



Федеральное государственное бюджетное образовательное учреждение высшего образования  
«Московский государственный технический университет имени Н.Э. Баумана  
(национальный исследовательский университет)»

# Разработка системы извлечения терминов

Студент: Сапожков Андрей Максимович ИУ7-63Б

Научный руководитель: Строганов Юрий Владимирович

# Цель и задачи

**Цель** – разработка системы извлечения многокомпонентных терминов и их переводных эквивалентов из параллельных научно-технических текстов.

## **Задачи:**

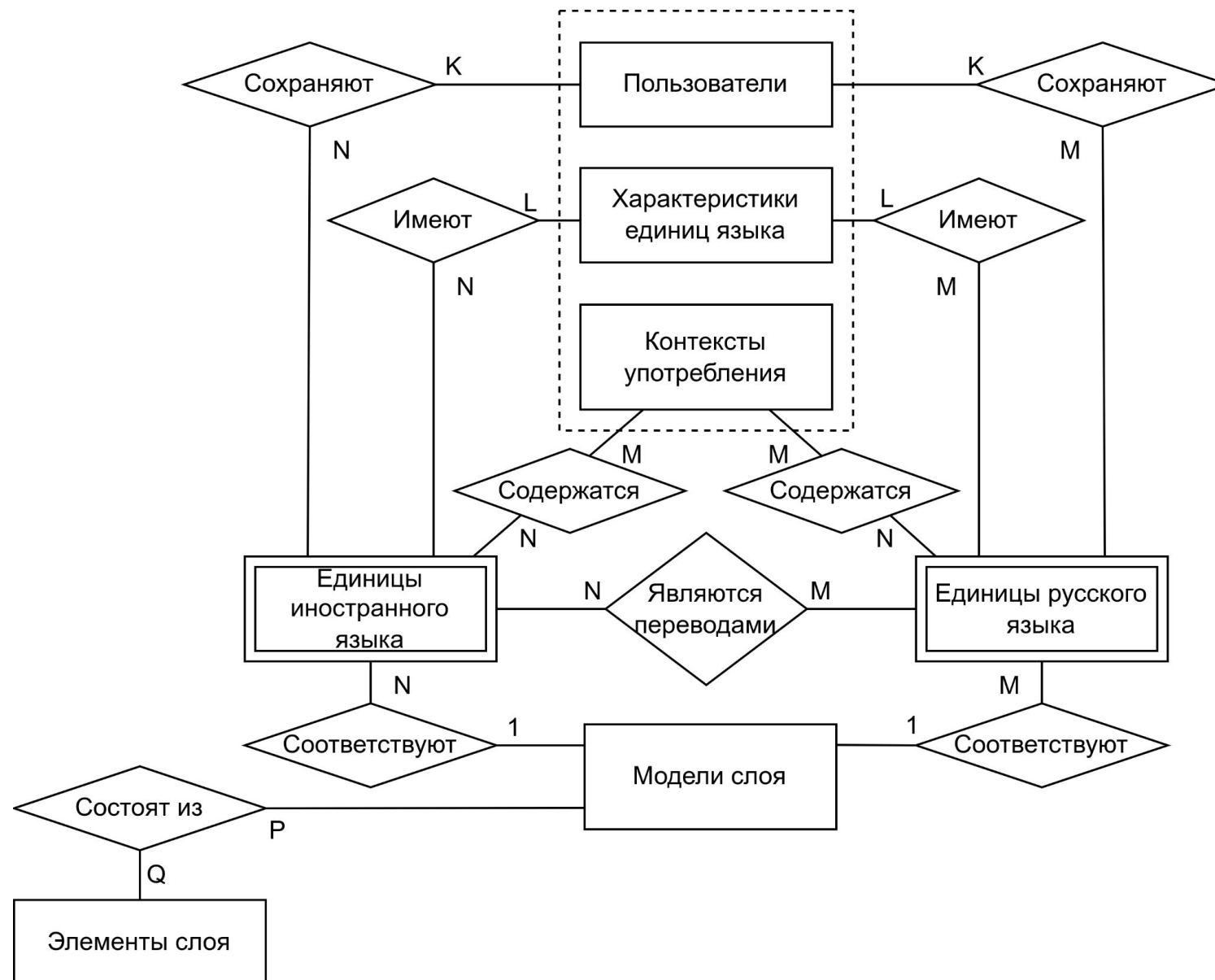
1. Провести анализ предметной области и формализовать задачу.
2. Спроектировать базу данных и структуру программного обеспечения.
3. Реализовать интерфейс для доступа к базе данных.
4. Реализовать ПО, которое позволит пользователю создавать, получать и изменять сведения из разработанной базы данных.
5. Провести исследование зависимости времени выполнения запросов от использования кеширования данных текущей сессии пользователя.

