УДК 004.89

ПОДХОД К АВТОМАТИЗАЦИИ СОЗДАНИЯ КОНТЕНТА ДЛЯ ОБРАЗОВАТЕЛЬНЫХ ПЛАТФОРМ НА ОСНОВЕ ДАННЫХ РЫНКА ТРУДА

Фомочкина А.Д. (ИТМО)
Научный руководитель – к.т.н. Болдырева Е.А.
(ИТМО)

Введение. Контент образовательных платформ необходимо адаптировать под актуальные требования рынка труда. С помощью интеллектуального анализа данных можно выявить востребованные навыки и компетенции, а применение нейронных сетей для генерации текста позволит автоматизировать обновление контента, снизив затраты на поддержание его актуальности.

Основная часть. Целью данной работы является разработка подхода к автоматизации создания контента для образовательных платформ на основе данных рынка труда с использованием методов машинного обучения. Для этого был проведен анализ предметной области, рассмотрены современные методы адаптации образовательных курсов, а также возможности интеллектуального анализа данных в сфере образования и рынка труда.

Исследование основано на данных вакансий с сайта hh.ru, относящихся к профессии «аналитик». Основное внимание уделено полю с ключевыми навыками. Для анализа текстовых данных проведена предварительная обработка, включающая удаление пунктуации, лемматизацию, токенизацию и фильтрацию стоп-слов. В статье [2] метод TF-IDF показал высокие результаты при работе с похожими данными ключевых навыков из вакансий. Поэтому для выделения значимых терминов использовался метод TF-IDF, который оценивает важность слов и исключает малозначимые термины. Для кластеризации навыков были исследованы три метода: метод K-means, Латентное распределение Дирихле (LDA) и BERTopic.

Метод K-means показал наилучшие результаты. Оптимальное число кластеров ($k = 10$) было определено с помощью метода локтя. В результате выделено 8 осмысленных кластеров, связанных с ключевыми компетенциями аналитиков, такими как SQL, бизнес-аналитика, маркетинг и другие. LDA оказался менее эффективным, так как некоторые темы было сложно интерпретировать. BERTopic выделил множество тем, но после их сокращения до 10 результаты примерно совпали с другими методами. Однако некоторые темы, выделенные BERTopic, оказались очень похожими друг на друга. Таким образом, метод K-means был выбран как наиболее подходящий благодаря высокой скорости работы, возможности автоматического определения числа кластеров и четкой интерпретируемости результатов.

Для генерации образовательного контента использовались модели нейронных сетей YandexGPT 4 и DeepSeek-V3. Обе модели продемонстрировали высокую эффективность в задачах присвоения названий кластерам и создания учебных материалов, включая задания, тестовые вопросы, списки литературы и краткие лекции. Однако DeepSeek-V3 показала существенное преимущество, так как она не только формирует перечень рекомендованных источников, но и предоставляет ссылки на книги и интернет-ресурсы, что значительно упрощает процесс дальнейшего изучения темы.

Детальность запроса к текстовой генеративной модели играет ключевую роль в качестве полученных образовательных материалов. В статье [1] авторами предложен план составления запроса для создания тестовых материалов. Он был использован для генерации тестовых заданий по выделенным тематикам.

На основе проведенного исследования разработана технология актуализации образовательного контента. Она включает следующие этапы: извлечение ключевых навыков из вакансий, кластеризация навыков с помощью метода K-means, определение областей знаний по кластерам, генерация образовательных материалов с использованием модели DeepSeek-V3.

В результате исследования была разработана технология обновления контента образовательной платформы с учетом анализа данных рынка труда. Метод K-means показал себя как наиболее эффективный инструмент для кластеризации навыков, а модель DeepSeek-V3 доказала свою полезность в генерации учебных материалов и рекомендаций по дополнительным ресурсам.

Выводы. Предложен подход к автоматизации создания контента для образовательных платформ на основе данных рынка труда. Преимущество данного подхода заключается в быстроте создания актуального контента, соответствующего определённой профессии, но перед использованием в образовательных целях контент должен быть проверен экспертами предметной области.

Список использованных источников:

1. Калинин А. А., Королева Н. Ю., Рыжова Н. И., Фёдорова Ю. В. Искусственный интеллект в образовательном контенте: актуальный тренд и практические аспекты эволюции учебного процесса // Наука и школа. 2024. № 5. С. 98–113. URL: https://cyberleninka.ru/article/n/iskusstvennyy-intellekt-v-obrazovatelnom-kontente-aktualnyy-trend-i-prakticheskie-aspekty-evolyutsii-uchebnogo-protsessa
2. Boldyreva E.A., Kholoshnia V.D. Ontological Approach to Modeling the Current Labor Market Needs for Automated Workshop Control in Higher Education // CEUR Workshop Proceedings, 2020, Vol. 2590, pp. 1-13 Scopus URL: https://ceur-ws.org/Vol-2590/paper29.pdf