

Московский Государственный Технический Университет имени Н.Э. Баумана

Факультет Информатика и системы управления

**Кафедра ИУ-5 «Системы обработки информации и управления»**

**Отчёт по рубежному контролю № 2 ：**

# Методы обработки текстов

# По дисциплине

**«**Методы Машинного Обучения**»**

Выполнила студентка Хэ Синьчэнь

Группа ИУ5И-24М

**Москва 2024г**

# 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:

# ИУ5-24М, ИУ5И-24М GradientBoostingClassifier LogisticRegression

1. Подготовка данных

Загрузите и прочитайте набор данных о рецензиях на фильмы IMDb.

Преобразуйте данные в формат, подходящий для обучения и тестирования моделей машинного обучения.

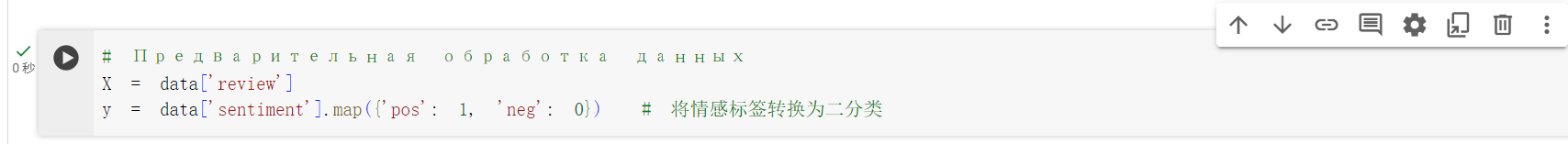




2.Предварительная обработка данных

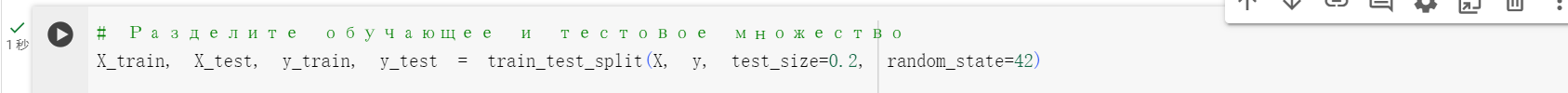
Текст рецензии на фильм был извлечен в качестве признаков (X).

Метки настроения (положительные или отрицательные) были преобразованы в бинарные метки (1 для положительных рецензий и 0 для отрицательных).



3. Разделение набора данных

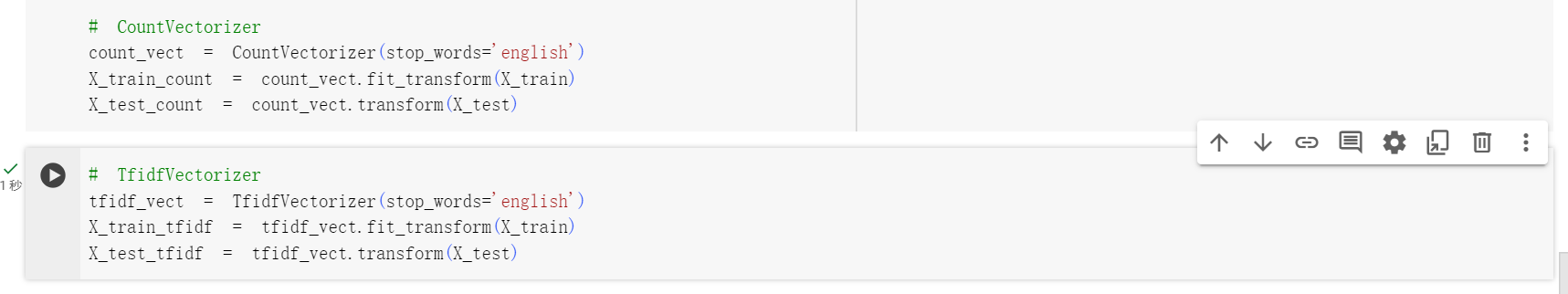
Набор данных делится на обучающий и тестовый, причем тестовый набор составляет 20%.



4. Извлечение признаков

Извлеките текстовые признаки с помощью векторизатора CountVectorizer, который вычисляет частоту встречаемости каждого слова в тексте.

Извлечение признаков текста с помощью TfidfVectorizer, который вычисляет частоту слов и обратную частоту документов (TF-IDF) - компромисс между частотой слова и частотой его появления в документе.



5. обучение и оценка моделей

Созданы две модели классификации: LinearSVC и LogisticRegression.

Обучите модели на основе признаков CountVectorizer и TfidfVectorizer, соответственно.

Оцените производительность каждой модели, используя отчеты о точности и классификации (включая точность, отзыв и F1).

