

**ЗАДАНИЕ ДЕМОНСТРАЦИОННОГО ЭКЗАМЕНА**

| **Код комплекта оценочной документации** | КОД 1.2-2023-2025 |
| --- | --- |
| **Год действия задания** | 2023 |
| **Номер варианта задания** | 1 |

**ОПИСАНИЕ ЗАДАНИЯ**

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

Ежегодно в России проводится Национальная премия за вклад в развитие российского сегмента сети Интернет. Каждый раз экспертное жюри осматривает вручную все заявки и выбирает из них самых достойных, а экспертный совет путем голосования выбирает в каждой номинации победителя. Премия Рунета является желаемой наградой не только для бизнеса, но и для представителей госорганов и ведомств. Данная премия является престижнейшей в Рунете и самой значимой, ежегодно собирая более 10 000 зрителей.

Задача: на основе аналитических статей, связанных с информационными технологиями, бизнесом и интернетом с Хабра (системы тематических коллективных блогов с элементами новостного сайта) обобщить информацию по публичной активности каждой организации-номинанта и разработать алгоритм, способный самостоятельно определять номинацию компании.

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

1. Подготовка набора данных для дальнейшего построения прогнозирующей модели.

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

1. Выполнить парсинг данных для сбора информации о компаниях с указанных веб-ресурсов;
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 статьи о компаниях (Data.zip)

Приложение 2: Список номинантов конкурса (Condidates.doc)

**ЗАДАНИЕ**

| 1.1 Парсинг данных |
| --- |
| На основе имеющихся данных из архива и аналитических статей, связанных с информационными технологиями, бизнесом и интернетом с сайта (https://habr.com/ru/search/) необходимо построить исходный набор данных (.csv или .xml). Набор данных должен включать названия, описание, рейтинг и сферу деятельности компаний, дату публикации, а также текст статей из Интернет-ресурсов. Подготовленный набор данных должен содержать сведения о всех номинантах конкурса. Разработанный парсер должен извлекать гиперссылки из начальной страницы с последующим обходом всех страниц по полученным ссылкам и извлечением их содержимого.  Можно дополнить набор какими-либо другими данными, если они могут быть полезны для дальнейшего исследования. |

| 1.2 Формирование структуры набора данных |
| --- |
| Задача заключается в определении класса(кластера) – номинации премии Рунета. 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 данных по сходству компаний 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 номинаций компаний-конкурсантов Премии Рунета.

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

1. Провести разведочный (визуальный) анализ данных;
2. Выполнить разбиение выборки на обучающую и валидационную;
3. Осуществить построение моделей классификации;
4. Оценить качество полученных моделей в соответствии со спецификой решаемой задачи;
5. 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 статьи о компаниях (Data.zip)

Приложение 2: Список номинантов конкурса (Condidates.doc)

Приложение 3: Информация по номинациям для представленных организаций (Target.json)

**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 темы от временных признаков, рейтинга и ключевых слов/n-грамм. Также провести визуальный анализ статистики публикаций. |

| 2.2 Разбиение выборки |
| --- |
| Выполните разбиение полученной выборки на обучающую и тестовую. Проведите обучение моделей, основанных на различных алгоритмах. Сделайте предсказание номинации организации на тестовой выборке. Выполните оценку моделей разной степени сложности в соответствии с выбранной метрикой. Определите модель, показавшую лучшее качество. |

| 2.3 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. |

| 2.4 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. |