**Описание выборки:**

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

**Библиографическое описание (БО) представляет собой сведения о документе для его краткого описания и быстрого поиска. Представлены в одном небольшом абзаце.**

**Названия соответственно представляют собой названия документов, 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

**Пример документа:**

**Название:**

**ПЕРСПЕКТИВЫ ВНЕДРЕНИЯ ТЕХНОЛОГИЙ DATA MINING В ТАМОЖЕННУЮ ДЕЯТЕЛЬНОСТЬ**

**ЖУРНАЛ:**

**АКАДЕМИЧЕСКИЙ ВЕСТНИК РОСТОВСКОГО ФИЛИАЛА РОССИЙСКОЙ ТАМОЖЕННОЙ АКАДЕМИИ**

**АННОТАЦИЯ:**

**В статье проведен анализ перспективных направлений внедрения технологий Data Mining в деятельность таможенных органов. Рассмотрены классификационные методы машинного обучения с учителем и без учителя, применение которых может автоматизировать решение сложных задач по отнесению поставок товаров к рисковым или выявлению потенциальных рисков. Особое внимание уделено кластерному анализу и программным платформам, которые поддерживают его реализацию.**

**Автор: Кудрявцев О. Е.**

**Стандартная процедура предварительной обработки текста**

Предварительная обработка текста, используется для очистки текстовых данных. Текстовые данные содержат различный «шум» — например, эмодзи, знаки препинания, чередование строчных и прописных символов. Машины не понимают текст, поэтому после предварительной обработки текстовые данные представляются в виде чисел, для этого используется матрица документ-термин или Tf-idf.

В работе предварительная обработка состоит из следующих шагов:

• преобразование в нижний регистр

• удаление знаков препинания

• удаление стоп-слов

• токенизация (разделение на слова и словосочетания)

• лемматизация

• векторное представление слов с использованием CountVectorizer и TfidfVectorizer (в ходе исследования было принято решение использовать TF-IDF)

,

где в числителе количество отдельных слов в документе, а в знаменателе общее количество слов в документе.

,

где в числителе количество документов, а в знаменателе количество документов, в которых содержится интересующее слово.

**Сокращение размерности словаря 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.

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**

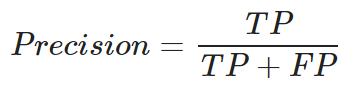
Сравнение результатов классификации по названиям

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 |

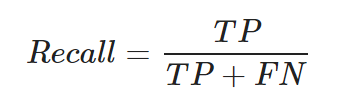
Сравнение результатов классификации по библиографическому описанию

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 |

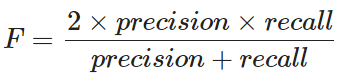
precision (точность) – доля правильных ответов.



recall (полнота) – доля истинно положительных классификаций.



f1-score (F1-мера) представляет собой гармоническое среднее между точностью и полнотой.

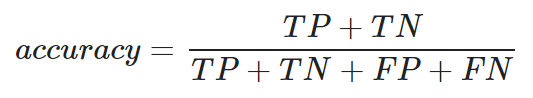


support – общее количество классификаций.

macro avg – среднее арифметическое

weighted avg – среднее взвешенное

accuracy – доля правильных ответов алгоритма.



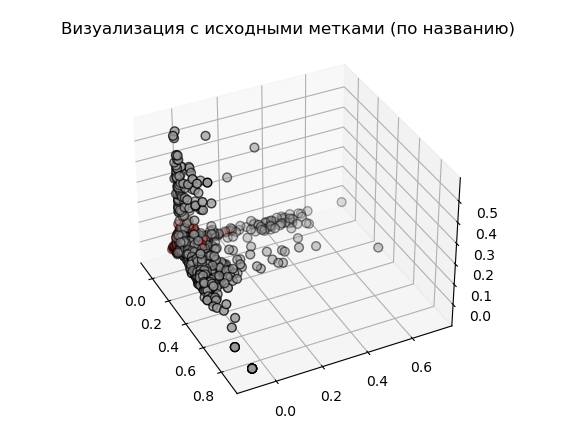
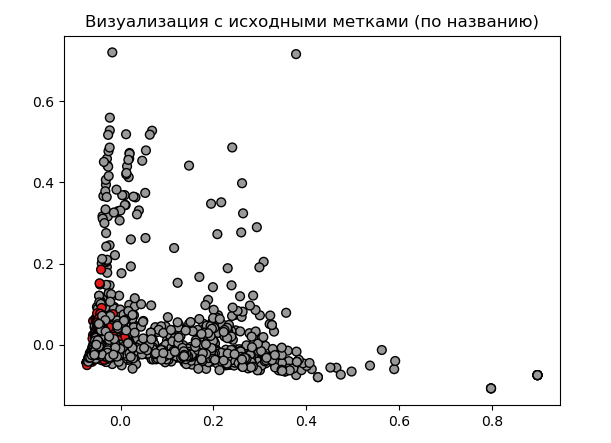
Вывод:

