1. Рубежный контроль №2

Бурашников Владимир Владимирович, группа ИУ5-22М. Вариант №1.

1.1. Задание

Необходимо решить задачу классификации текстов на основе любого выбранного Вами датасета. Классификация может быть бинарной или многоклассовой. Целевой признак из выбранного Вами датасета может иметь любой физический смысл, примером является задача анализа тональности текста.

Необходимо сформировать признаки на основе `CountVectorizer` или `TfidfVectorizer`.

В качестве классификаторов необходимо использовать один из классификаторов, не относящихся к наивным Байесовским методам (например, `LogisticRegression`), а также Multinomial Naive Bayes (MNB), Complement Naive Bayes (CNB), Bernoulli Naive Bayes. Для каждого метода необходимо оценить качество классификации с помощью хотя бы одной метрики качества классификации (например, `accuracy`).

Сделайте выводы о том, какой классификатор осуществляет более качественную классификацию на Вашем наборе данных.

1.2. Решение

1.2.1. Загрузка и предобработка данных

[22]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
import pandas as pd
```

[23]:
```
#rcv_train = fetch_rcv1(subset='train')
#rcv_test = fetch_rcv1(subset='test')
df = pd.read_json('./news.json', lines=True)
```

[24]:
```
df.head()
```

[24]:

```
            authors       category       date  \
0  Melissa Jeltsen           CRIME 2018-05-26
1    Andy McDonald  ENTERTAINMENT 2018-05-26
2       Ron Dicker  ENTERTAINMENT 2018-05-26
3       Ron Dicker  ENTERTAINMENT 2018-05-26
4       Ron Dicker  ENTERTAINMENT 2018-05-26

                                            headline  \
0  There Were 2 Mass Shootings In Texas Last Week…
1  Will Smith Joins Diplo And Nicky Jam For The 2…
2    Hugh Grant Marries For The First Time At Age 57
3  Jim Carrey Blasts 'Castrato' Adam Schiff And D…
4  Julianna Margulies Uses Donald Trump Poop Bags…

                                                link  \
0  https://www.huffingtonpost.com/entry/texas-ama…
1  https://www.huffingtonpost.com/entry/will-smit…
2  https://www.huffingtonpost.com/entry/hugh-gran…
```

```
3  https://www.huffingtonpost.com/entry/jim-carre…
4  https://www.huffingtonpost.com/entry/julianna-…

                                  short_description
0  She left her husband. He killed their children…
1                           Of course it has a song.
2  The actor and his longtime girlfriend Anna Ebe…
3  The actor gives Dems an ass-kicking for not fi…
4  The "Dietland" actress said using the bags is …
```

[87]:
```
headline = np.array(df['headline'])
category = np.array(df['category'])
# build train and test datasets

# Train/test splitting for 41 categories of news
from sklearn.model_selection import train_test_split
headline_train, headline_test, category_train, category_test =␣
 ↪train_test_split(headline, category, test_size=0.2, random_state=41)
```

[88]:
```
from sklearn.feature_extraction.text import CountVectorizer,␣
 ↪TfidfVectorizer

## Build Bag-Of-Words on train phrases
cv = CountVectorizer(stop_words='english',max_features=10000)
cv_train_features = cv.fit_transform(headline_train)
```

[89]:
```
# build TFIDF features on train reviews
tv = TfidfVectorizer(min_df=0.0, max_df=1.0, ngram_range=(1,2),
                     sublinear_tf=True)
tv_train_features = tv.fit_transform(headline_train)
```

[91]:
```
cv_test_features = cv.transform(headline_test)
tv_test_features = tv.transform(headline_test)
```

1.2.2. Обучение моделей

[110]:
```
from sklearn.metrics import accuracy_score
```

[113]:
```
from sklearn import metrics
import numpy as np

def accuracy(classifier,
                        train_features, train_labels,
                        test_features, test_labels):
    classifier.fit(train_features, train_labels)
    print('Accuracy:', metrics.accuracy_score(test_labels, classifier.
 ↪predict(test_features)))
```

[114]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB, ComplementNB, BernoulliNB
```

[115]:
```
lr = LogisticRegression(solver='lbfgs',penalty='l2', max_iter=100,␣
↪C=1,multi_class='auto')

lr_accuracy = accuracy(classifier=lr,train_features=cv_train_features,␣
↪train_labels=category_train,
                       test_features=cv_test_features,␣
↪test_labels=category_test)
```

```
Accuracy: 0.5750665903263549

/home/vladimir/PycharmProjects/giis_lab1/env/lib/python3.6/site-
packages/sklearn/linear_model/logistic.py:947: ConvergenceWarning: lbfgs␣
↪failed
to converge. Increase the number of iterations.
  "of iterations.", ConvergenceWarning)
```

[116]:
```
mu = MultinomialNB()

mu_accuracy = accuracy(classifier=mu,train_features=cv_train_features,␣
↪train_labels=category_train,
                       test_features=cv_test_features,␣
↪test_labels=category_test)
```

```
Accuracy: 0.5516915187573125
```

[117]:
```
co = ComplementNB()

co_accuracy = accuracy(classifier=co,train_features=cv_train_features,␣
↪train_labels=category_train,
                       test_features=cv_test_features,␣
↪test_labels=category_test)
```

```
Accuracy: 0.5401906848223843
```

[118]:
```
be = BernoulliNB()

be_accuracy = accuracy(classifier=be,train_features=cv_train_features,␣
↪train_labels=category_train,
                       test_features=cv_test_features,␣
↪test_labels=category_test)
```

```
Accuracy: 0.5414104702397252
```

1.2.3. Вывод

Метод LogisticRegression, лучше всего решает поставленную задачу многоклассовой классификации. Но значительно больше затрачивает ресурсов и времени.