УДК 004.9

ИССЛЕДОВАНИЕ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ ДИСТРИБУТИВНОЙ СЕМАНТИКИ

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

В данной работе рассматриваются методы тематического моделирования и построение тематических моделей с использованием векторных представлений слов, применяемые в рамках исследования интегральных методов обучения для систем автоматического распознавания речи.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР №620173 «Исследование интегральных методов обучения для систем автоматического распознавания речи».

Тематическое моделирование используется для обнаружения скрытой семантической структуры, обычно называемой темами, в большой коллекции документов. В традиционных методах, таких как LSA, PLSA и LDA, отсутствует лингвистическое обоснование, чтобы исправить этот недостаток предлагается использование векторных представлений слов вместо различных способов предварительной обработки данных.

Для проведения исследования используется набор данных «Сборник новостей Lenta.ru», он содержит корпус русскоязычных новостных статей, собранных на сайте Lenta.ru. Данный корпус составляет около 800 тысяч новостей, которые по тематикам «Библиотека», «Россия» и «Мир».

В работе рассматривается модификация метода lda2vec, основанного на создании фрагментов текста размером с документ и разделяя векторы документа на два разных компонента. Подобным же образом, что и реализация метода LDA, вектор документа раскладывается на вектор веса документа и матрицу тем. Вектор веса документа представляет процент различных тем, тогда как матрица тем состоит из разных векторов тем. Таким образом, контекстный вектор создается путем объединения различных тематических векторов, которые встречаются в документе.

В качестве альтернативы рассматривается модификация метода top2vec, который использует семантическое пространство, состоящее из векторов слов и документов, представляющих собой непрерывное представление тем, в отличие от LDA, где темы отбираются из дискретного пространства. Алгоритм работает на основе предположения, что семантически похожие документы следует размещать близко друг к другу в семантическом пространстве, а различные документы следует размещать дальше друг от друга. Так же top2vec позволяет уменьшить количество тем путем слияния векторов тем, которые были наиболее похожи друг на друга.

В данной работе был выполнен обзор существующих методов тематического моделирования с использованием векторных представлений слов lda2vec и top2vec, что является необходимой основой для осуществления дальнейших этапов исследования интегральных методов обучения для систем автоматического распознавания речи.