Санкт-Петербургский государственный электротехнический университет им. В.И. Ульянова (Ленина)

Разработка инструмента анализа научных статей методами text mining

Выполнил:
Нгуен Куанг Хуи, гр. 7304
Руководитель:
Заславский Марк Маркович, к.т.н., доцент

Санкт-Петербург, 2020

Цель

Актуальность:

- Проверка научной статьи требует много времени и усилий
- Существующие инструменты не имеют расширенных функций, которые используют text mining для обеспечения полезного понимания научной статьи

Цель: Разработайте веб-сервис, которая предоставляет функции оценки ключевых слов, обнаружения некогерентных предложений и обнаружения отсутствующих предложений-связки

Задачи

1. Обзор моделей оценки ключевых слов и обнаружения некогерентности
2. Разработка сервиса, обеспечивающего реализацию указанных функций
3. Интеграция разработанного сервиса в существующую систему
4. Тестирование разработанного сервиса и анализ потребности в ресурсах (с точки зрения памяти и времени обработки)

Определение проблемы

1. Выбор ключевых слов
2. Некогерентность в научной статье
    1. Некогерентные предложения
    2. Отсутствующие предложения-связки

Обзор моделей оценки ключевых

	Рассмотрение отношения по смыслу между ключевой фразой и статьей	Рассмотрение разницы между различными частями статьи	Предоставление абсолютной оценки	Относительное упорядочение точности в бенчмарке
TF-IDF	-	-	-	3
TextRank	-	-	-	5
EmbedRank	+	-	-	-
PositionRank	-	+	-	4
KEA	-	+	-	2
Метод Nguyen и Kan	-	+	-	-
SBERT-MLP	+	+	+	1

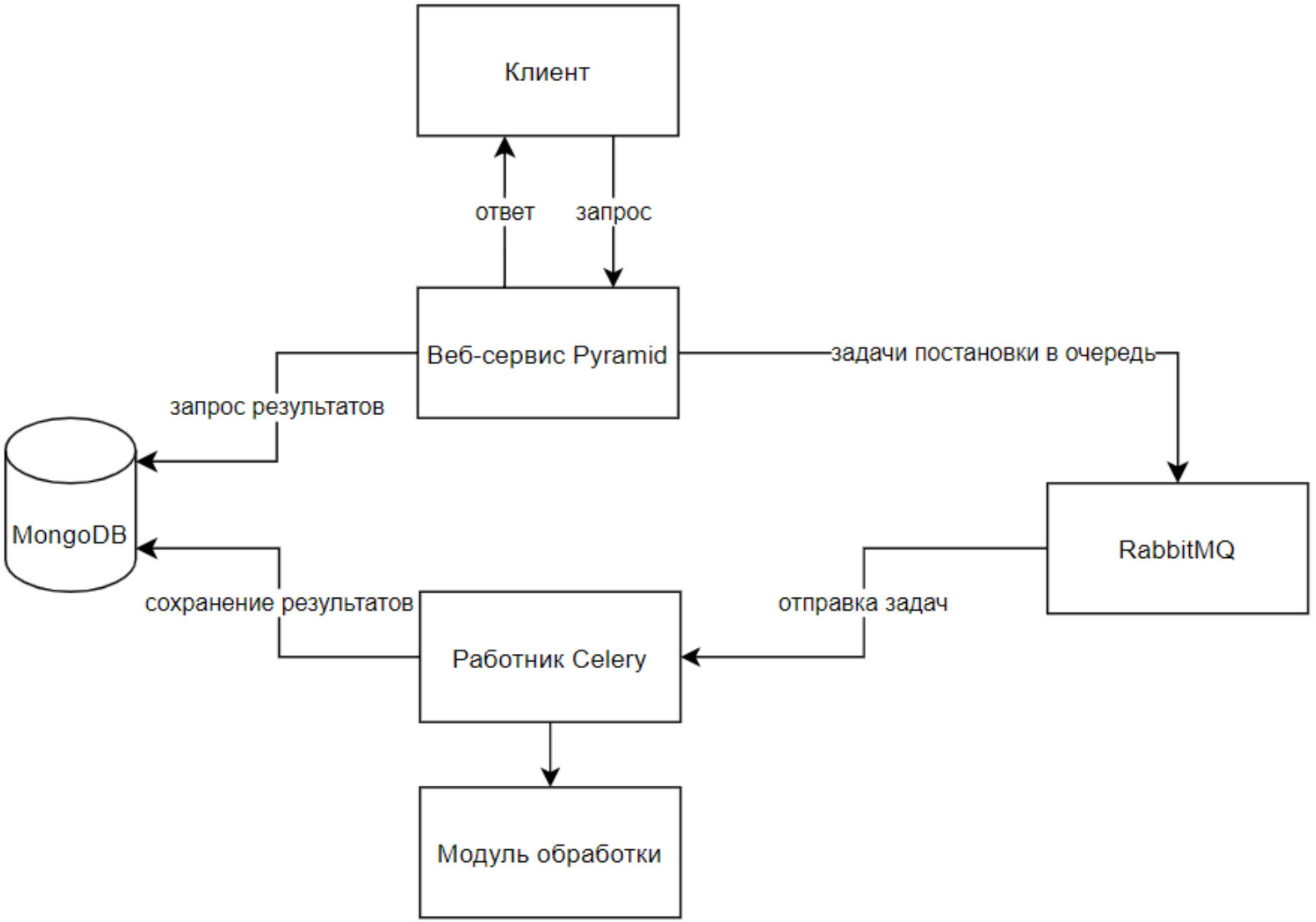
Таблица 1 - Сравнение методов извлечения ключевых слов

