РК2 ММО Савельев Алексей ИУ5-24М

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.

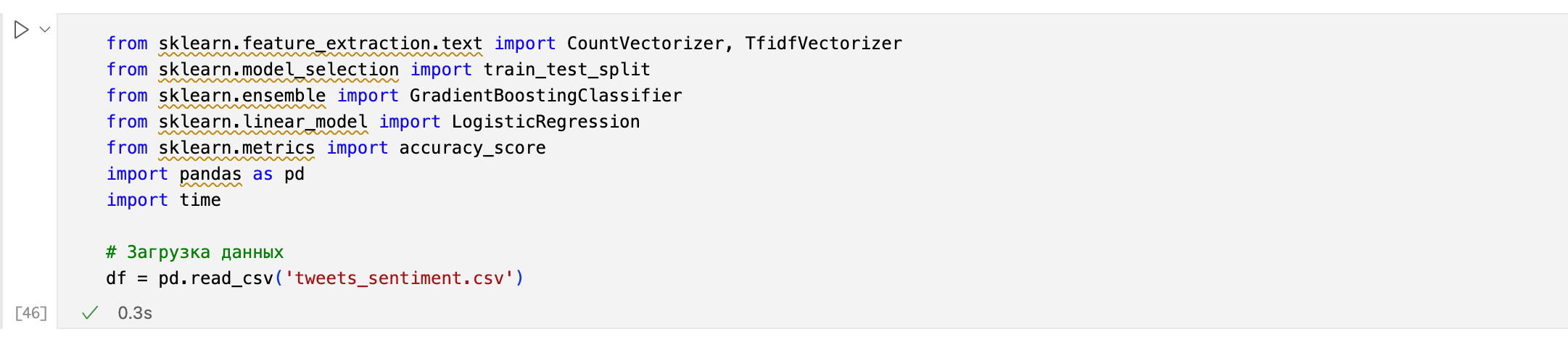
Необходимо сформировать два варианта векторизации признаков - на основе CountVectorizer и на основе TfidfVectorizer.

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

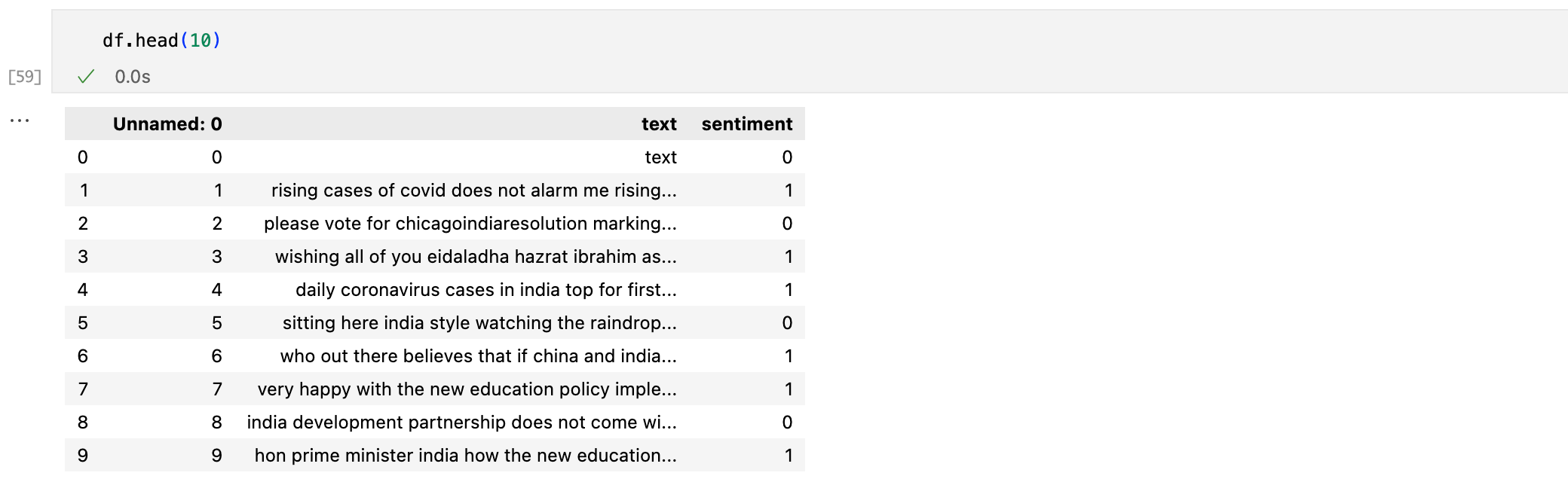
ИУ5-24М, ИУ5И-24М GradientBoostingClassifier LogisticRegression