# Существующие аналоги

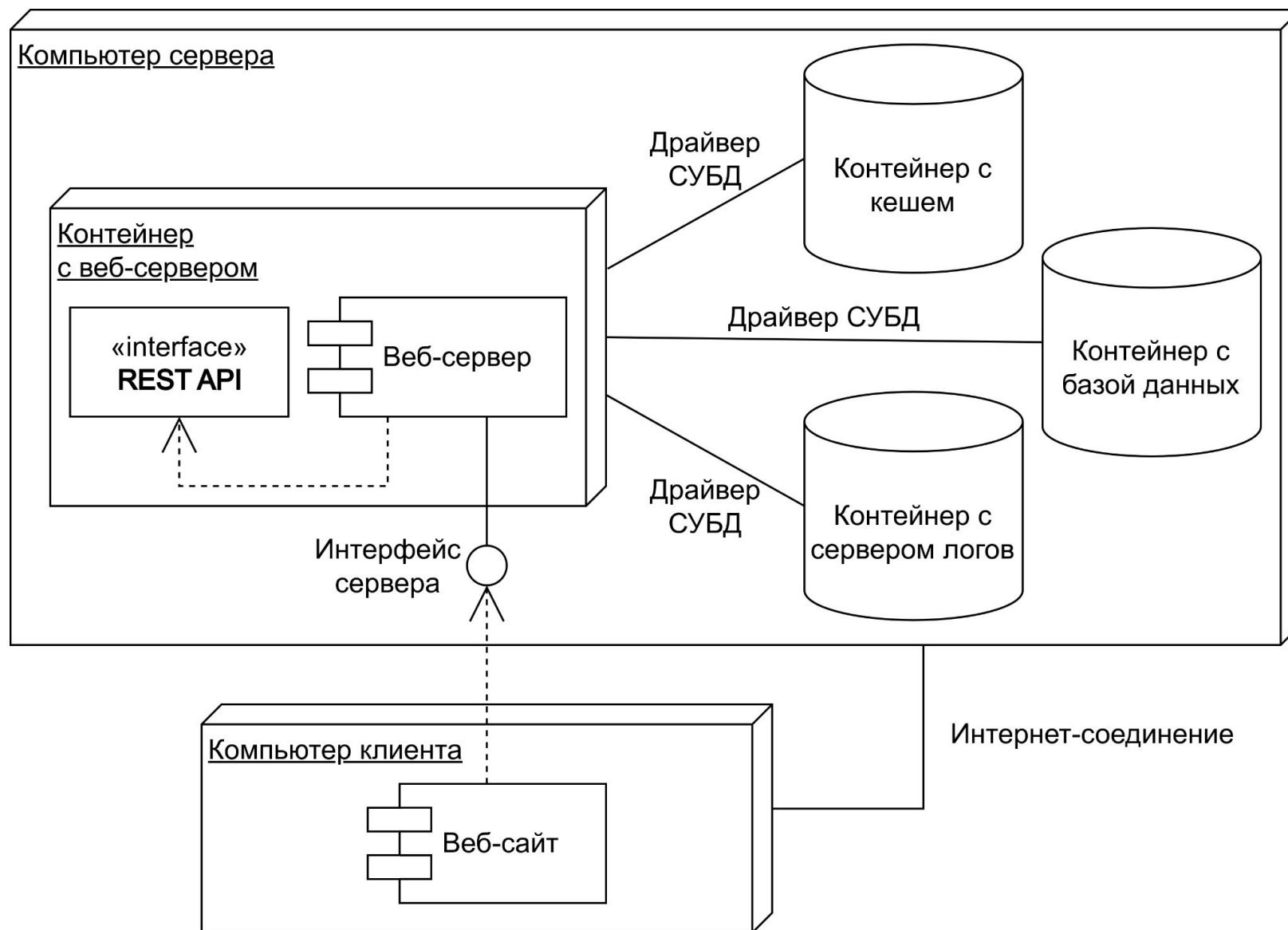
1. **Переводчики** (Google, Яндекс, DeepL) - позволяют размечать тексты, но не дают работать с терминологией. Также нет возможности редактировать "разметку".
2. **Словари** (Thesaurus) позволяют в какой-то степени изучать терминологию, но не дают возможности дополнять её на основе размеченных текстов.



