УДК 004.912

ИЗВЛЕЧЕНИЕ ТЕМ ИЗ ТЕКСТОВ ЛЕКЦИЙ С ПРИМЕНЕНИЕМ ГРАФОВЫХ МОДЕЛЕЙ

Коробова П.И.
(Университет ИТМО)
Научный руководитель – к. т. н. Махныткина О.В.
(Университет ИТМО)

Данная работа посвящена теме извлечения тем и фактов из текстов лекций с применением графовых моделей. Рассмотрены этапы предварительной обработки, алгоритмы и методы решения задачи извлечения тем из текстов на русском языке.

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 и обработка текста представляют собой большой интерес для широкого круга организаций, исследователей и специалистов данной области. Он позволяет выделить ценную информацию из неструктурированных текстов. Такие тексты не могут в исходном виде использоваться для дальнейшей обработки компьютерами, поскольку содержат огромное количество информации в неструктурированном виде. Одной из важных задач при обработке и анализе текстовых конспектов лекций является выявление основных тем и фактов.

Существует несколько подходов для извлечения информации:

- на основе онтологий (англ. *ontology-based*);
- основанные на правилах (англ. *rule-based*);
- основанные на машинном обучении (англ. *machine learning (ML*);
- гибридные подходы.

Необходимым этапом для применения алгоритмов извлечения информации является предварительная обработка текстов. Базовая предобработка как правило включает выявление и удаление списка стоп слов, токенизацию, лемматизацию, также дополнительно можно использовать исправление орфографических ошибок. Список стоп слов должен быть качественно составлен, чтобы исключить слова, которые не относятся к тематике текста. Токенизация используется для того, чтобы разбить текст на более мелкие части, токены. К токенам относятся как слова, так и знаки препинания. Для токенизации могут быть использованы различные функции и инструменты. Лемматизация используется для приведения слова к его базовой форме. Для лемматизации используется морфологический анализатор *pymorphy2*.

После предварительной обработки применяется алгоритм для извлечения информации. Подход на основе графовых моделей является эффективным для решения задачи извлечения ключевых тем. Семантический граф состоит из взвешенного графа, вершинами которого являются термины документа, а наличие ребра между двумя вершинами означает семантическую связь между терминами. Численным значением семантической близости двух терминов является вес ребра, вершины которого соединяют данное ребро. Затем в графе происходит поиск групп и их ранжирование. Обычно ранжирование происходит в соответствие с плотностью групп и его информативностью, которая может быть посчитана с помощью любой статистической метрики. Соответственно слова с наиболее высоким рангом являются ключевыми темами.

Для реализации на русском языке была использована упрощенная графовая модель на основе метода *TextRank,* позволяющая осуществлять обобщение текста, извлечение ключевых слов и терминов.

В данной работе были рассмотрены методы предварительной обработки текста, алгоритм на основе графовых моделей для извлечения тем из лекций. В дальнейшей работе планируется разработка алгоритма для извлечения фактов из конспектов на основе графовых моделей.