**УДК 004.934**

# ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ СЛОВОСОЧЕТАНИЙ В ЗАДАЧЕ АВТОМАТИЧЕСКОЙ ТЕМАТИЗАЦИИ

**Залуская В.С.** (Университет ИТМО), **Астапов С.** (Университет ИТМО), **Кабаров В.И.** (Университет ИТМО)
**Научный руководитель – PhD, Астапов С.**
(Университет ИТМО)

В работе описываются методы извлечения ключевых слов и словосочетаний из текстовых документов в задаче кластеризации текстов по тематикам. Рассматриваются статистические, графовые и нейросетевые модели извлечения ключевых словосочетаний, а также их влияние на восстановление тематических кластеров.

**Введение.** Задача извлечения ключевых словосочетаний направлена на формирование набора фраз, который будет содержать максимальную информацию о документе при малом объеме. Извлечение словосочетаний является предварительным этапом во многих задачах обработки и интерпретации естественных языков: распознавание именованных сущностей, классификация текстовых документов, извлечение информации из текста. Одной из основных сложностей является выявление нетипичных ключевых словосочетаний, присущих конкретной доменной области или отдельному корпусу данных. В связи с этим, в ряде решений возможно использование локального и глобального контекста, чтобы выявить как общепринятые ключевые слова, так и использующиеся исключительно в выбранном наборе данных. Широкий набор ключевых слов позволяет более гибко настраивать тематическое моделирование на основе кластеризации.

**Основная часть.** В работе рассматриваются методы извлечения ключевых слов и словосочетаний из набора документов в задаче кластеризации текстов по тематикам. По архитектуре методы можно разделить на статистические, графовые и нейросетевые. Процесс имплементации и анализа разделен на следующие стадии: предобработка текста, извлечение словосочетаний, извлечение признаков, ранжирование словосочетаний на основе признаков, восстановление тематик на основе ТОП-N ключевых словосочетаний. При использовании статистической модели извлекаемые признаки опираются на частотность конкретного слова или словосочетания как в отдельном документе, так и в целом корпусе. Графовые модели, в свою очередь, позволяют использовать теорию графов и показатель центральности при ранжировании словосочетаний, представленных в виде вершин. Нейросетевые модели расширяют контекст и позволяют использовать не только знания из исследуемого набора документов, но и знания из других доменных областей, использовавшихся при предварительном обучении моделей.

**Выводы.** В результате исследования были выделены признаки на основе статистических, графовых и нейросетевых архитектур, а также использованы алгоритмы ранжирования и кластеризации для выделения тематик. При использовании статистических и графовых архитектур наблюдалось выявление закономерностей конкретного документа и корпуса данных, тогда как предобученные нейросетевые модели выделяли в качестве ключевых слов более широкий набор словосочетаний, что отразилось на конечной кластеризации документов по тематикам.