# ER-модель базы данных



# Архитектура ПО



# Анализ СУБД

СУБД	Характеристики	Назначение
PostgreSQL	<ul style="list-style-type: none"><li>● Объектно-реляционная модель</li><li>● Открытый исходный код</li><li>● Сертификация ФСТЭК России</li></ul>	Долговременное хранение данных
Redis	<ul style="list-style-type: none"><li>● Хранилище типа ключ-значение</li><li>● Хранение данных в оперативной памяти</li></ul>	Кеширование данных
InfluxDB	<ul style="list-style-type: none"><li>● Ориентированность на хранение и обработку временных рядов</li><li>● Оптимизация записи данных</li><li>● Встроенные графические средства визуализации данных</li></ul>	Хранение логов

# Исследование

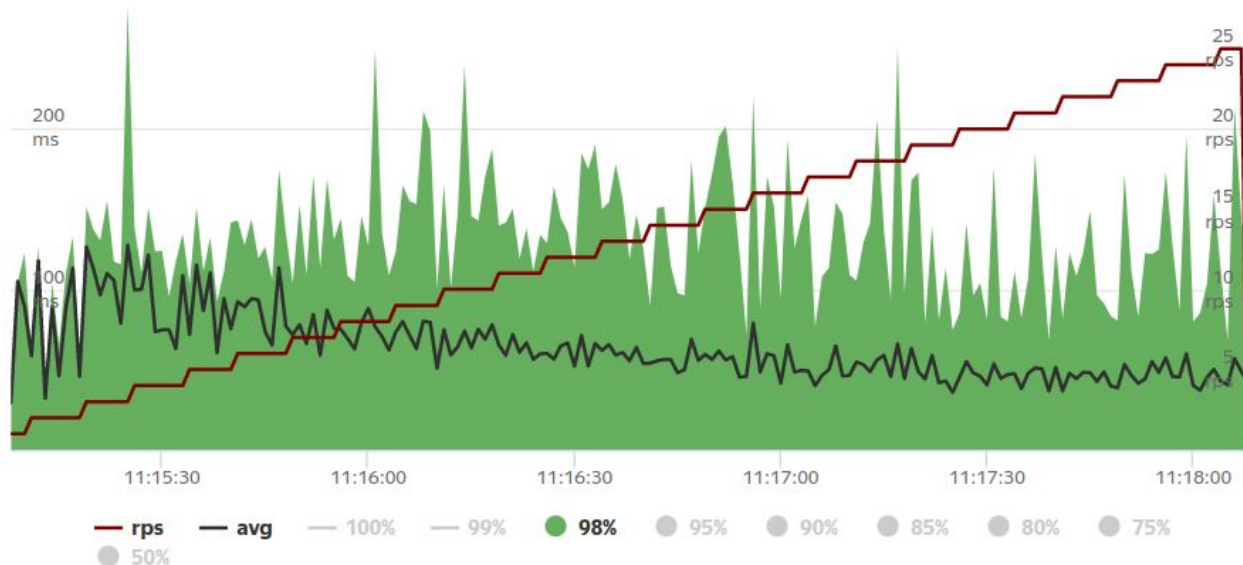
**Цель** – проведение нагрузочного тестирования и сравнение производительности веб-сервера при обработке запросов на сохранение терминов с использованием кеширования и без него.

