Рубежный контроль №2

Тема: Методы обработки текстов.

Решение задачи классификации текстов.

Необходимо решить задачу классификации текстов на основе любого выбранного Вами датасета (кроме примера, который рассматривался в лекции). 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 два варианта векторизации признаков - на основе CountVectorizer и на основе TfidfVectorizer.

Группа: ИУ5-23М

Вариант: LinearSVC, LogisticRegression

```
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.svm import LinearSVC
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
import pandas as pd
import time
```

```
# Загрузка данных
df = pd.read_csv('cryptonews.csv')
```

```
df.head(10)
```

	date	sentiment	source	subject	text	title	url
0	2023-12-19 06:40:41	{'class': 'negative', 'polarity': -0.1, 'subje...	CryptoNews	altcoin	Grayscale CEO Michael Sonnenshein believes the...	Grayscale CEO Calls for Simultaneous Approval ...	https://cryptonews.comhttps://cryptonews.com/n...
1	2023-12-19 06:03:24	{'class': 'neutral', 'polarity': 0.0, 'subject...	CryptoNews	blockchain	In an exclusive interview with CryptoNews, Man...	Indian Government is Actively Collaborating Wi...	https://cryptonews.comhttps://cryptonews.com/n...
2	2023-12-19 05:55:14	{'class': 'positive', 'polarity': 0.05, 'subje...	CryptoNews	blockchain	According to the Federal Court ruling on Decem...	Judge Approves Settlement: Binance to Pay $1.5...	https://cryptonews.comhttps://cryptonews.com/n...
3	2023-12-19 05:35:26	{'class': 'positive', 'polarity': 0.5, 'subjec...	CoinTelegraph	blockchain	Some suggest EVM inscriptions are the latest w...	Why a gold rush for inscriptions has broken ha...	https://cointelegraph.com/news/inscriptions-ev...
4	2023-12-19 05:31:08	{'class': 'neutral', 'polarity': 0.0, 'subject...	CoinTelegraph	ethereum	A decision by bloXroute Labs to start censorin...	‘Concerning precedent’ — bloXroute Labs' MEV r...	https://cointelegraph.com/news/concerning-prec...
5	2023-12-19 05:25:00	{'class': 'negative', 'polarity': -0.01, 'subj...	CryptoPotato	bitcoin	Yonsei found that during BTC’s rally in early ...	Is This Why Bitcoin’s Price Rally Was Halted? ...	https://cryptopotato.com/is-this-why-bitcoins-...
6	2023-12-19 04:50:11	{'class': 'positive', 'polarity': 0.3, 'subjec...	CryptoNews	bitcoin	Cathie Wood led ARK Invest fund sold around 80...	Cathie Wood’s Ark Invest Sells $27.6 Million i...	https://cryptonews.comhttps://cryptonews.com/n...
7	2023-12-19 04:10:00	{'class': 'neutral', 'polarity': 0.0, 'subject...	CryptoPotato	bitcoin	Bitcoin's 150% surge pales in comparison to th...	Bitcoin Soared 150% in 2023 But These Companie...	https://cryptopotato.com/bitcoin-soared-150-in...
8	2023-12-19 04:00:01	{'class': 'neutral', 'polarity': 0.0, 'subject...	CryptoNews	blockchain	The South Korean city of Busan is edging close...	South Korean City Busan Names Digital Exchange...	https://cryptonews.comhttps://cryptonews.com/n...
9	2023-12-19 02:59:59	{'class': 'negative', 'polarity': -0.08, 'subj...	CoinTelegraph	bitcoin	The SEC has pushed back its decision on a rost...	SEC delays several Ethereum ETFs, pushing fina...	https://cointelegraph.com/news/sec-delays-ethe...