Обзор моделей обнаружения некогерентности

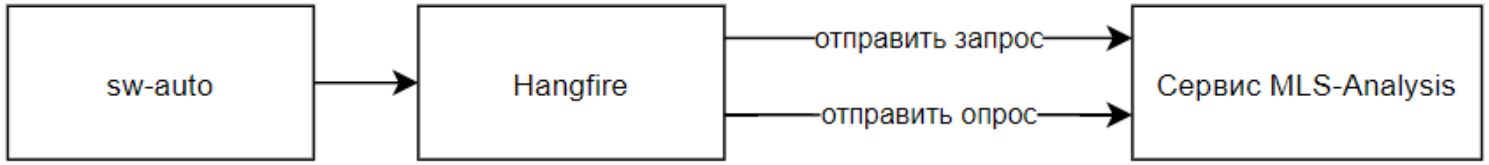
	Фиксированное потребление памяти	Ограничение длины абзаца	Бенчмарк для обнаружения некогерентных предложений	Бенчмарк для обнаружения отсутствующих предложений-связки
Токен-ориентированный метод	1 модель	+	2	1
Предложение-ориентированный метод	2 модели	-	3	2
BERT-Clustering	1 модел	-	1	3

Таблица 2 - Сравнение методов обнаружения некогерентности

Разработка сервиса, обеспечивающего реализацию указанных функций



Интеграция разработанного сервиса в существующую систему



Ключевые слова

управление ресурсами 1.00

менеджер задач 0.01

система управления проектами 0.99

трудозатраты 0.00

аналитика трудозатрат 0.96

Качество ключевых слов

Критерий отражает стастическую и смысловую важность ключевого слова относительно текста работы.

Значение: 0.590

Требования: Значение критерия должно находиться в интервале [0.5, 1]