**Способ проведения** – нагрузочное тестирование с помощью Yandex.Tank.

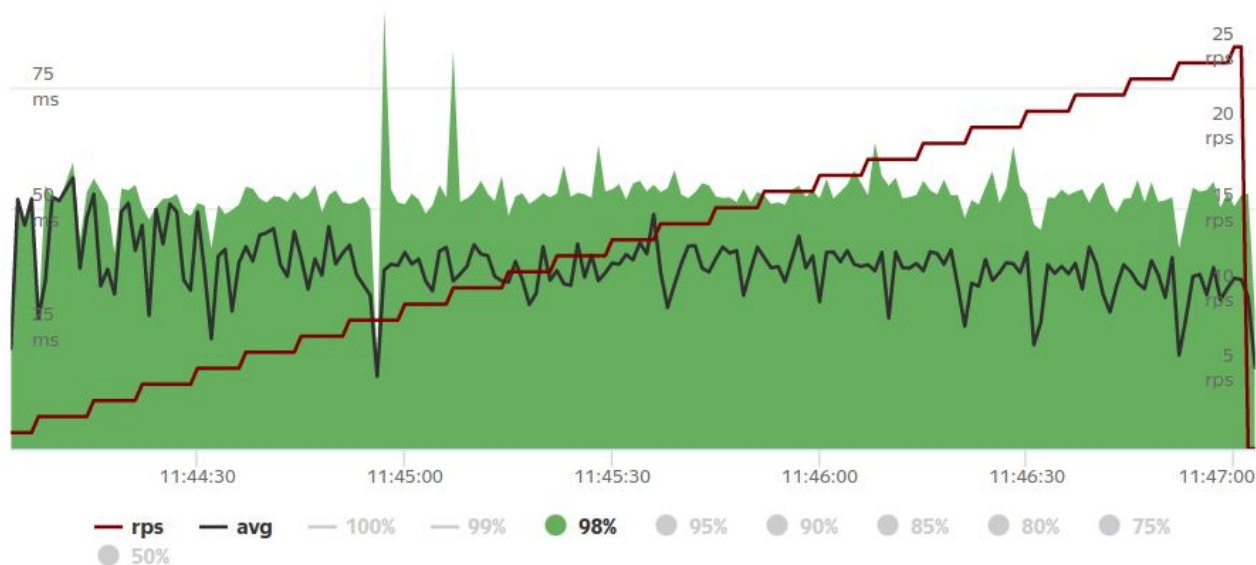
**Технические характеристики:**

1. Операционная система: Manjaro Linux x86-64, версия ядра 5.15.32.
2. Объём оперативной памяти: 16 Гб.
3. Процессор: Intel i5-9300H 2.4 ГГц.

