**Описание выборки:**

**Для исследования используется выборка текстов с сайта elibrary.ru, состоящая из признаков текстов различной тематики. В выборке содержится 3419 элементов, каждый из которых имеет метку класса «ИАД» или «не ИАД». Из основного имеем название, аннотацию, год выпуска, автора и ключевые слова. Мы же будем работать с названиями и библиографическим описанием(аннотацией).**

**Библиографическое описание(БО) представляет собой очень краткий пересказ общего текста, отсюда можно понять тематику и примерное содержание теста, имеет довольно небольшой размер представленный в одном абзаце. Надо дать внятное определение – оно стандартное – см.Интернет**

**Названия соответственно представляют собой названия текстов, дают понимание тематики и имеют малую размерность.**

**Предварительно можно дать анализ сбалансированности выборки:**

# Проверяем сбалансированность выборки

print("Количество текстов по теме ИАД", labels.count(True))

print("Количество текстов по теме не ИАД", labels.count(False))

Количество текстов по теме ИАД 1583

Количество текстов по теме не ИАД 1836

Отсюда можно сказать, что классы сбалансированы

**Надо привести пример документа.**

**Стандартная процедура предварительной обработки текста**

Метод, известный как предварительная обработка текста, используется для очистки текстовых данных перед их передачей в Machine Learning. Текстовые данные содержат различный «шум» — например, эмодзи, знаки препинания, чередование строчных и прописных символов. Это только начало трудностей, с которыми мы столкнемся. Машины не понимают слов, им нужны числа, поэтому нужно найти быстрый и эффективный способ преобразования текста в числа. Бакалаврскую работу будут проверять на Антиплагиате – надо стараться писать собственный текст (и делать это не в разговорном стиле).

Следующие шаги выполняются с использованием стандартных методов предварительной обработки:

• преобразование в нижний регистр

• удаление знаков препинания

• удаление стоп-слов

• токенизация корпуса

• стемминг и лемматизация

• векторное представление слов с использованием CountVectorizer и TF-IDF

Проведем классификацию со всеми терминами и с удалением малозначимых терминов.

Для удаления терминов используем алгоритм «Использование частоты документа (Document Frequency thresholding)»

Под частотой документа понимают число документов из обучающей коллекции, в которых появляется данный терм. Метод заключается в следующем: для каждого терма вычисляется частота документа, а затем из вектора признаков удаляются те термы, частота документов для которых ниже некоторого, заранее заданного порогового значения. Этот подход, основан на том предположении, что редко встречающиеся в документах термы не могут оказать влияние на результат классификации.

Нужны формулы TF-IDF, что использовал стемминг или лемматизацию – не ясно.

Словарь терминов, составленный по названиям, имел размерность 25 612, после сокращения 2443.

Словарь терминов, составленный по БО, имел размерность 170 666, после сокращения 15082.

3

**Проводим классификацию с помощью логистической регрессии**

Сравнение результатов классификации по названиям

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | С удалением малозначимых терминов | | | | Без удаления малозначимых терминов | | | |
| Точность | 0.8358395989974937 | | | | 0.8341687552213868 | | | |
| Матрица ошибок | [1281 18]  [ 375 720] | | | | [1281 18]  [ 379 716] | | | |
|  | precision | recall | f1-score | support | precision | recall | f1-score | support |
| False | 0.77 | 0.99 | 0.87 | 1299 | 0.77 | 0.99 | 0.87 | 1299 |
| True | 0.98 | 0.66 | 0.79 | 1095 | 0.98 | 0.65 | 0.78 | 1095 |
| accuracy |  |  | 0.84 | 2394 |  |  | 0.83 | 2394 |
| macro avg | 0.87 | 0.82 | 0.83 | 2394 | 0.87 | 0.82 | 0.82 | 2394 |
| weighted avg | 0.87 | 0.84 | 0.83 | 2394 | 0.86 | 0.83 | 0.83 | 2394 |

Сравнение результатов классификации по библиографическому описанию

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | С удалением малозначимых терминов | | | | Без удаления малозначимых терминов | | | |
| Точность | 0.8813700918964077 | | | | 0.8788638262322472 | | | |
| Матрица ошибок | [1274 25]  [ 259 836] | | | | [1275 24]  [ 266 829] | | | |
|  | precision | recall | f1-score | support | precision | recall | f1-score | support |
| False | 0.83 | 0.98 | 0.90 | 1299 | 0.83 | 0.98 | 0.90 | 1299 |
| True | 0.97 | 0.76 | 0.85 | 1095 | 0.97 | 0.76 | 0.85 | 1095 |
| accuracy |  |  | 0.88 | 2394 |  |  | 0.88 | 2394 |
| macro avg | 0.90 | 0.87 | 0.88 | 2394 | 0.90 | 0.87 | 0.87 | 2394 |
| weighted avg | 0.90 | 0.88 | 0.88 | 2394 | 0.89 | 0.88 | 0.88 | 2394 |

Надо разъяснить все меры качества, приведенные в таблицах.

Вывод:

Во всех 4-х случаях можно судить о неплохом качестве классификации. Опираясь на Accuracy и F1-метрику, можно сказать, что библиографическое описание подходит лучше, с точки зрения точности.

При удалении малозначимых терминов мы практически не теряем в качестве, сама же классификация проходит примерно в 4 раза быстрее.

Стоит заметить, что из-за большой размерности библиографического описания уменьшения размерности списка терминов занимает примерно в 5 раз больше времени, чем для списка, составленного по названиям.