**Решение**

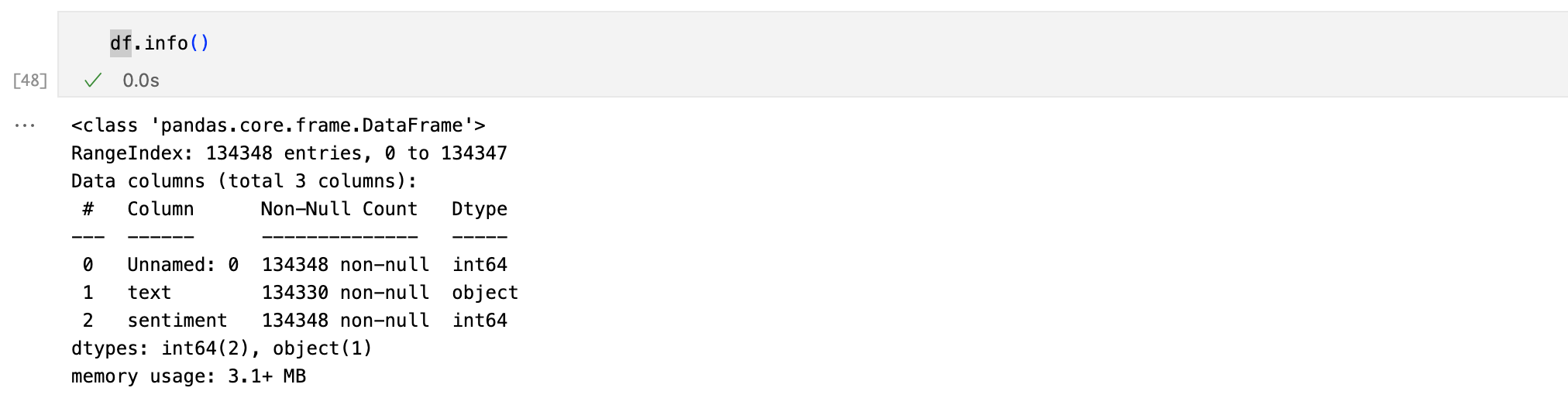
Загрузка набора данных (Набор твитов (с обозначенным настроением для каждого))



