Московский государственный технический университет им. Н.Э. Баумана Кафедра «Системы обработки информации и управления»



Рубежный контроль №2 по дисциплине

«Методы машинного обучения»

на тему

**«Методы обработки текстов.»**

Выполнил:

студент группы ИУ5и-22М Лун Сыхань

Москва — 2024 г.

[Оглавление «Методы обработки текстов.» 1](#_Toc2952)

[Варианты заданий 3](#_Toc2953)

[Текстовое описание набора данных: 3](#_Toc2954)

[Предварительная обработка данных и извлечение признаков 4](#_Toc2955)

[Обучение и оценка модели 5](#_Toc2956)

[случайный классификатор леса 5](#_Toc2957)

[классификатор логистической регрессии 5](#_Toc2958)

[Распечатать результаты 6](#_Toc2959)

[Вывод： 6](#_Toc2960)

# Варианты заданий

Решайте проблемы классификации текста с любым набором данных по вашему выбору. Классификация может быть бинарной или многоуровневой. Целевые объекты в выбранном вами наборе данных могут иметь любое физическое значение; одним из примеров является задача анализа тональности текста.

Необходимо сгенерировать два варианта векторизации признаков — на основе CountVectorizer и на основе TfidfVectorizer.

В качестве классификатора вы должны использовать два классификатора в зависимости от опций вашей группы:

|  |  |  |
| --- | --- | --- |
| **Группа** | **Классификатор №1** | **Классификатор №2** |
| ИУ5И-22М | [RandomForestClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html) | [LogisticRegression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html) |

# Текстовое описание набора данных:

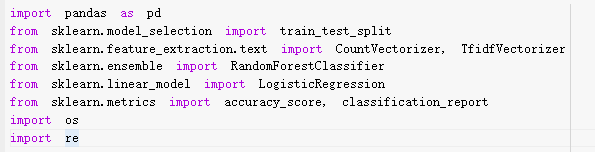
Было решено использовать набор данных для анализа настроения кинорецензий IMDB, широко применяемый на Kaggle. Этот набор данных содержит тексты рецензий на фильмы и соответствующие им метки настроения (положительные или отрицательные). Текстовые данные в наборе данных были разделены на обучающий и тестовый наборы.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

# Предварительная обработка данных и извлечение признаков

Сначала были импортированы необходимые библиотеки и модули, включая экстракторы текстовых объектов (CountVectorizer и TfidfVectorizer), два классификатора (случайный лес и логистическая регрессия) и индикаторы оценки (точность). Далее, путем инициализации объектов CountVectorizer и TfidfVectorizer, текстовые данные преобразуются в матрицу частот слов (CountVectorizer) и матрицу TF-IDF (TfidfVectorizer). Затем эти матрицы функций используются для извлечения функций из текстовых данных обучающего набора и тестового набора. Наконец, целевые функции (метки категорий) обучающего набора и тестового набора извлекаются для обучения и оценки модели.



文本

描述已自动生成 文本

描述已自动生成

# Обучение и оценка модели.

### ⚫ случайный классификатор леса

Классификатор случайного леса — это алгоритм машинного обучения, основанный на деревьях решений. Он состоит из нескольких деревьев решений и делает прогнозы путем объединения этих деревьев.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

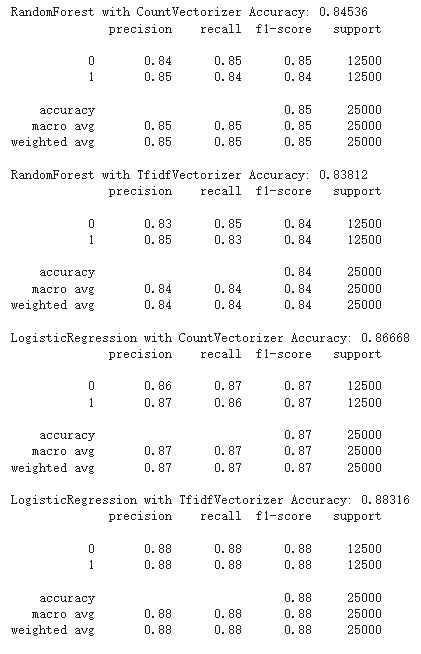
## ⚫ классификатор логистической регрессии

Логистическая регрессия использует логистическую функцию (также называемую сигмовидной функцией) для преобразования линейной комбинации признаков в значение вероятности, которое представляет вероятность принадлежности выборки к определенной категории.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

# Распечатать результаты



# Вывод：

Эти результаты показывают точность использования различных представлений признаков (CountVectorizer и TfidfVectorizer) и различных классификаторов (Random Forest и Logistic Regression) на данном наборе данных. В частности, точность классификатора случайного леса с использованием CountVectorizer и TfidfVectorizer составляет 0,845 и 0,838 соответственно, а точность классификатора логистической регрессии с использованием CountVectorizer и TfidfVectorizer - 0,867 и 0,883 соответственно, что несколько выше, чем точность классификатора логистической регрессии с использованием TfidfVectorizer для классификатора случайного леса. Feature: показывает точность 0,883, что является самым высоким показателем среди всех моделей. Это говорит о том, что модель логистической регрессии превзошла модель случайного леса в этом наборе данных, а представление признаков с помощью TfidfVectorizer достигло наивысшей точности среди всех моделей, возможно, потому, что оно лучше представляет признаки в текстовых данных. Эти результаты подчеркивают влияние выбора подходящего представления признаков на производительность модели и тот факт, что при обработке текстовых данных крайне важно учитывать особенности данных.