Набрано 9.46 баллов

Кандидаты ключевых слов

системах управления проектами 0.99

работе 0.98

системах управления 0.98

управления проектами 0.97

трудозатрат 0.97

аналитикой трудозатрат 0.96

сотрудников 0.96

аналитики трудозатрат 0.96

веб-сервиса 0.95

сравнительного анализа 0.94

Предложение, не связанное по смыслу с остальным текстом

Набрано 9.46 из 9.46 возможных баллов

Найдено ошибок: 1

Убедитесь, что предложение действительно по смыслу необходимо в данном фрагменте

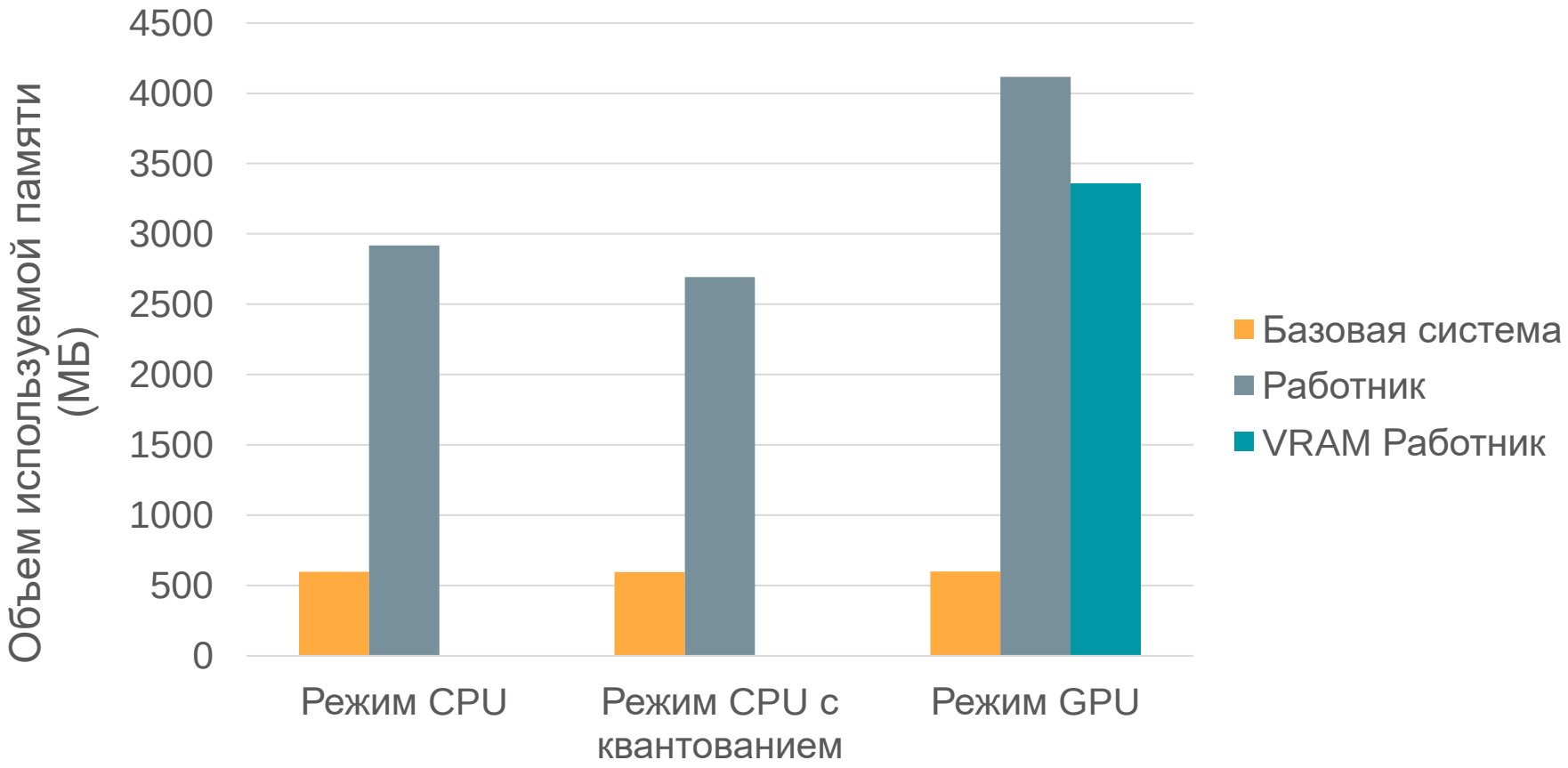
Отсутствуют предложения-связки

Набрано 9.46 из 9.46 возможных баллов

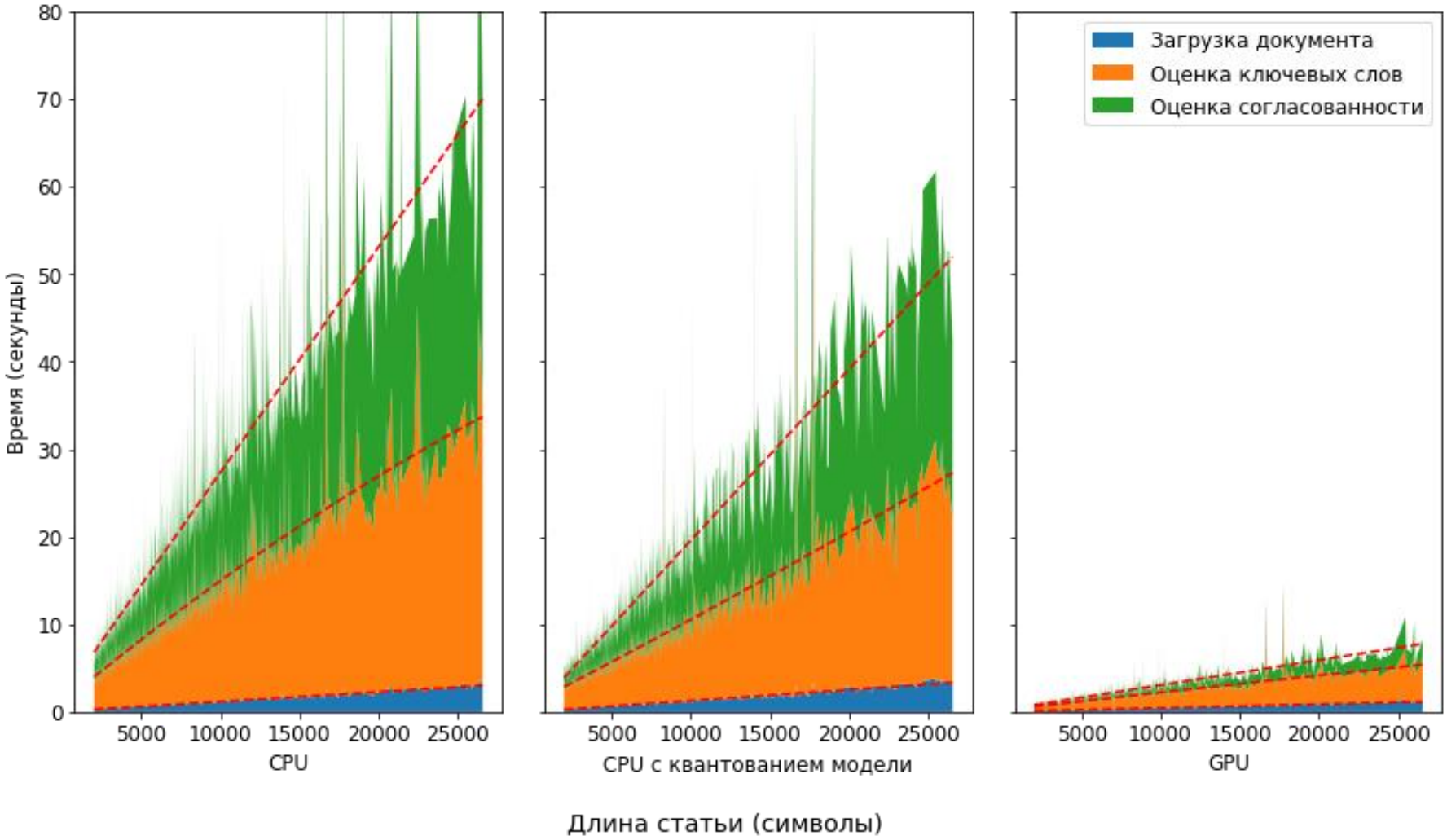
Найдено ошибок: 1

Добавьте предложения-связки, которые привяжут предложение к абзацу/разделу

Анализ потребности ресурсах



Анализ потребности ресурсах



Заключение

- Был разработан сервис для предоставления расширенных функций (оценка ключевых слов и обнаружение некогерентности) для помощи рецензентам в проверке научных статей.
- Сервис успешно интегрирован в существующую систему.

Апробация работы

- [SCOPUS] «Keyphrase Extraction in Russian and English Scientific Articles Using Sentence Embeddings» // Конференция FRUCT 28th, 2021.
- «Incoherent Sentence Detection in Scientific Articles in Russian and English» // Конференция FRUCT 29th, 2021.
- Репозиторий проекта https://github.com/moevm/bsc_nguyen_quang_hui

Запасные слайды