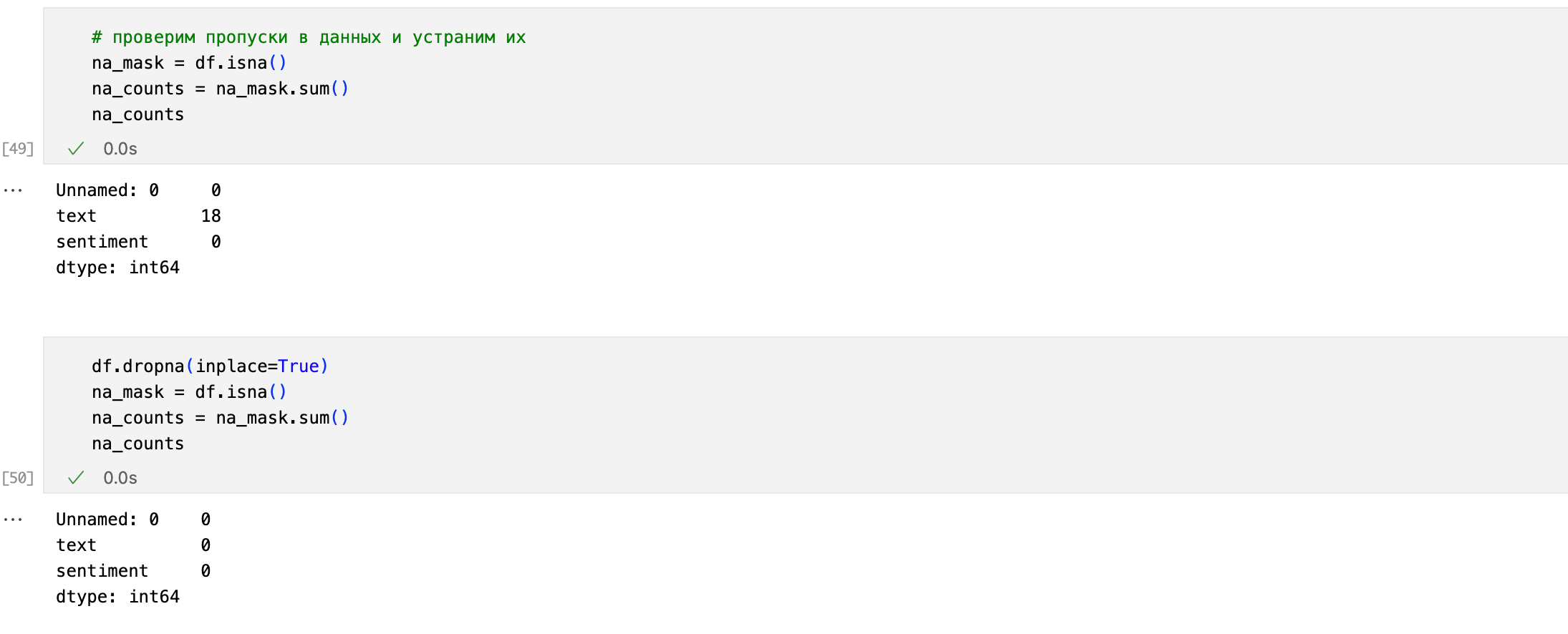
Данные датасета:



Информация по столбцам набора данных



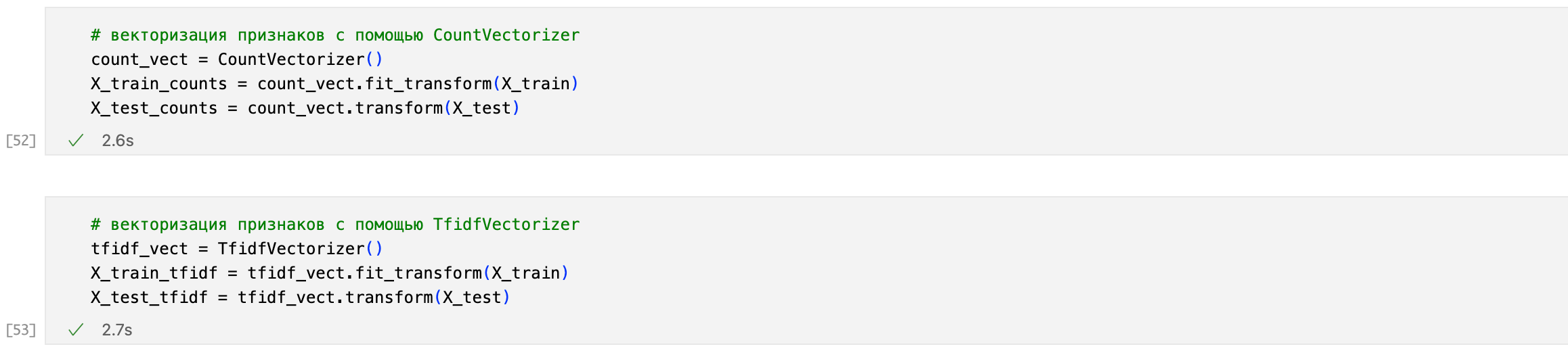
Количество пропусков по столбцам (и удаление их):



