

**ЗАДАНИЕ ДЕМОНСТРАЦИОННОГО ЭКЗАМЕНА**

|  |  |
| --- | --- |
| **Код комплекта оценочной документации** | КОД 1.2-2023-2025 |
| **Год действия задания** | 2023 |
| **Номер варианта задания** | 2 |

**ОПИСАНИЕ ЗАДАНИЯ**

**Описание модуля 1: «****Парсинг и предобработка данных»**

Архив содержит файлы с информацией, полученной с сайта регистрации жалоб граждан. Один файл соответствует одной жалобе. Необходимо автоматизировать маршрутизацию обращений и жалоб граждан, поступающих с сайта «умного города». С точки зрения постановки задачи в терминах машинного обучения, – решить задачу классификации текстов.

В этом модуле необходимо сформировать набор данных, содержащий данные по всем жалобам.

**При выполнении модуля 1 ставятся следующие цели:**

1. Подготовка набора данных для дальнейшего построения прогнозирующей модели;
2. Проведение разведочного анализа для выявления закономерностей и факторов, влияющих на целевую переменную.

**При выполнении данного модуля 1 ставятся следующие задачи:**

1. Выполнить парсинг данных для формирования единого набора данных;
2. Формирование структуры набора данных;
3. Провести предварительную обработку данных;
4. Выполнить построение и отбор признаков;
5. Выполнить кластеризацию данных

**Требования к оформлению письменных материалов**

Письменный материал отсутствует.

**Представление результатов работы**

Результат выполнения Модуля 1«Парсинг и предобработка данных»: предобработанные данные (архив Data.zip), отчет о проделанной работе (Report\_C1.html, Report\_C1.ipynb), дополнительные комментарии коду (Readme.txt).

**Необходимые приложения**

Необходимые приложения смотреть в папке «КОД 1.2 Приложения к вариантам».

Приложение 1: Архив, содержащий информацию по каждой жалобе граждан (Data.zip)

**ЗАДАНИЕ**

|  |
| --- |
| 1.1 Парсинг данных |
| Архив содержит в себе файлы с информацией, полученной с сайта регистрации жалоб граждан. Один файл соответствует одной жалобе. Необходимо проанализировать данные каждого файла и определить итоговый набор полей для создания структуры данных, в которую преобразованные записи будут импортированы.  Необходимо создать структуру данных, которая позволит сохранить записи, состоящие из итогового набора полей. |

|  |
| --- |
| 1.2 Формирование структуры набора данных |
| Задача заключается в определении класса(кластера) – категории обращения граждан. Исходя из этого, необходимо определить, какие атрибуты имеют наибольшее влияние на определение классов(кластеров), оставить только их для последующего обучения. Также необходимо обосновать выбор дополнительных атрибутов и причину исключения каких-либо данных из исходного набора документов. |

|  |
| --- |
| 1.3 Предварительная обработка текстовых данных |
| Проведите предварительную обработку текста с помощью методов NLP: токенизацию, лемматизацию, выделение значимых частей речи, а также удаление стоп-слов, пунктуации, спецсимволов. Обоснуйте выбор методов предварительной обработки данных. |

|  |
| --- |
| 1.4 Поиск ключевых слов/n-грамм. Векторизация текстов |
| Выполните поиск ключевых слов/биграмм/триграмм в тексте различными способами. Обоснуйте выбор алгоритмов поиска ключевых слов/биграмм/триграмм. Добавьте ключевые слова/биграммы/триграммы, как новые признаки в набор данных.  Преобразовать документы в векторные представления, к которым можно применить численное машинное обучение. |

|  |
| --- |
| 1.5 Тематическое моделирование |
| Выполните тематическое моделирование различными способами (не менее трех) и визуализируйте его результаты. Обоснуйте выбор алгоритмов тематического моделирования. |

|  |
| --- |
| 1.6 Кластеризация |
| Выполнить кластеризацию текстов по сходству обращений несколькими способами (не менее трех). Выберите метрику оценки качества кластеризации. Обоснуйте выбор методов и приемов. Выполните визуальный анализ кластерных структур и оценки качества кластеризации. Определите лучший алгоритм кластеризации на основе выбранной метрики. |

|  |
| --- |
| 1.7 Подготовка отчета |
| Подготовьте отчет о проделанной работе по итогам сессии, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. Результаты работы должны состоять из отчетов в формате .html и исходников с возможностью перекомпиляции. Архив Data.zip должен содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. В файле Readme.txt необходимо описать содержимое результирующих файлов архива Data.zip. |

**Описание модуля 2:** **«Построение, обучение и оптимизация модели»**

В этом модуле продолжается работа с данными, подготовленными в предыдущей сессии. Требуется осуществить выбор алгоритма классификации, построить модель и провести оптимизацию полученной модели машинного обучения в контексте исследуемой задачи.

**При выполнении модуля 2 ставятся следующие цели:**

Построение классификатора обращений граждан по категориям обращений.

**При выполнении модуля 2 ставятся следующие задачи:**

1. Выполнить разбиение выборки на обучающую и валидационную;
2. Осуществить построение моделей классификации;
3. Оценить качество полученных моделей в соответствии со спецификой решаемой задачи;
4. Выполнить оптимизацию лучшей модели.

**Требования к оформлению письменных материалов**

Письменный материал отсутствует.

**Представление результатов работы**

Результат выполнения Модуля 2 «Построение, обучение и оптимизация модели»: результирующие файлы (архив Data.zip), отчет о проделанной работе (Report\_C1.html, Report\_C1.ipynb), дополнительные комментарии коду (Readme.txt).

**Необходимые приложения**

Необходимые приложения смотреть в папке «КОД 1.2 Приложения к вариантам».

Приложение 1: Архив, содержащий информацию по каждой жалобе граждан (Data.zip)

Приложение 2: Готовый набор данных с петициями для обучения (Petitions.csv)

**ЗАДАНИЕ**

|  |
| --- |
| 2.1 Разведочный анализ |
| Проведите анализ плотности распределения атрибутов и целевой переменной набора данных. Дайте интерпретацию полученных результатов.  Выполнить визуализацию пространства текстовых признаков различными способами. Визуализация должна отражать зависимости категории обращения от временных признаков, ключевых слов/n-грамм, а также как минимум одну – на Ваш выбор. |

|  |
| --- |
| 2.1 Разбиение выборки |
| Выполните разбиение полученной (либо представленной Petitions.csv) выборки на обучающую и тестовую. Проведите обучение моделей, основанных на различных алгоритмах. Сделайте предсказание категории обращения на тестовой выборке. Выполните оценку моделей разной степени сложности в соответствии с выбранной метрикой. Определите модель, показавшую лучшее качество. |

|  |
| --- |
| 2.2 Оптимизация модели |
| Выполните настройку полученной модели уменьшив вычислительную сложность модели (выбор значимых признаков, понижение размерности). Оцените качество полученной модели, сделайте вывод.  Необходимо оптимизировать полученную модель под решаемую задачу, настраивая гиперпараметры. Выполните оценку данной модели после настройки гиперпараметров.  Построить для данной модели кривые валидации и обучения, интерпретируйте полученные результаты. |

|  |
| --- |
| 2.3 Подготовка отчета |
| Подготовьте отчет о проделанной работе по итогам сессии, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. Результаты работы должны состоять из отчетов в формате .html и исходников с возможностью перекомпиляции. Архив Data.zip должен содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. В файле Readme.txt необходимо описать содержимое результирующих файлов архива Data.zip. |