0.1. *Шашок Н.А., Кожемякина Э.Д.* Разработка структуры документов с пересекающейся сегментацией в системе Elasticsearch

Один из ключевых модулей системы автоматизированного комплексного анализа русских поэтических текстов [1,2], разработанной в ФИЦ ИВТ, - модуль автоматического составления словарей авторского языка и конкордансов. В процессе создания словарей такого типа возникает задача определения контекста употребления лексики.

Термин «контекст» может обозначать различные уровни сегментации текста: в качестве сегмента, включающего контекст, выступает строфа, строка, или предложение. Строфы и строки явно связаны иерархическими отношениями, поскольку строфы состоят из строк. Однако строки и предложения, а также предложения и строфы иерархических отношений не имеют: предложение может быть частью строки или строфы, занимать несколько строк, начинаться на одной строфе и заканчиваться на другой; такие сегменты поэтического текста будем называть пересекающимися.

Модуль должен предоставлять возможность работы с любым доступным контекстом, на любом уровне сегментации. Поскольку форматы хранения и передачи текстовых данных имеют, как правило, иерархичный характер, практический интерес представляет разработка принципов структуризации текстов с учетом выявленных пересекающихся сегментов в рамках задачи поиска контекста с предварительно заданным уровнем сегментации.

Для решения поставленной задачи целесообразно использование поисковой системы Elasticsearch, документы в которой хранятся в формате JSON. В работе представлена структура документов JSON, использование которой в рамках индекса Elasticsearch позволяет осуществлять поиск контекста употребления лексики в корпусе поэтических текстов, хранящегося в индексе. Разработка имеет общие черты с решениями проблемы пересекающихся сегментов в языках разметок [3,4], в частности, с решением типа Segmentation. Принципиальна возможность применения разработанной структуры также к корпусам не только поэтических текстов, что расширяет область потенциального использования модуля автоматического составления словарей в других информационных системах.

Научный руководитель — д.т.н., к.филол.н. Кожемякина О. Ю.

Список литературы

[1] Система комплексного анализа поэтических текстов. [Электронный ресурс]. URL: www.poem.ict.nsc.ru (дата обращения 29.08.2023).

[2] Кожемякина О. Ю. Программная система комплексного анализа русских поэтических текстов: модели и алгоритмы: дис. ... д-ра т. наук. ФГБОУ ВО «Сибирский государственный университет телекоммуникаций и информатики», Новосибирск, 2022. 288 с.

[3] Di Iorio A., Peroni S., Vitali F. Towards markup support for full GODDAGs and beyond: the EARMARK approach // Proc. Intern. Conf. «Balisage: The Markup Conference». Montr?al: Mulberry Tecnologies, Inc., 2009. Vol. 3.

[4] Schmidt D. The role of markup in the digital humanities // Historical Social Research. 2012. Vol. 27. N. 3. P. 125–146.