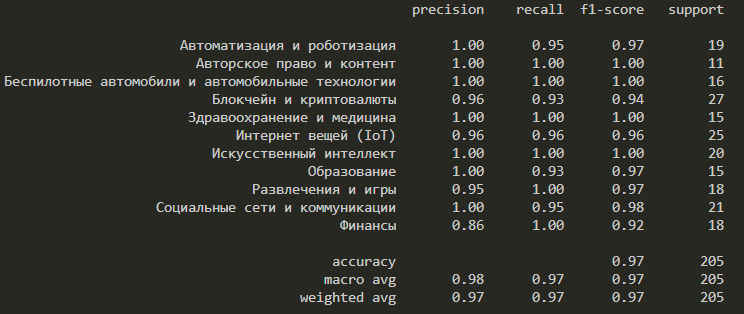
Пайплайн (pipeline) в scikit-learn представляет собой последовательность шагов обработки данных, которые выполняются последовательно, один за другим. Каждый шаг пайплайна является отдельным этапом обработки данных, который может включать в себя преобразование данных или модель машинного обучения.

В данном случае пайплайн состоит из двух шагов:

1. **TfidfVectorizer()**: Этот компонент представляет собой преобразователь данных, который преобразует коллекцию текстовых документов в матрицу TF-IDF (Term Frequency-Inverse Document Frequency). Этот метод вычисляет важность каждого слова в документе относительно всего корпуса документов. Он используется для преобразования текстовых данных в числовой формат, который может быть использован для обучения модели машинного обучения.
2. **MultinomialNB()**: Этот компонент представляет собой модель машинного обучения - наивный байесовский классификатор с мультиномиальным распределением. Он используется для классификации текстовых данных на основе частот слов в документах. Как правило, этот метод хорошо работает с данными, представленными в виде частот слов или TF-IDF.

Таким образом, в пайплайне сначала выполняется преобразование текстовых данных в числовой формат с помощью **TfidfVectorizer()**, а затем эти числовые данные подаются на вход модели классификатора **MultinomialNB()** для обучения и предсказания.

Пайплайн составлен именно так, потому что TF-IDF преобразование хорошо подходит для работы с текстовыми данными, а наивный байесовский классификатор является одним из наиболее распространенных методов классификации текста, который хорошо работает с большими объемами текстовых данных. Кроме того, использование пайплайна позволяет соблюдать стандартный интерфейс scikit-learn и облегчает работу с моделью.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| Автоматизация и роботизация | 1.00 | 0.95 | 0.97 | 19 |
| Авторское право и контент | 1.00 | 1.00 | 1.00 | 11 |
| Беспилотные автомобили и автомобильные технологии | 1.00 | 1.00 | 1.00 | 16 |
| Блокчейн и криптовалюты | 0.96 | 0.93 | 0.94 | 27 |
| Здравоохранение и медицина | 1.00 | 1.00 | 1.00 | 15 |
| Интернет вещей (IoT) | 0.96 | 0.96 | 0.96 | 25 |
| Искусственный интеллект | 1.00 | 1.00 | 1.00 | 20 |
| Образование | 1.00 | 0.93 | 0.97 | 15 |
| Развлечения и игры | 0.95 | 1.00 | 0.97 | 18 |
| Социальные сети и коммуникации | 1.00 | 0.95 | 0.98 | 21 |
| Финансы | 0.86 | 1.00 | 0.92 | 18 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| accuracy |  |  | 0.97 | 205 |
| macro avg | 0.98 | 0.97 | 0.97 | 205 |
| weighted avg | 0.97 | 0.97 | 0.97 | 205 |