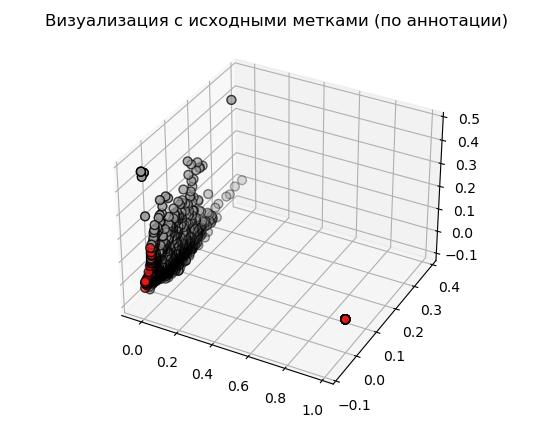
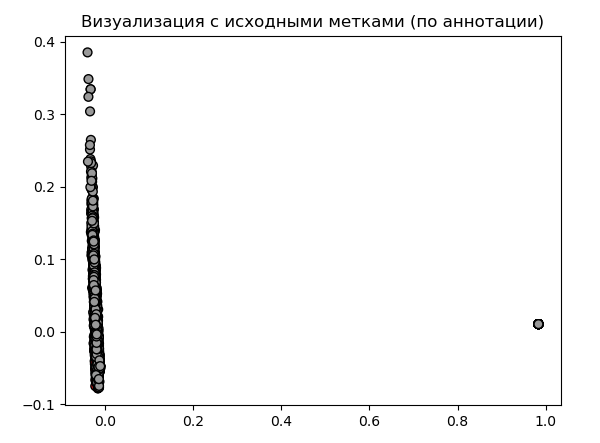
Во всех 4-х случаях можно судить о неплохом качестве классификации. Опираясь на Accuracy и F1-метрику, можно сказать, что библиографическое описание подходит лучше, с точки зрения точности.

