

РАЗРАБОТКА БАЗЫ ДАННЫХ ДЛЯ АВТОМАТИЗАЦИИ РАБОЧЕГО МЕСТА РАЗМЕТЧИКОВ ПАРАЛЛЕЛЬНОГО КОРПУСА ТЕХНИЧЕСКИХ ТЕКСТОВ

Выполнил:

студент 3 курса
группы ИУ7-64Б

Рунов Константин Алексеевич

Руководитель:

Строганов Юрий Владимирович

Москва, 2024 г.

Цель: Разработать базу данных для автоматизации рабочего места разметчиков параллельного корпуса технических текстов.

Задачи:

- Провести анализ предметной области корпусов текстов;
- Спроектировать и разработать базу данных, описать ее сущности, ограничения целостности, ролевую модель на уровне базы данных и используемые триггеры;
- Разработать приложение для доступа к базе данных;
- Исследовать зависимость времени ответа от количества запросов в секунду и сравнить эффективности реализаций приложения с использованием дополнительного кеширования и без него.

Параллельные корпуса — корпуса, представляющие собой множество текстов-оригиналов, написанных на каком-либо исходном языке, и текстов — переводов этих исходных текстов на один или несколько других языков.

7. Gregory Berns. What emotions look like in a dog's brain [TED Talks] (2015) | Грегори Бёрнс. Как нас любят собаки (Екатерина Юсупова) ⓘ

Все примеры — 56

английский:

What emotions look like in a dog's brain 00: 17 ⓘ ↔

английский:

00: 17 How many of you are dog people? A show of hands. ⓘ ↔

английский:

(Laughter) So, of the dog people and the cat people who want to be dog people, (Laughter) how many of you have thought, "Wouldn't it be great to know what my dog is thinking?" ⓘ ↔

русский:

Как нас любят *собаки* 00: 17 ⓘ ↔

русский:

00: 17 Сколько в зале любителей *собак*? Поднимите руки. ⓘ ↔

русский:

(Смех) Сколько собачников и кошатников, желающих стать собачниками, (Смех) когда-либо думали: «Было бы здорово знать, о чём думает моя *собака*?» ⓘ ↔

Анализ предметной области

Пример разметки текста в OpenCorpora.

OpenCorpora Разметка ▼ Словарь Статистика Скачать О проекте Бейджи Войти ▼

Предложение разобрано автоматически.

Весь текст: Махнул он рукой, да и отошел в сторону.

[Показать исходный текст](#)

Источник: Горький, Максим (весь текст)

Отменить правки

История

Комментировать

Махнул ✓	он ✓	рукой ✓	✓	да ✓	и ✓	отошел ✓	в ✓	сторону ✓	✓
<div>✓ v махнул VERB, perf, intr. masc, sing, past, inde</div>	<div>✓ v он NPRO, masc, 3per, Anph, sing, nomn</div>	<div>✓ v рука NOUN, inan, femn, sing, ablt</div>	<div>✓ v . PNCT</div>	<div>✓ v да CONJ</div>	<div>✓ v и CONJ</div>	<div>✓ v отошел VERB, perf, intr, masc, sing, past, inde</div>	<div>✓ v в PREP</div>	<div>✓ v сторону NOUN, inan, femn, sing, accs</div>	<div>✓ v . PNCT</div>
				<div>✓ v да PRCL</div>	<div>✓ v и INTJ</div>		<div>✓ v в NOUN, inan, masc, Fixd, Abbr, plur, ablt</div>		
				<div>✓ v да INTJ</div>	<div>✓ v и PRCL</div>		<div>✓ v в NOUN, inan, masc, Fixd, Abbr, plur, accs</div>		
					<div>✓ v и NOUN, anim, masc, Fixd, Abbr, plur, ablt</div>				

Существующие аналоги



An overview of the OPUS collection

1,210 CORPORA
45,945,946,108 TOTAL SENTENCE PAIRS
744 LANGUAGES AVAILABLE
THIS MAP DISPLAYS 10 CORPORA, WHICH MAKE UP A TOTAL 93% OF THE OPUS COLLECTION

OUR CONTRIBUTORS



Лицензия цены
Скачать до 40%

Открыть AMPUS
на переводчике Словесник

Организация и ИТ
Бесплатно для ИТ



MyMemory
by translated LABS



Corpora
Language corpora, multi-billion word collections of texts, provide source data for all features of Sketch Engine. Our corpora are tagged and annotated to be ready for complex searches of phrases and language structures. Parallel and multilingual corpora are also available.