# Открытая линейная нагрузка



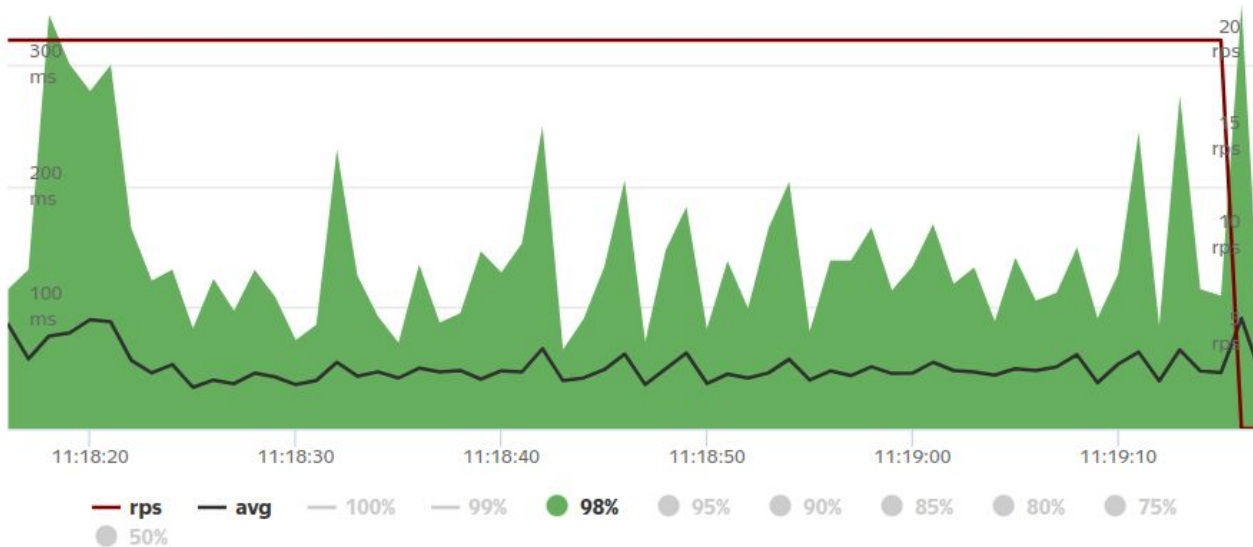
Без кеширования



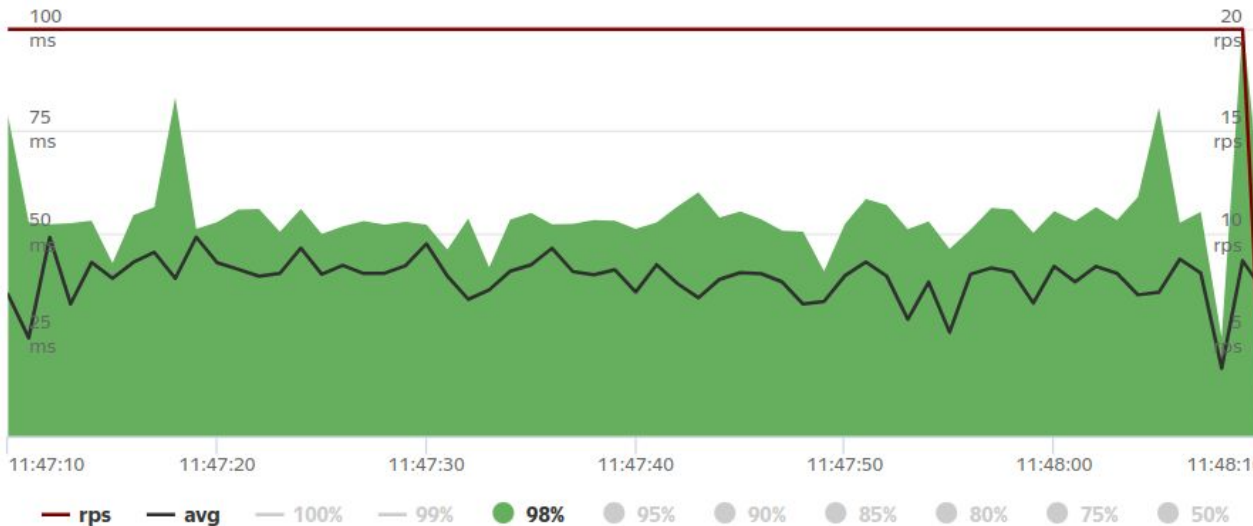
С кешированием



# Открытая постоянная нагрузка

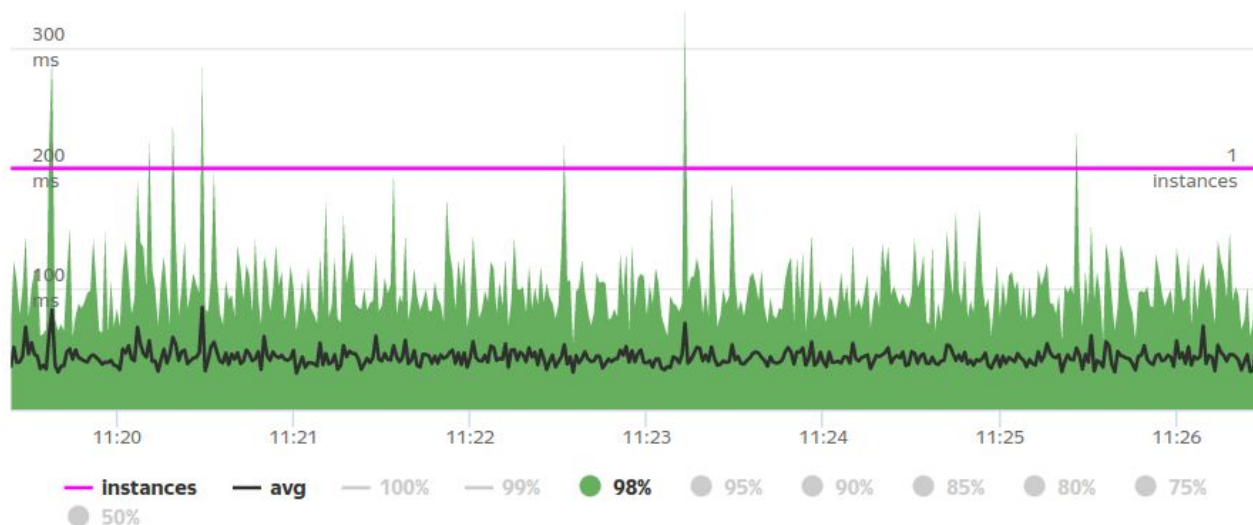


Без кеширования

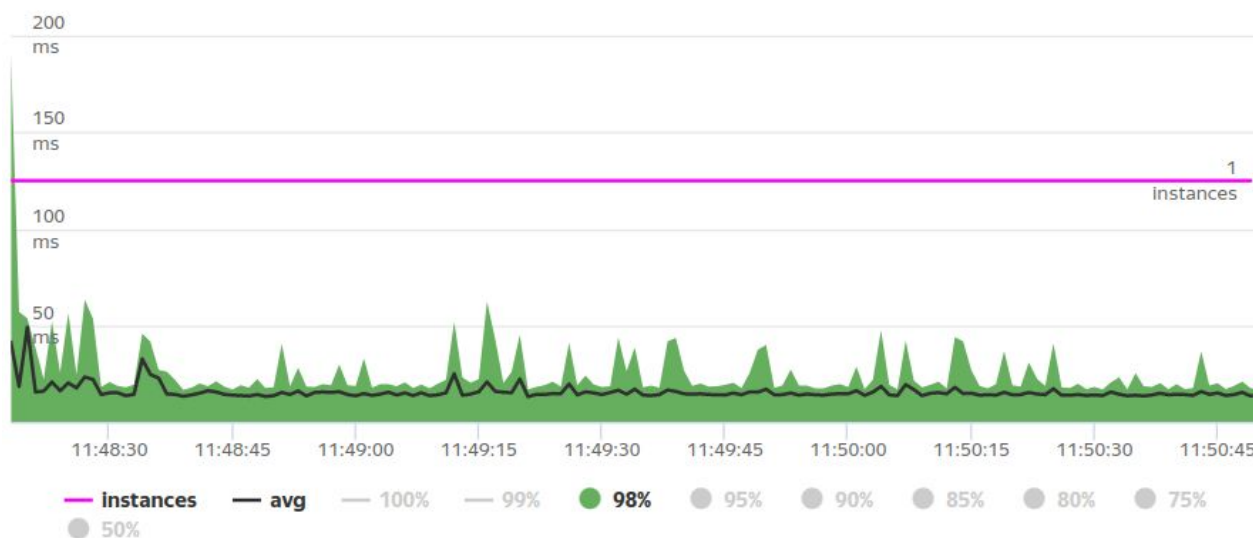


С кешированием

# Закрытая нагрузка



Без кеширования



С кешированием

# Результаты исследования

С использованием кеширования

- 1) при линейной открытой нагрузке среднее время ответа сервера уменьшилось с 63 мс до 37 мс (на 41%);
- 2) при линейной постоянной нагрузке среднее время ответа сервера уменьшилось с 49 мс до 38 мс (на 22%);
- 3) при закрытой нагрузке среднее время ответа сервера уменьшилось с 44 мс до 15 мс (на 66%);
- 4) уменьшился разброс времён ответов, то есть сервер стал отвечать на запросы стабильнее.

# Заключение

В рамках курсовой работы была разработана система извлечения многокомпонентных терминов и их переводных эквивалентов из параллельных научно-технических текстов. Для достижения этой цели были решены следующие задачи:

1. Проведён анализ предметной области и формализована задача.
2. Спроектирована база данных и структура ПО.
3. Реализован интерфейс для доступа к базе данных.
4. Реализовано ПО, которое позволяет пользователю создавать, получать и изменять сведения из разработанной базы данных.
5. Проведено исследование зависимости времени выполнения запросов от использования кеширования данных текущей сессии пользователя.