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31037 entries, 0 to 31036
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   date       31037 non-null  object
 1   sentiment  31037 non-null  object
 2   source     31037 non-null  object
 3   subject    31037 non-null  object
 4   text       31037 non-null  object
 5   title      31037 non-null  object
 6   url        31037 non-null  object
dtypes: object(7)
memory usage: 1.7+ MB
```

```
# проверим пропуски в данных и устраним их
na_mask = df.isna()
na_counts = na_mask.sum()
na_counts
```

```
date         0
sentiment    0
source       0
subject      0
text         0
title        0
url          0
dtype: int64
```

```
df.dropna(inplace=True)
na_mask = df.isna()
na_counts = na_mask.sum()
na_counts
```

```
date         0
sentiment    0
source       0
subject      0
text         0
title        0
url          0
dtype: int64
```

```
# Разделим набор данных на обучающую и тестувую выборки
X, Y = df['text'], df['source']
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=42)

time_arr = []
```

```
# векторизация признаков с помощью CountVectorizer
count_vect = CountVectorizer()
X_train_counts = count_vect.fit_transform(X_train)
X_test_counts = count_vect.transform(X_test)
```

```
# векторизация признаков с помощью TfidfVectorizer
tfidf_vect = TfidfVectorizer()
X_train_tfidf = tfidf_vect.fit_transform(X_train)
X_test_tfidf = tfidf_vect.transform(X_test)
```

```
# Произведем обучения вдух классификаторов (по варианту) для CountVectorizer

# LinearSVC
gbc = LinearSVC()
start_time = time.time()
gbc.fit(X_train_counts, y_train)
train_time = time.time() - start_time
time_arr.append(train_time)
pred_gbc_counts = gbc.predict(X_test_counts)
print("Точность (CountVectorizer + LinearSVC):", accuracy_score(y_test, pred_gbc_counts))

# Logistic Regression
lr = LogisticRegression(max_iter=1000)
start_time = time.time()
lr.fit(X_train_counts, y_train)
train_time = time.time() - start_time
time_arr.append(train_time)
pred_lr_counts = lr.predict(X_test_counts)
print("Точность (CountVectorizer + LogisticRegression):", accuracy_score(y_test, pred_lr_counts))
```

```
/Users/peterpechenkin/anaconda3/lib/python3.10/site-packages/sklearn/svm/_base.py:1244: ConvergenceWarning: Libline
ar failed to converge, increase the number of iterations.
  warnings.warn(
```

```
Точность (CountVectorizer + LinearSVC): 0.6659149484536082
Точность (CountVectorizer + LogisticRegression): 0.7007087628865979
```

```python
# Произведем обучения вдух классификаторов (по варианту) для TfidfVectorizer

# LinearSVC
gbc = LinearSVC()
start_time = time.time()
gbc.fit(X_train_tfidf, y_train)
train_time = time.time() - start_time
time_arr.append(train_time)
pred_gbc_tfidf = gbc.predict(X_test_tfidf)
print("Точность (TfidfVectorizer + LinearSVC):", accuracy_score(y_test, pred_gbc_tfidf))

# Logistic Regression
lr = LogisticRegression(max_iter=1000)
start_time = time.time()
lr.fit(X_train_tfidf, y_train)
train_time = time.time() - start_time
time_arr.append(train_time)
pred_lr_tfidf = lr.predict(X_test_tfidf)
print("Точность (TfidfVectorizer + LogisticRegression):", accuracy_score(y_test, pred_lr_tfidf))
```

```
Точность (TfidfVectorizer + LinearSVC): 0.6968427835051546
Точность (TfidfVectorizer + LogisticRegression): 0.7116623711340206
```

```python
from tabulate import tabulate

data = [
    ["(CountVectorizer + LogisticRegression)", accuracy_score(y_test, pred_lr_counts), time_arr[0]],
    ["(CountVectorizer + LinearSVC)", accuracy_score(y_test, pred_gbc_counts), time_arr[1]],
    ["(TfidfVectorizer + LogisticRegression)", accuracy_score(y_test, pred_lr_tfidf), time_arr[2]],
    ["(TfidfVectorizer + LinearSVC)", accuracy_score(y_test, pred_gbc_tfidf), time_arr[3]]
]

sorted_data = sorted(data, key=lambda x: x[1], reverse=True)

# Вывод отсортированных данных в виде таблицы
print(tabulate(sorted_data, ['Связка','Точность валидации', 'Время обучения'], tablefmt="grid"))
```

Связка	Точность валидации	Время обучения
(TfidfVectorizer + LogisticRegression)	0.711662	0.214736
(CountVectorizer + LogisticRegression)	0.700709	1.88814
(TfidfVectorizer + LinearSVC)	0.696843	1.65225
(CountVectorizer + LinearSVC)	0.665915	4.09705