

МЕТОД СЕМАНТИЧЕСКОГО АНАЛИЗА АРХИТЕКТУРНЫХ МАТЕРИАЛОВ

Усман Т. Г. (ИТМО),

Научный руководитель – ассистент преподавателя, ведущий инженер Антонов А.С. (ИТМО)

Введение. В настоящее время существует большое разнообразие открытых градостроительных конкурсов. В их процессе подготавливаются текстовые и визуальные материалы, в которых описывается заявка и основа предлагаемого проекта. В результате объем информации, который необходимо обработать экспертам для предварительной оценки проекта, может превышать имеющийся объем временных ресурсов и внимания. В такой ситуации особенности предлагаемых проектов могут быть упущены. Кроме того, анализ заявок как набора данных может предоставить новые социологические и антропологические знания о существующей идентичности объекта проектирования. Для решения этих задач был разработан метод автоматизированного семантического анализа текстовых проектных материалов на основе BERTopic [1] в сфере градостроительства.

Основная часть. Современные подходы к анализу текстовых данных в градостроительстве включают методы машинного обучения, такие как тематическое моделирование (Latent Dirichlet Allocation, LDA) [2], анализ тональности с использованием методов на основе BERT [3], а также кластеризацию текстов с применением методов снижения размерности, таких как UMAP [4]. Однако эти методы часто требуют значительной предобработки данных и не всегда учитывают контекстную семантику. BERTopic, основанный на трансформерных моделях, позволяет эффективно выделять темы и кластеризовать тексты с учетом контекста [1].

Для анализа использовались 53 архитектурные концепции студентов, относящиеся к тюрьме “Кресты” и представленные в виде многостраничных PDF-файлов, содержащих текстовую и графическую информацию. Текстовые данные были извлечены с использованием OCR (ABBY FineReader) и предобработаны для дальнейшего анализа.

Описание метода:

1. Извлечение текста из PDF-файлов и его очистка – на этом этапе текстовые данные извлекаются из многостраничных PDF-файлов с использованием OCR. После извлечения текст очищается от общих и стоп-слов, которые могут повлиять на качество анализа.
2. Векторизация текста с использованием моделей BERT – для преобразования текста в числовые векторы используется предобученная модель BERT для русскоязычных текстов (sbert_large_nlu_ru). Этот этап позволяет сохранить контекстную семантику текста, что особенно важно для анализа сложных градостроительных проектов;
3. Кластеризация текстов и выделение 5 тем с помощью BERTopic – на заключительном этапе тексты кластеризуются с использованием BERTopic, который сочетает методы снижения размерности (UMAP) и кластеризации (HDBSCAN) для выделения ключевых тем. Фиксированное количество тем обусловлено тем, что тексты в основе достаточно похожи друг на друга, и HDBSCAN не справлялся с задачей выделения кластеров.

В качестве результатов были получены 5 кластеров, соотнесенных с темой их элементов. Специфика исходных данных (студенческие работы) объясняет их содержание: музей репрессий, ревитализация и приспособление, общая тема, переосмысленный символ общества, мемориальное культурно-просветительское пространство. Также представлена столбчатая диаграмма частоты встречаемости слов в темах, которая помогает определить

наиболее значимые элементы, которые характеризуют содержание кластеров, и иерархическая кластеризация.

Выводы. Разработанный метод автоматизированного семантического анализа на основе BERTopic позволяет эффективно обрабатывать большие объемы текстовых данных, выделять ключевые темы и интерпретировать их в контексте градостроительных проектов. В дальнейшем возможно расширить анализ за счет включения визуальных данных и применения методов мультимодального обучения, таких как CLIP [5], [6].

Список использованных источников:

1. Грутендорст М. П. BERTopic. — 04.2024. — URL: <https://maartengr.github.io/BERTopic>
2. Блей Д. М., Нг А. Я., Джордан М. И. Латентное размещение Дирихле // Журнал исследований машинного обучения. — 2003.
3. Девлин Дж., Чанг М.-В., Ли К., Тутанова К. BERT: Предварительное обучение глубоких двунаправленных трансформеров для понимания языка // arXiv:1810.04805. — 2019.
4. МакИннес Л., Хили Дж., Мелвилл Дж. UMAP: Равномерное приближение и проекция многообразия для снижения размерности // arXiv:1802.03426. — 2018.
5. Грутендорст М. П. BERTopic. Multimodal Topic Modeling — 04.2024. — URL: https://maartengr.github.io/BERTopic/getting_started/multimodal/multimodal.html
6. Радфорд А., Ким Дж. У., Халласи К. и др. Обучение переносимым визуальным моделям с использованием естественно-языкового контроля // arXiv:2103.00020. — 2021.

Автор _____ Усман Т.Г.

Научный руководитель _____ Антонов А.С.