УДК 004.62

Исследование оптимального количества тем при построении тематического профиля онлайн-сообщества

А.Д. Телевной, Университет ИТМО, Санкт-Петербург
Научный руководитель – к.ф.-м.н., доцент С.Е. Иванов
Университет ИТМО, Санкт-Петербург

Введение. Одним из вариантов кластеризации текстовых данных является использование алгоритма латентного размещения Дирихле (LDA). При использовании данного алгоритма от исследователя требуется указать определенное количество тем (кластеров), на которое будет разбиваться анализируемый корпус данных. Оптимальное количество тем, в большинстве случаев, на первоначальном этапе исследователю неизвестно, поэтому оно, как правило, задается субъективно, и в дальнейшем может изменяться, что приводит к перестроению всей тематической модели. Описанный в данной работе алгоритм определения количества тем с помощью вычисления среднего гармонического значения позволяет определить оптимальное значение количества кластеров, тем самым значительно упрощая работу исследователю при решении задач по анализу социальных сетей.

Цель работы. Целью данной работы является изучение способа определения оптимального количества тем (кластеров) при построении тематического профиля онлайн-сообщества с помощью алгоритма LDA путем вычисления среднего гармонического значения.

Базовые положения исследования. Исследование проводилось с помощью пакета vkR языка программирования R, морфологического анализатора pymorphy2, пакета lda, а также дополнительных пакетов языка программирования R, используемых для обработки данных.

Этап 1. Постановка задачи. Исследовать способ определения оптимального количества тем (кластеров) на основе алгоритма LDA путем вычисления среднего гармонического значения.

Этап 2. Сбор данных. Собрать текстовые корпуса данных (тексты постов и комментариев со стен онлайн-сообществ) с помощью пакета vkR.

Этап 3. Подготовка данных. Нормализация собранных текстовых корпусов данных с помощью вспомогательных пакетов обработки данных и морфологического анализатора pymorphy2.

Этап 4. Моделирование. Построение и визуализация модели с помощью пакетов lda и LDAvis языка программирования R на произвольном количестве тем.

Этап 5. Вычисление оптимального количества тем (кластеров) путем вычисления среднего гармонического значения, сравнение результатов данного этапа с результатами этапа 4.

Этап 5. Объединение текстовых корпусов данных различных онлайн-сообществ в единый корпус, повторение п. 4,5.

Промежуточные результаты работы.

1. Построены тематические профили различных онлайн-сообществ с помощью произвольного количества тем (кластеров), заданных вручную.
2. Предложенный вариант изучения способа определения количества оптимальных тем (кластеров) путем вычисления среднего гармонического состояния был протестирован на конкретных наборах данных, осуществлено сравнение результатов обоих вариантов формирования количества тем.

Основной результат. Результатами данной работы являются построенные тематические профили онлайн-сообществ. Описанный в работе алгоритм определения оптимального количества тем (кластеров) при построении тематического профиля онлайн-сообществ был протестирован на собранных наборах данных и может быть рекомендован к использованию при анализе социальных сетей с целью оптимизации вычислительных ресурсов.