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

**Сравнение различных методов классификации.**

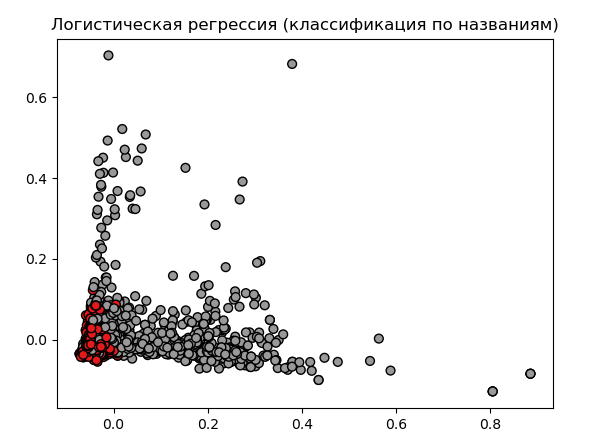
Визуализация с исходными метками

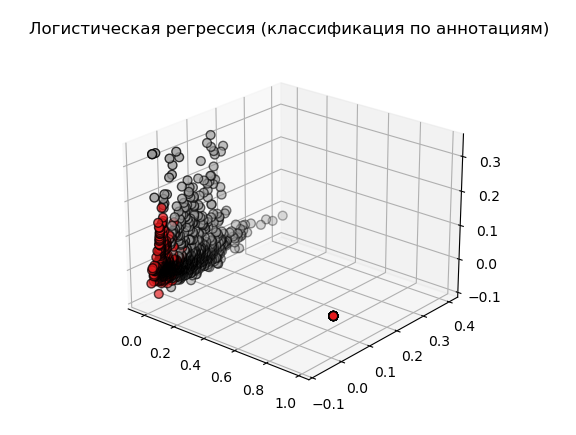
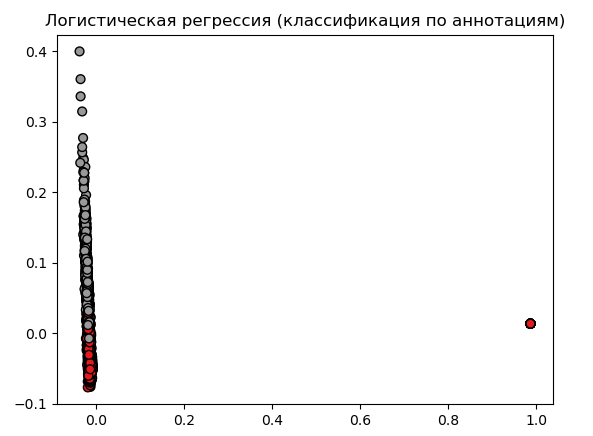
****

****

Логистическая регрессия.

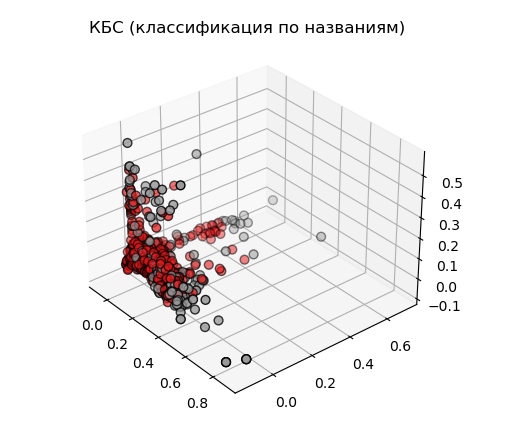
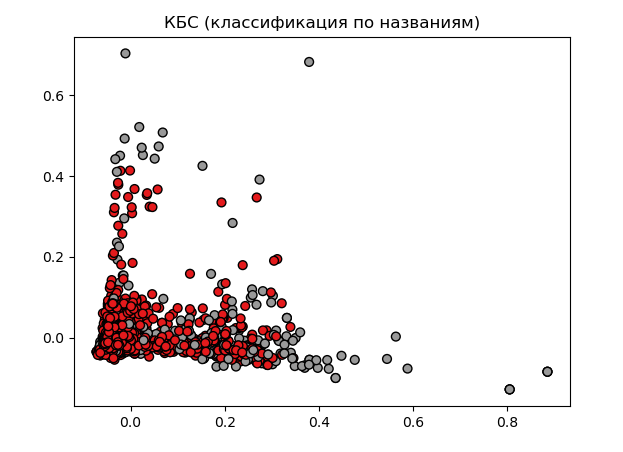
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | По названиям | | | | По библиографическому описанию | | | |
| Точность | 0.8358395989974937 | | | | 0.8813700918964077 | | | |
| Матрица ошибок | [1281 18]  [ 375 720] | | | | [1274 25]  [ 259 836] | | | |
|  | precision | recall | f1-score | support | precision | recall | f1-score | support |
| False | 0.77 | 0.83 | 0.83 | 0.83 | 0.83 | 0.99 | 0.87 | 1299 |
| True | 0.98 | 0.97 | 0.97 | 0.97 | 0.97 | 0.65 | 0.78 | 1095 |
| accuracy |  |  |  |  |  |  | 0.83 | 2394 |
| macro avg | 0.87 | 0.90 | 0.90 | 0.90 | 0.90 | 0.82 | 0.82 | 2394 |
| weighted avg | 0.87 | 0.90 | 0.90 | 0.90 | 0.90 | 0.83 | 0.83 | 2394 |

