Белоусов Евгений Александрович

ИУ5-23м

РК2

```
1 import numpy as np
2 import pandas as pd
```

```
1 !unzip /content/drive/MyDrive/Colab_data/MMO/fake_news.zip
```

```
Archive:  /content/drive/MyDrive/Colab_data/MMO/fake_news.zip
  inflating: news_articles.csv
```

```
1 data = pd.read_csv('news_articles.csv')
2 data.head()
```

	author	published	title	text	language	site_
0	Barracuda Brigade	2016-10-26T21:41:00.000+03:00	muslims busted they stole millions in govt ben...	print they should pay all the back all the mon...	english	100percentfedup.c
1	reasoning with facts	2016-10-29T08:47:11.259+03:00	re why did attorney general loretta lynch plea...	why did attorney general loretta lynch plead t...	english	100percentfedup.c
2	Barracuda Brigade	2016-10-31T01:41:49.479+02:00	breaking weiner cooperating with fbi on hillar...	red state \nfox news sunday reported this mor...	english	100percentfedup.c
3	Fed Up	2016-11-01T05:22:00.000+02:00	pin drop speech by father of daughter kidnappe...	email kayla mueller was a prisoner and torture...	english	100percentfedup.c
4	Fed Up	2016-11-01T21:56:00.000+02:00	fantastic trumps point plan	email healthcare reform to make	english	100percentfedup.c

```
1 data = data[data['language']=='english']
```

```
1 data['language'].unique()
```

```
    array(['english'], dtype=object)
```

```
1 data.keys()
```

```
    Index(['author', 'published', 'title', 'text', 'language', 'site_url',
           'main_img_url', 'type', 'label', 'title_without_stopwords',
           'text_without_stopwords', 'hasImage'],
          dtype='object')
```

```
1 data = data.drop(columns = ['author', 'published', 'title', 'language', 'site_url',
2        'main_img_url', 'type', 'title_without_stopwords',
3        'text_without_stopwords', 'hasImage'])
```

```
1 import sklearn
2 from sklearn.svm import LinearSVC
3 from sklearn.naive_bayes import MultinomialNB
4 from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
5 from sklearn.model_selection import cross_val_score
```

```
1 data = data.dropna()
```

```
1 tfidfv = TfidfVectorizer()
2 tfidf_features = tfidfv.fit_transform(data['text'])
3 tfidf_features
```

```
    <1972x42691 sparse matrix of type '<class 'numpy.float64'>'
            with 460362 stored elements in Compressed Sparse Row format>
```

```
1 countv = CountVectorizer()
2 countv_features = countv.fit_transform(data['text'])
3 countv_features
```

```
    <1972x42691 sparse matrix of type '<class 'numpy.int64'>'
            with 460362 stored elements in Compressed Sparse Row format>
```

```
1 y = data['label'].values
```

```
1 cross_val_score(LinearSVC(), tfidf_features, y, scoring='accuracy', cv=3).mean()
```

```
    0.5420820745798886
```

```
1 cross_val_score(LinearSVC(), countv_features, y, scoring='accuracy', cv=3).mean()
```

```
    /usr/local/lib/python3.7/dist-packages/sklearn/svm/_base.py:1208: ConvergenceWarning
```

```
  ConvergenceWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/svm/_base.py:1208: ConvergenceWarning
  ConvergenceWarning,
/usr/local/lib/python3.7/dist-packages/sklearn/svm/_base.py:1208: ConvergenceWarning
  ConvergenceWarning,
0.5055639600961664
```

```
1 cross_val_score(MultinomialNB(), tfidf_features, y, scoring='accuracy', cv=3).mean()
```

```
0.6075071824124577
```

```
1 cross_val_score(MultinomialNB(), countv_features, y, scoring='accuracy', cv=3).mean()
```

```
0.4792130265753116
```

Лучший accuracy достигается при сочитании MultinominalNB и tfidf vectorizer