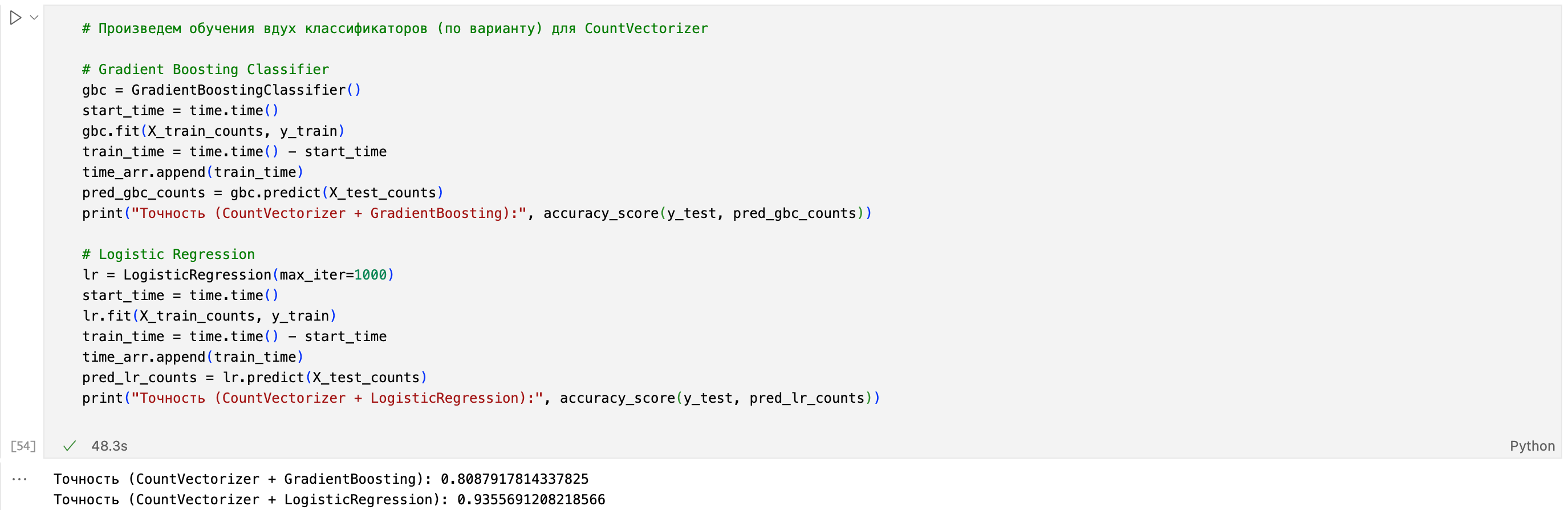
Разделение на тренировочную и тестовую выборки



