УДК 004.9

КЛАСТЕРИЗАЦИЯ ТЕКСТОВЫХ КОММЕНТАРИЕВ НА ОСНОВЕ МЕТОДОВ ВЕРОЯТНОСТНОГО ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Захарова А.А. (Университет ИТМО)
Научный руководитель – к.т.н. Махныткина О.В.
(Университет ИТМО)

В данной работе рассматривается корпус текстов, который используется для определения тематик на методах тематического моделирования. Для построения тематического моделирования используются такие методы как вероятностный латентный семантический анализ и латентное размещение Дирихле.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №619423 «Разработка виртуального диалогового помощника для поддержки проведения дистанционного экзамена на основе аргументационного подхода и глубокого машинного обучения».

Анализ комментариев пользователей является важной составляющей при оценке качества различных образовательных ресурсов и материалов. Использование оценок в комментариях на платформе онлайн-курсов и выявление ключевых слов в сообщениях позволит выявить отношение пользователей к различным темам курса.

Для проведения исследования был собран датасет текстов комментариев к онлайн – курсам по дисциплине «Машинное обучение». Корпус текстов содержит 1700 отзывов, которые имеют 5-ти бальную шкалу оценивания.

Существуют множество методов, которые используются для получения тематических моделей. Большинство из них принадлежат классу вероятностного моделирования. Темы представляются, как дискретные распределения на множестве слов, а тексты – как дискретное распределение на множество тем. Ведь тексты могут относиться сразу к нескольким темам, тогда моделирование осуществляет «нечеткую кластеризацию», то есть документ принадлежит нескольким темам в разных уровнях. Построение тематической модели может рассматриваться, как задача одновременной кластеризации документов и слов по одному и тому же множеству кластеров, называемых темами.

Одним из основных методов вероятностного тематического моделирования является метод вероятностный латентный семантический анализ (pLSA), он позволяет определить взаимосвязь между коллекциями документов и терминов, в них встречающимися. Основная задача метода, заключается на смешанном разложении и использовании вероятностей модели, что позволит более качественно определять возможные тематики документов. Алгоритм, выделяет следующие этапы: предобработка, нахождение весов слов любым методом, построение весовой матрицы и разложение матрицы методом сингулярного разложения. Преимущество данной модели заключается в том, что возможность нахождения вероятности отношения каждого документа к каждой из представленных тем, с последующей группировкой, что является достаточно трудоемкой задачей. К недостаткам можно отнести высокую вычислительную сложность и низкую скорость работы, а также линейную зависимость количества параметров от количества документов. Модель pLSA утверждает, что каждое слово документа происходит из случайно выбранной темы. Сами темы взяты из распределения тем по документам. Проблема в том, что для вычисления вероятностей нового документа, необходимо полностью переобучить тематическую модель для всего корпуса.

Большая часть моделей разрабатываются на основе латентного размещения Дирихле (LDA). Данная модель, позволяет уйти от недостатков первой рассмотренной модели, таких как отсутствие закономерности при генерации документов из набора полученных тем, что позволяет улучшить выборку. Подход к модели включает в себя два этапа. На первом шаге выполняется предобработка текста, которая содержит в себе графематический анализ, лемматизацию, удаление лишних элементов и создание корпуса текстов. После того, как мы получили все необходимое для обучения модели, где каждая тема будет представлять собой

комбинацию ключевых слов, и каждое слово вносит определенный вес в тему. Так же, модель способна распознавать скрытую семантическую информацию из набора документов. Она применяется в нескольких областях, таких как сегментация текста, рекомендация тегов, автоматизированная оценка эссе, идентификация темы.

На основе рассмотренных моделей были выявлены основные тематики в группах комментариев к онлайн курсам с различными оценками пользователей. Так же, был собран корпус текстов, а именно отзывов, которые оцениваются по определенной тематике.