**ЗАДАНИЕ Модуль Б**

|  |
| --- |
| 2.1 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. |

|  |
| --- |
| 2.2 Тематическое моделирование |
| Выполните тематическое моделирование различными способами (не менее трех) и визуализируйте его результаты. Обоснуйте выбор алгоритмов тематического моделирования. |

|  |
| --- |
| 2.3 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. |

|  |
| --- |
| 2.4 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. |

|  |
| --- |
| 2.5 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. |