**ЗАДАНИЕ Модуль А**

|  |
| --- |
| 1.1 Парсинг данных |
| На основе имеющихся аналитических статей, связанных с информационными технологиями, бизнесом и интернетом 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.  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 Подготовка отчета |
| Подготовьте отчет о проделанной работе, 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. |