! This class has been made inactive. No posts will be allowed until an instructor reactivates the class.

private note @106 3 views

HW4_2017310936_Md_Shirajum_Munir

1) Only use 10,000 documents for training and test sets.

Step 1: Read data , clean data and divided with training set and test set

```
import pandas as pd
import nltk
from nltk.stem.porter import PorterStemmer
from nltk.corpus import stopwords
from sklearn.svm import SVC
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import GridSearchCV

# Read Data
df = pd.read_csv('movie_data.csv', encoding='utf-8')
print(df.head(5))

# Cleaning text data
import re
def preprocessor(text):
    text = re.sub('<[^>]*>', '', text)
    emoticons = re.findall('(?::|;|=)(?:-)?(?:\)|\(|D|P)',
                           text)
    text = (re.sub('[\W]+', ' ', text.lower()) +
            ' '.join(emoticons).replace('-', ''))
    return text

df['review'] = df['review'].apply(preprocessor)
print(df.head())

# Data set for training and testing
X_train = df.loc[:10000, 'review'].values
y_train = df.loc[:10000, 'sentiment'].values
X_test = df.loc[10000:, 'review'].values
y_test = df.loc[10000:, 'sentiment'].values

Output:
0  In 1974, the teenager Martha Moxley (Maggie Gr...          1
1  OK... so... I really like Kris Kristofferson a...          0
2  ***SPOILER*** Do not read this, if you think a...          0
3  hi for all the people who have seen this wonde...          1
4  I recently bought the DVD, forgetting just how...          0
                                              review  sentiment
0  in 1974 the teenager martha moxley maggie grac...          1
1  ok so i really like kris kristofferson and his...          0
2   spoiler do not read this if you think about w...          0
3  hi for all the people who have seen this wonde...          1
4  i recently bought the dvd forgetting just how ...          0
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\munir\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
Fitting 5 folds for each of 12 candidates, totalling 60 fits
                                              review  sentiment
0  In 1974, the teenager Martha Moxley (Maggie Gr...          1
1  OK... so... I really like Kris Kristofferson a...          0
2  ***SPOILER*** Do not read this, if you think a...          0
3  hi for all the people who have seen this wonde...          1
4  I recently bought the DVD, forgetting just how...          0
                                              review  sentiment
0  In 1974, the teenager Martha Moxley (Maggie Gr...          1
1  OK... so... I really like Kris Kristofferson a...          0
2  ***SPOILER*** Do not read this, if you think a...          0
3  hi for all the people who have seen this wonde...          1
4  I recently bought the DVD, forgetting just how...          0
                                              review  sentiment
0  In 1974, the teenager Martha Moxley (Maggie Gr...          1
1  OK... so... I really like Kris Kristofferson a...          0
2  ***SPOILER*** Do not read this, if you think a...          0
3  hi for all the people who have seen this wonde...          1
4  I recently bought the DVD, forgetting just how...          0
                                              review  sentiment
0  In 1974, the teenager Martha Moxley (Maggie Gr...          1
1  OK... so... I really like Kris Kristofferson a...          0
2  ***SPOILER*** Do not read this, if you think a...          0
3  hi for all the people who have seen this wonde...          1
4  I recently bought the DVD, forgetting just how...          0
                                              review  sentiment
0  in 1974 the teenager martha moxley maggie grac...          1
1  ok so i really like kris kristofferson and his...          0
2   spoiler do not read this if you think about w...          0
3  hi for all the people who have seen this wonde...          1
4  i recently bought the dvd forgetting just how ...          0
```

```
                                              review  sentiment
0  in 1974 the teenager martha moxley maggie grac...          1
1  ok so i really like kris kristofferson and his...          0
2   spoiler do not read this if you think about w...          0
3  hi for all the people who have seen this wonde...          1
4  i recently bought the dvd forgetting just how ...          0
                                              review  sentiment
0  in 1974 the teenager martha moxley maggie grac...          1
1  ok so i really like kris kristofferson and his...          0
2   spoiler do not read this if you think about w...          0
3  hi for all the people who have seen this wonde...          1
4  i recently bought the dvd forgetting just how ...          0
                                              review  sentiment
0  in 1974 the teenager martha moxley maggie grac...          1
1  ok so i really like kris kristofferson and his...          0
2   spoiler do not read this if you think about w...          0
3  hi for all the people who have seen this wonde...          1
4  i recently bought the dvd forgetting just how ...          0
```

Step 2: Processing documents into tokens

```
# Processing documents into tokens
porter = PorterStemmer()
def tokenizer(text):
    return text.split()

def tokenizer_porter(text):
    return [porter.stem(word) for word in text.split()]

# Download the nltk package
nltk.download('stopwords')
stop = stopwords.words('english')

tfidf = TfidfVectorizer(strip_accents=None,
                        lowercase=False,
                        preprocessor=None)

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

2) Use the classification SVM.

Step 3: Training a SVM model for document classification, Fit and Performance Measurement

```
# Training a SVM model for document classification

param_grid = [{'vect__ngram_range': [(1, 1)],
               'vect__stop_words': [stop, None],
               'vect__tokenizer': [str.split] ,
               'clf__C': [1.0, 10.0, 100.0]},
              {'vect__ngram_range': [(1, 1)],
               'vect__stop_words': [stop, None],
               'vect__tokenizer': [str.split],
               'vect__use_idf':[False],
               'vect__norm':[None],
               'clf__C': [1.0, 10.0, 100.0]},
              ]

svm_tfidf = Pipeline([('vect', tfidf),
                      ('clf', SVC(random_state=1))])

gs_svm_tfidf = GridSearchCV(svm_tfidf, param_grid,
                            scoring='accuracy',
                            cv=5,
                            verbose=1,
                            n_jobs=-1)

# Fit and Performance Measurement
if __name__ == '__main__':
    gs_svm_tfidf.fit(X_train, y_train)
    print('Best parameter set: %s ' % gs_svm_tfidf.best_params_)
    print('CV Accuracy: %.3f' % gs_svm_tfidf.best_score_)
    clf = gs_svm_tfidf.best_estimator_
    print('Test Accuracy: %.3f' % clf.score(X_test, y_test))

Output:
```

```
[Parallel(n_jobs=-1)]: Done  42 tasks      | elapsed: 37.9min
[Parallel(n_jobs=-1)]: Done  60 out of  60 | elapsed: 47.7min finished
Best parameter set: {'clf__C': 100.0, 'vect__ngram_range': (1, 1), 'vect__norm': None, 'vect__stop_words': None, 'vect__tokenizer': <method 's
plit' of 'str' objects>, 'vect__use_idf': False}
CV Accuracy: 0.863
Test Accuracy: 0.872
```

hw4

This private post is only visible to Instructors and MD SHIRAJUM MUNIR

Updated 2 years ago by MD SHIRAJUM MUNIR

followup discussions *for lingering questions and comments*