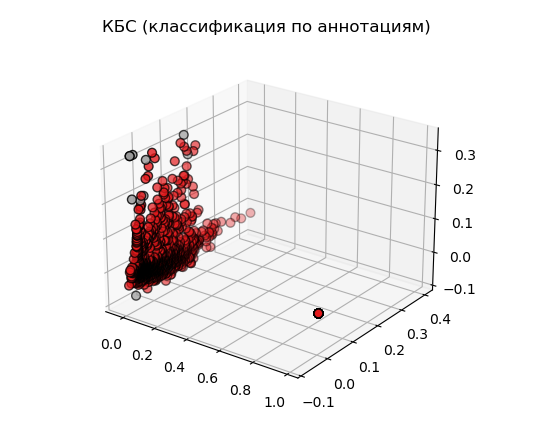
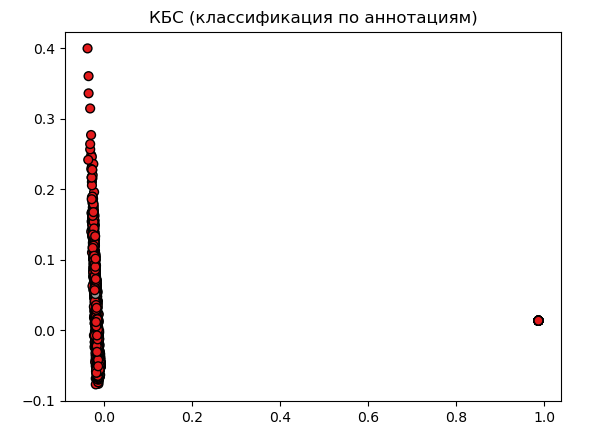




К-ближайших соседей.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | По названиям | | | | По библиографическому описанию | | | |
| Точность | 0.633249791144528 | | | | 0.551796157059315 | | | |
| Матрица ошибок | [1292 7]  [ 871 224] | | | | [1299 0]  [1073 22] | | | |
|  | precision | recall | f1-score | support | precision | recall | f1-score | support |
| False | 0.60 | 0.99 | 0.75 | 1299 | 0.55 | 1.00 | 0.71 | 1299 |
| True | 0.97 | 0.20 | 0.34 | 1095 | 1.00 | 0.02 | 0.04 | 1095 |
| accuracy |  |  | 0.63 | 2394 |  |  | 0.55 | 2394 |
| macro avg | 0.78 | 0.60 | 0.54 | 2394 | 0.77 | 0.51 | 0.37 | 2394 |
| weighted avg | 0.77 | 0.63 | 0.56 | 2394 | 0.75 | 0.55 | 0.40 | 2394 |





Деревья решений.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | По названиям | | | | По библиографическому описанию | | | |
| Точность | 0.8149540517961571 | | | | 0.7581453634085213 | | | |
| Матрица ошибок | [1135 164]  [ 279 816] | | | | [1056 243]  [ 336 759] | | | |
|  | precision | recall | f1-score | support | precision | recall | f1-score | support |
| False | 0.80 | 0.87 | 0.84 | 1299 | 0.76 | 0.81 | 0.78 | 1299 |
| True | 0.83 | 0.75 | 0.79 | 1095 | 0.76 | 0.69 | 0.72 | 1095 |
| accuracy |  |  | 0.81 | 2394 |  |  | 0.76 | 2394 |
| macro avg | 0.82 | 0.81 | 0.81 | 2394 | 0.76 | 0.75 | 0.75 | 2394 |
| weighted avg | 0.82 | 0.81 | 0.81 | 2394 | 0.76 | 0.76 | 0.76 | 2394 |

Случайный лес.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | По названиям | | | | По библиографическому описанию | | | |
| Точность |  | | | |  | | | |
| Матрица ошибок |  | | | |  | | | |
|  | precision | recall | f1-score | support | precision | recall | f1-score | support |
| False |  |  |  |  |  |  |  |  |
| True |  |  |  |  |  |  |  |  |
| accuracy |  |  |  |  |  |  |  |  |
| macro avg |  |  |  |  |  |  |  |  |
| weighted avg |  |  |  |  |  |  |  |  |