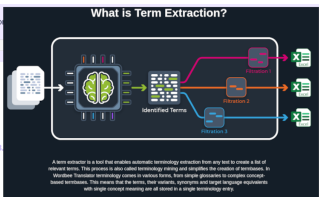
hunalign – sentence aligner

ce level. Its input is tokenized and sentence-segmented text in 1 sentences).

Human contributions

From professional translators, enterprises, web pages and freely available translation repositories.

Russian	English	
собака	dog	Last Update Usage Freq



Переводчик Write Словарь

русский ↔ английский

слова

Перевести текст
Перевести файлы
Улучшить текст

Wordfast Pro (WFP)

Wordfast Classic (WFC)

Wordfast Anywhere (WFA)

Wordfast Server (WFS)

Wordfast Aligner (WFL)

Главная
Решения ABBYY
Партнеры, поставщики
Помощь
Способы оплаты
Контакты

Поиск по каталогу

ИНФОРМАЦИЯ для ПОСЕТИТЕЛЕЙ САЙТА: Поставки лицензий, Каталог ПО размещен в справочных целях.

Каталог продуктов

ABBYY Aligner

ABBYY Aligner 2.0 – это удобный инструмент для выравнивания параллельных текстов и создания качественных баз Translation Memory. Программа находит соответствующие друг другу предложения в текстах на разных языках, сопоставляет их между собой и позволяет сохранить результат в базе Translation Memory или в файле формата RTF.

Варианты покупки

Спец. цены

Диаграмма сущностей

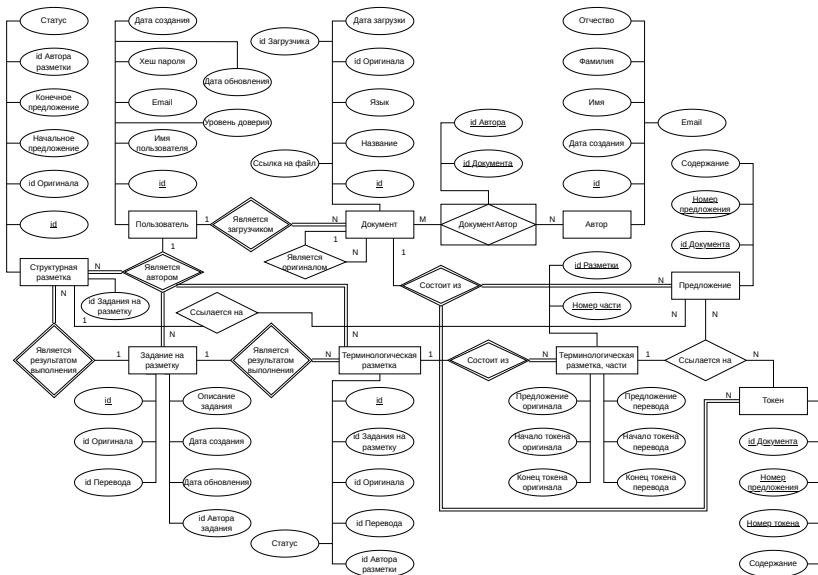


Диаграмма вариантов использования

- Администратор имеет полный доступ к данным: может добавлять, удалять и изменять тексты и разметки.
- Модераторы создают задания на разметку и производят проверку разметок, осуществленных пользователями.
- Пользователи осуществляют разметку и выполняют поиск по корпусу.

