|  |  |
| --- | --- |
| lu135925on3bu_tmp_3360867a00ce4d37 | **Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования** **«Московский государственный технический университет имени Н.Э. Баумана** **(национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ Информатика и системы управления и искусственный интеллект

КАФЕДРА                  Системы обработки информации и управления

**Рубежный контроль №2** **по курсу «Методы машинного обучения в автоматизированных системах обработки информации и управления»**

Подготовили:

**Е Чжэмин**

**ИУ5И-24М**

08.05.2024

Проверил:

**Гапанюк Ю. Е.**

*2024 г*.

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.

В качестве классификаторов необходимо использовать два классификатора по варианту для Вашей группы:

Группа Классификатор №1 Классификатор №2

ИУ5И-24М GradientBoostingClassifier LogisticRegression

Для каждого метода необходимо оценить качество классификации. Сделайте вывод о том, какой вариант векторизации признаков в паре с каким классификатором показал лучшее качество.

(WELFake) — это набор данных из 72 134 новостных статей, из которых 35 028 реальных и 37 106 фейковых новостей. Для этого авторы объединили четыре популярных набора новостных данных (например, Kaggle, McIntire, Reuters, BuzzFeed Political), чтобы предотвратить переобучение классификаторов и предоставить больше текстовых данных для лучшего обучения машинному обучению.

Набор данных содержит четыре столбца: Серийный номер (начиная с 0); Заголовок (о заголовке текстовой новости); Текст (о содержании новости); и Метка (0 = подделка и 1 = настоящая).

### Загрузите набор данных WELFake и выполните начальное исследование

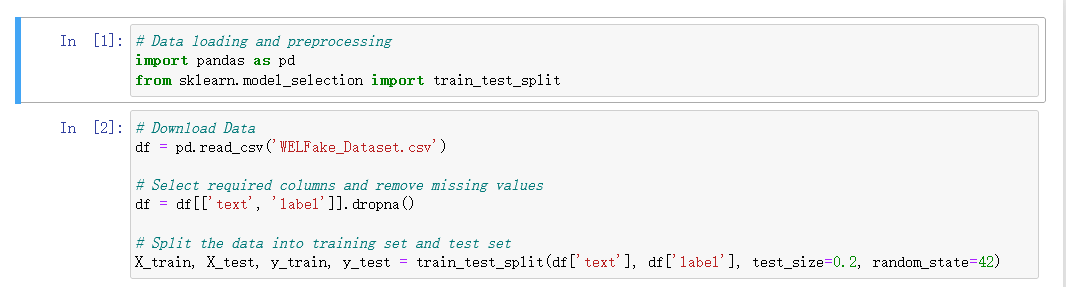


Рис.1

### Выполните предварительную обработку текста и извлечение признаков

Используйте CountVectorizer и TfidfVectorizer для преобразования текста в векторы признаков.

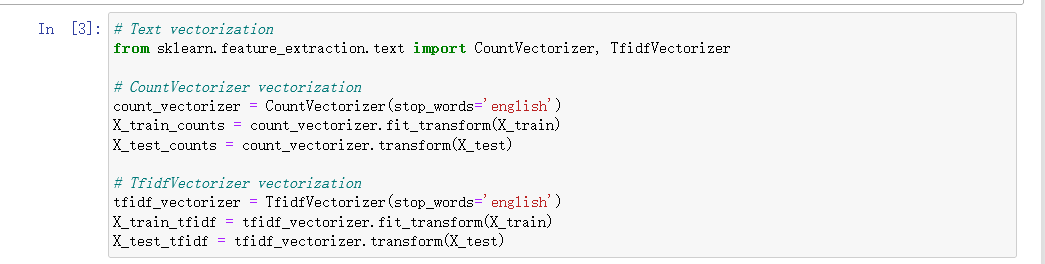


Рис. 2

### Используя классификаторы GradientBoostingClassifier и LogisticRegression для классификации векторов признаков соответственно, классификаторы будут обучены и их производительность будет оценена на тестовом наборе

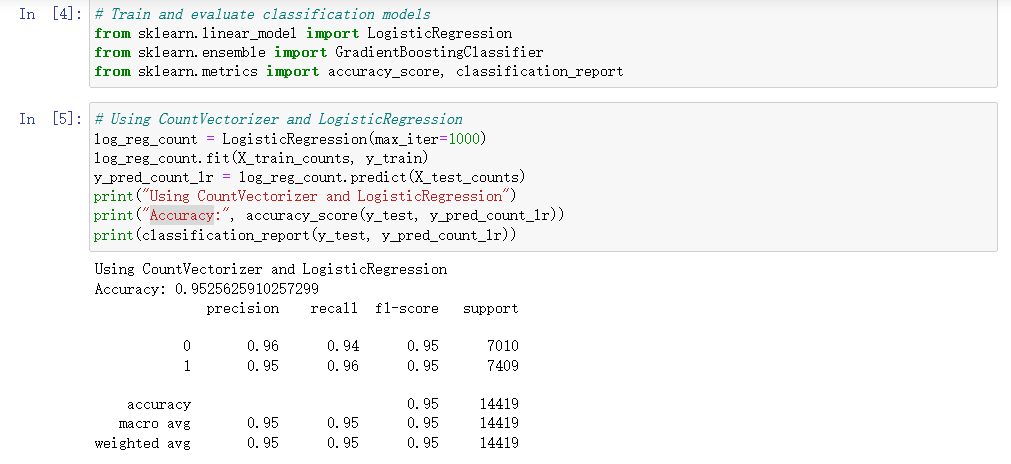


Рис.4

На основании отчета о результатах классификации можно сделать следующие выводы:

Для классификатора GradientBoostingClassifier, использующего признаки CountVectorizer, точность составляет 0,92, в то время как точность классификатора LogisticRegression, использующего признаки TfidfVectorizer, равна 0,95. Таким образом, сочетание признаков TfidfVectorizer с классификатором LogisticRegression работает лучше.

По другим показателям (например, recall, F1 score и т. д.) комбинация признаков TfidfVectorizer с классификатором LogisticRegression также работает лучше.

В целом, комбинация функций TfidfVectorizer и классификатора LogisticRegression имеет лучшие показатели для этой задачи.