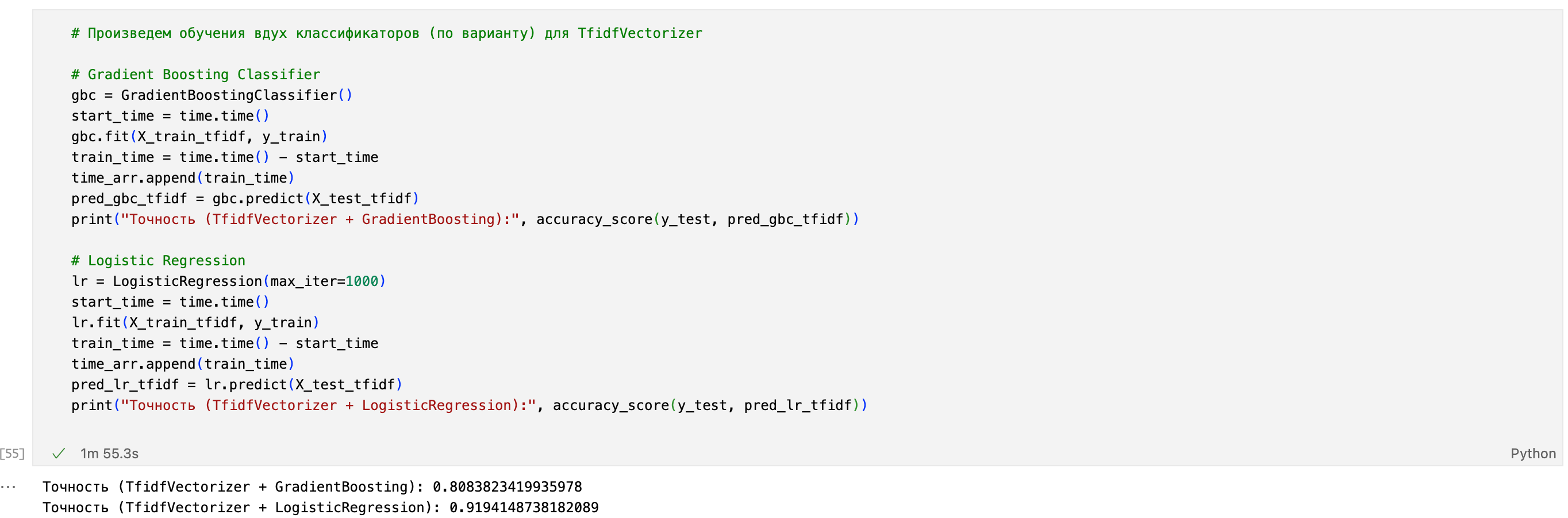
Формирование двух вариантов векторизации признаков:



Обучения для CountVectorizer:



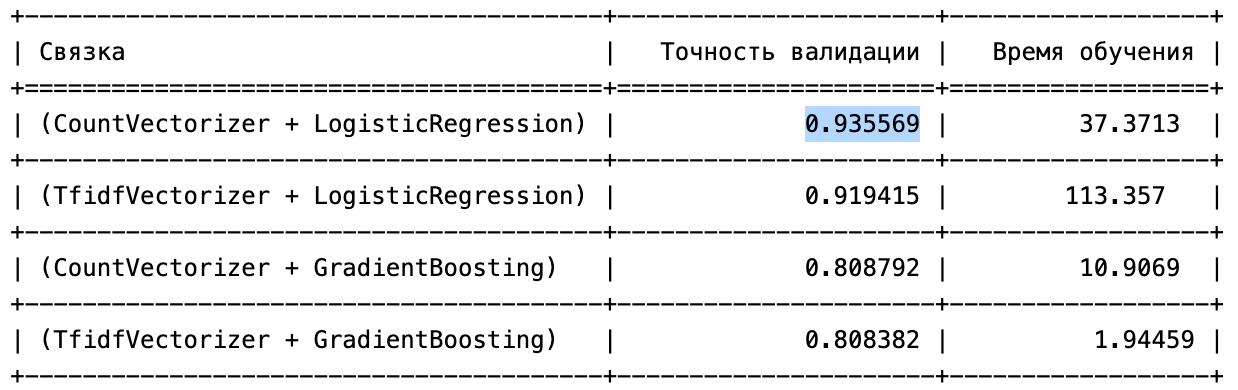
Обучения для TfidfVectorizer:



Вывод результатов:



**Сравнение и анализ результатов обучения:**

****

Наибольшую точность показал классификатор LogisticRegression в месте с векторизацией признаков CountVectorizer, а именно 93,5% на валидационной выборке. При этом время выполнения находится в удовлетворимом интервале, в отличие от связки TfidVectorizer с LogisticRegression, где обучение длилось значительно дольше остальных.

**В результате сравнения можно сказать,** что, независимо от варианта векторизации, качество обучения классификатора Логической Регрессии выше, чем Градиентного Бустинга.