Приложение Б

Министерство науки
и высшего образования РФ
ФГБОУ ВО «НИУ «МЭИ»
Институт дистанционного
и дополнительного образования

АННОТАЦИИ РАБОЧИХ ПРОГРАММ ДИСЦИПЛИН (МОДУЛЕЙ) ДОПОЛНИТЕЛЬНОЙ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

повышения квалификации
«Основы анализа текстовых данных»,

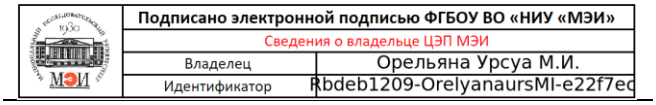
Раздел(предмет) ***Основы анализа текстовых данных***

Наименование дисциплин (модулей)	Содержание дисциплин (модулей)	Форма ТК	Количество часов
Постановка задачи машинного обучения	Типы задач, решаемых методами машинного обучения (МО). Формальная постановка задачи машинного обучения. Задачи анализа тестовой информации, решаемые методами МО.Критерии качества: аккуратность, полнота, точность, F-мера, площадь под кривой ошибок, матрица неточностей. Способы формирование выборок. Обучающая, экзаменационная и тестовая выборки. Обучение моделей. Явление переобучения модели. Несбалансированные классы. Методы борьбы с несбалансированностью: oversampling, undersampling. Специальные стратегии сэмплинга в условиях несбалансированных классов	*Нет*	*70*
Особенности и задачи обработки	Text Mining и особенности задач, связанных с анализом текстов. Онтологии и	*Тестирование*	

Наименование дисциплин (модулей)	Содержание дисциплин (модулей)	Форма ТК	Количество часов
текстовых данных	тезаурусы. Статистический подход к анализу текстовой информации. Проблема снижения размерности в задачах Text Mining. Предварительная обработка данных: стемминг, лемматизация. Слова, не несущие информации. Выявление информативных признаков. Взвешивание как способ выявления информативных терминов. Статистический подход к выявлению информативных терминов. Теоретико-информационный подход к выявлению информативных терминов. Модель представления текстовых данных в математическом виде. Модель «Мешок слов» (Bag of words). Частично- и полностью структурированные модели.		
Методы классификации данных	Систематизация и обзор методов классификации. Метод ближайших соседей. Метод деревьев решений. Метод опорных векторов. Метод логистической регрессии. Наивный байесовский метод. Профильные методы классификации. Ансамблевые методы классификации. Оценка разнородности методов классификации. Бустинг, бэггинг. Метод случайного леса деревьев решений.	*Нет*	
Кластеризация данных, выявление дубликатов текстовых	Постановка задачи кластеризации, ее особенности. Иерархическая кластеризация. EM-алгоритм кластеризации.	*Нет*	

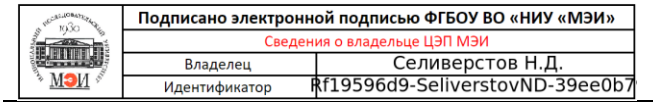
Наименование дисциплин (модулей)	Содержание дисциплин (модулей)	Форма ТК	Количество часов
документов	Семейство алгоритмов k-means, другие алгоритмы кластеризации. Виды дубликатов: полные дубликаты, явные дубликаты, нечеткие дубликаты. Коэффициент ассоциативности Жаккара. Семейство методов шинглов. Методы выявления дубликатов Winnowing, SpotSigs, I-Match, коэффициент Джаро-Винклера		

Руководитель ОДПО, ЦПП УВО



М.И. Орельяна Урсуа

Начальник ОДПО



Н.Д. Селиверстов