Домашнее задание 4

Возьмите код, который мы писали на занятии, используйте его при выполнении задания.
https://colab.research.google.com/drive/1vKWpG4xlo_691URRDWCSfs9VmUSDzcJu

1.
Выберите тексты двух похожих тематик. Можно взять из 20newsgroups, как мы делали на занятии, можно взять свои. Сформируйте обучающую и тестовую выборку.

Задание на 8 баллов
a) Векторизуйте тексты при помощи `TfidfVectorizer()`, обучите классификатор и оцените его качество работы (надо посчитать `accuracy_score()`).

Задание на 9 баллов
b) Возьмите те же самые тексты и векторизуйте при помощи `CountVectorizer()`, это векторизатор на обычных частотах. Обучите классификатор (возьмите, например, `LogisticRegression()`, как мы делали на занятии), оцените его качество работы и сравните полученный результат с предыдущим.
c) Замените одну из тематик на тематику из другой области (менее похожую). Используйте `TfidfVectorizer()`, обучите классификатор, оцените его качество работы и сравните полученный результат с результатом для похожих тематик.

2.
Задание на 10 баллов
Возьмите датасет “Ирисы”. Обучите классификатор `RandomForestClassifier()`.
Какие признаки оказались наиболее важными для классификации?
Воспользуйтесь атрибутом `feature_importances_`, это массив, который содержит значения важности признаков. Порядок элеменов массива со значениями важности признаков соотносится с порядком элементов в массиве названий признаков.

```
clf = RandomForestClassifier()
clf.fit(X_train, y_train)
print(clf.feature_importances_)
```

Чтобы узнать, какому признаку какое измерение соответствует, нужно обратиться к атрибуту `feature_names`.

```
iris = datasets.load_iris()
print(iris.feature_names)
```