss: 0.016480
Total training time: 0.00 seconds.
-- Epoch 5
Norm: 14.92, NNZs: 4141, Bias: 0.100163, T: 670, Avg. loss: 0.013405
Total training time: 0.00 seconds.
-- Epoch 6
Norm: 14.06, NNZs: 4230, Bias: 0.100478, T: 804, Avg. loss: 0.015114
Total training time: 0.00 seconds.
-- Epoch 7
Norm: 12.90, NNZs: 4340, Bias: 0.101213, T: 938, Avg. loss: 0.007200
Total training time: 0.00 seconds.
-- Epoch 8
Norm: 12.68, NNZs: 4502, Bias: 0.118036, T: 1072, Avg. loss: 0.010035
Total training time: 0.00 seconds.
-- Epoch 9
Norm: 12.70, NNZs: 4551, Bias: 0.110413, T: 1206, Avg. loss: 0.013927
Total training time: 0.00 seconds.
-- Epoch 10
Norm: 12.30, NNZs: 4609, Bias: 0.116907, T: 1340, Avg. loss: 0.008047
Total training time: 0.00 seconds.
-- Epoch 11
```

```
Norm: 11.82, NNZs: 4651, Bias: 0.104012, T: 1474, Avg. loss: 0.004059
Total training time: 0.00 seconds.
-- Epoch 12
Norm: 11.47, NNZs: 4785, Bias: 0.109941, T: 1608, Avg. loss: 0.004251
Total training time: 0.00 seconds.
-- Epoch 13
Norm: 11.47, NNZs: 4820, Bias: 0.115342, T: 1742, Avg. loss: 0.009165
Total training time: 0.00 seconds.
-- Epoch 14
Norm: 11.53, NNZs: 4901, Bias: 0.115505, T: 1876, Avg. loss: 0.007314
Total training time: 0.00 seconds.
-- Epoch 15
Norm: 11.21, NNZs: 4901, Bias: 0.115688, T: 2010, Avg. loss: 0.003342
Total training time: 0.00 seconds.
-- Epoch 16
Norm: 10.96, NNZs: 4901, Bias: 0.120173, T: 2144, Avg. loss: 0.005454
Total training time: 0.01 seconds.
Convergence after 16 epochs took 0.01 seconds
```

[35]:
```
Pipeline(steps=[('vect',
                 CountVectorizer(input=['aber', 'alle', 'allem', 'allen',
                                        'aller', 'alles', 'als', 'also', 'am',
                                        'an', 'ander', 'andere', 'anderem',
                                        'anderen', 'anderer', 'anderes',
                                        'anderm', 'andern', 'anderr', 'anders',
                                        'auch', 'auf', 'aus', 'bei', 'bin',
                                        'bis', 'bist', 'da', 'damit', 'dann',
...])),
                ('tfidf', TfidfTransformer()),
                ('clf',
                 SGDClassifier(alpha=0.001, random_state=42, verbose=1))])
```

[36]:
```
predicted = text_clf.predict(X_test)
```

[37]:
```
print (np.mean(predicted == y_test))
```

```
0.7777777777777778
```

[38]:
```
print(metrics.classification_report(y_test, predicted,
    target_names=docs_to_train.target_names))
```

```
              precision    recall  f1-score   support

    Balladen       0.76      0.97      0.85        59
       Lyrik       0.87      0.42      0.57        31

    accuracy                           0.78        90
   macro avg       0.81      0.69      0.71        90
```

```
weighted avg       0.80      0.78      0.75        90
```

[ ]: