Министерство науки и высшего образования Российской Федерации

федеральное государственное автономное образовательное учреждение высшего образования

«Национальный исследовательский университет ИТМО»

Факультет инфокоммуникационных технологий

**ПРАКТИЧЕСКАЯ РАБОТА №4**

**Выполнил**:

студент 4 курса ИМРиП

Группы D34101

Ф.И.О. Лю Ифэн

Лю Юе

**Руководитель**:

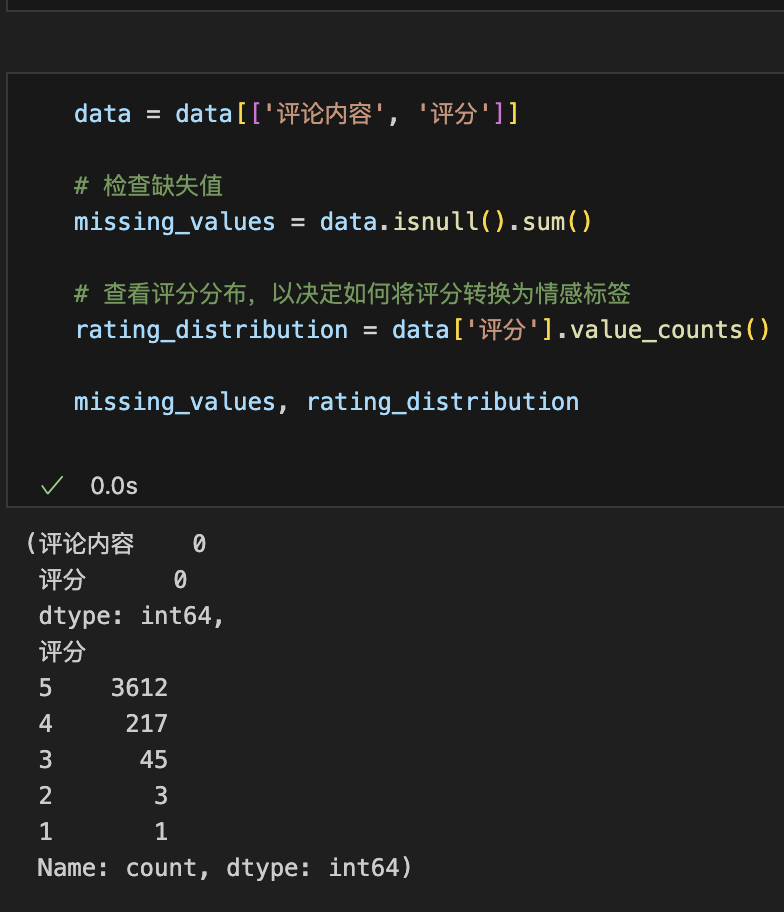
Чернышева Анастасия Вадимовна

Санкт-Петербург

2024

Сначала я выполняю необходимую предварительную обработку данных.

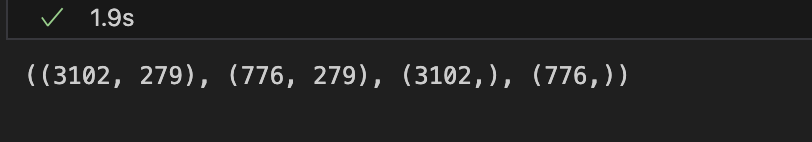
Мы удаляем ненужные столбцы, и нам достаточно просмотреть контент и рейтинги только в качестве образцов.



Результаты очистки данных показывают, что в нашем наборе данных нет пропущенных значений, поэтому мы можем сразу перейти к следующему этапу обработки. Содержание обзора и рейтинги завершены.

Очки в основном сосредоточены на 5 очках, за которыми следуют 4 очка, а количество 3, 2 и 1 очка относительно невелико. На основе этого распределения мы можем преобразовать рейтинги в две категории меток эмоций: положительные эмоции (4 балла и 5 баллов) и отрицательные эмоции (1 балл, 2 балла и 3 балла). Такое разделение может помочь нам упростить задачу классификации настроений.

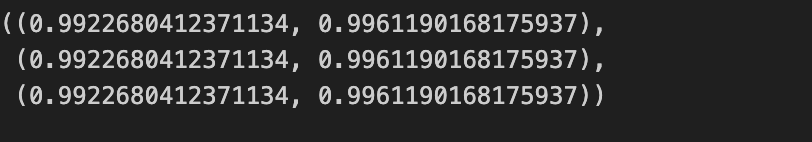
Мы предварительно обработали и векторизовали текст, а затем разделили данные на обучающую и тестовую выборку.



Данные были успешно предварительно обработаны и векторизованы, и мы получили 279 признаков. Набор данных разделен на обучающий набор и тестовый набор, где обучающий набор содержит 3102 выборки, а тестовый набор — 776 выборок.

Изначально набор данных содержал 3878 образцов (т.е. комментариев). После разделения в соотношении примерно 80-20 обучающий набор содержит 3102 выборки (приблизительно 80%), а тестовый набор содержит 776 выборок (приблизительно 20%). Этот метод сегментации гарантирует наличие достаточного количества данных для обучения модели.

Мы использовали несколько различных моделей машинного обучения (логистическую регрессию, машину опорных векторов (SVM) и модели случайного леса) для обучения и оценки нашей задачи классификации настроений.



После обучения и тестирования трех разных моделей мы получили следующие результаты:

Логистическая регрессия: точность 99,23%, показатель F1 99,61%.

Машина опорных векторов (SVM): точность 99,23 %, показатель F1 99,61 %.

Случайный лес: точность 99,23%, показатель F1 99,61%.

Производительность этих моделей на тестовом наборе очень близка, демонстрируя чрезвычайно высокую точность и оценки F1. Эти проблемы могут возникнуть, потому что наш набор данных относительно прост.В следующей задаче я буду использовать набор данных одинакового размера для обучения.

**Результаты эксперимента**

Основываясь на результатах экспериментов, мы можем сделать следующие выводы:

Используя TF-IDF для векторизации текста в сочетании с логистической регрессией, машиной опорных векторов и моделями случайного леса, можно эффективно выполнять классификацию отзывов по настроению.

Все модели продемонстрировали чрезвычайно высокую точность и баллы F1, что указывает на то, что модели смогли очень хорошо различать положительные и отрицательные отзывы.

Следующие шаги

Для дальнейшего улучшения производительности модели я буду использовать следующие методы.

Попробуйте более сложные методы векторизации текста, такие как Word2Vec.

Изучите более сложные модели глубокого обучения, такие как сверточные нейронные сети (CNN) или рекуррентные нейронные сети (RNN), особенно для обработки последовательных данных, таких как текст.

Выполните настройку гиперпараметров для оптимизации производительности модели.