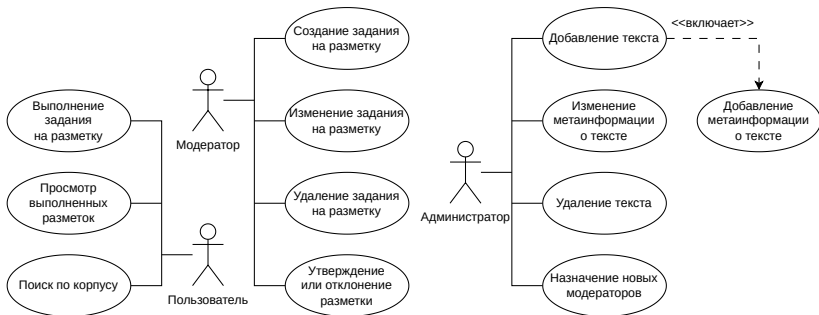


Диаграмма проектируемой БД

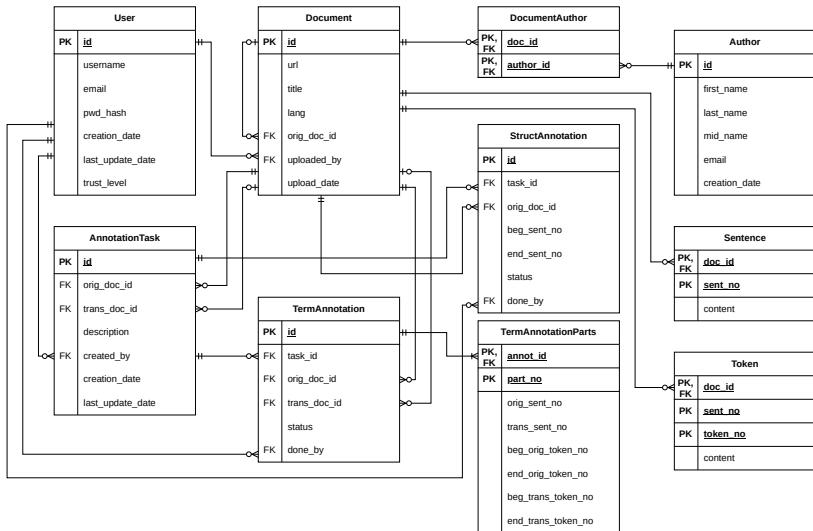
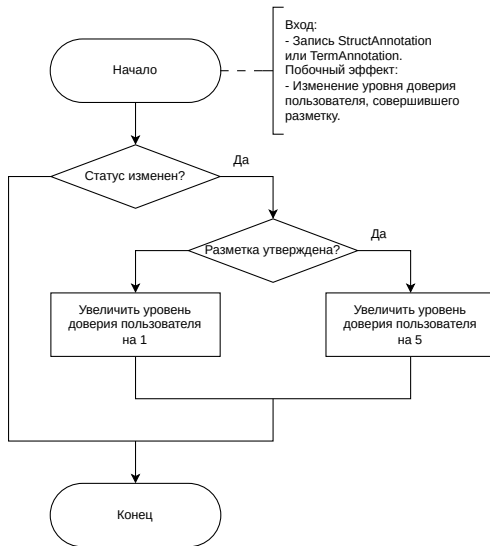


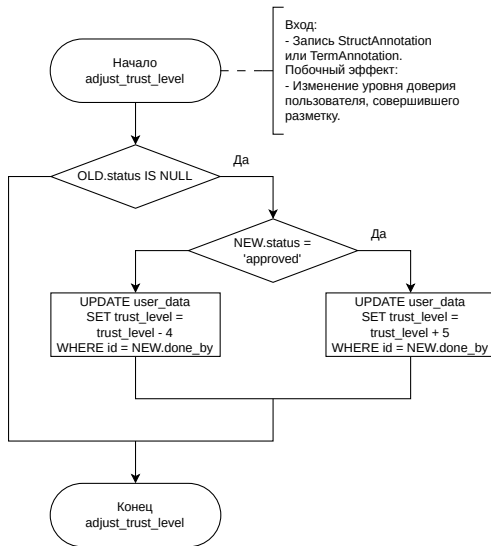
Схема проектируемого триггера



Была выбрана PostgreSQL по следующим причинам:

- Реализует реляционную модель, которая, как было выяснено, является наиболее подходящей для базы данных, разрабатываемой в настоящей работе;
- Открытый исходный код;
- Поддержка полнотекстового поиска;
- Имеется опыт работы с данной СУБД.

Схема реализованного триггера




Проведение исследования, locustfile.py

```
words = ["the", "be", "of", "and", "a", ...]
with open("document.ids", 'r') as file:
    document_ids = file.read().splitlines()

class UserBehavior(TaskSet):
    @task(2)
    def post_search(self):
        payload = {"content":
                    fake.random_element(words)}
        self.client.post("/search", json=payload)
    # ...

    @task(1)
    def get_document(self):
        id = fake.random_element(document_ids)
        self.client.get(f"/d/{id}")
```


Проведение исследования, интерфейс Locust

**Locust**HOST
http://localhost:8080

STATUS
READY

RPS
0

FAILURES
0%



Start new load test


Number of users (peak concurrency)
4000

Ramp up (users started/second)
100

Host
http://localhost:8080

Advanced options
Run time (e.g. 20, 20s, 3m, 2h, 1h20m, 3h30m10s, etc.)
40

START

**Locust**HOST
http://localhost:8080


STATUS
SPAWNING

USERS
3600

RPS
618.7

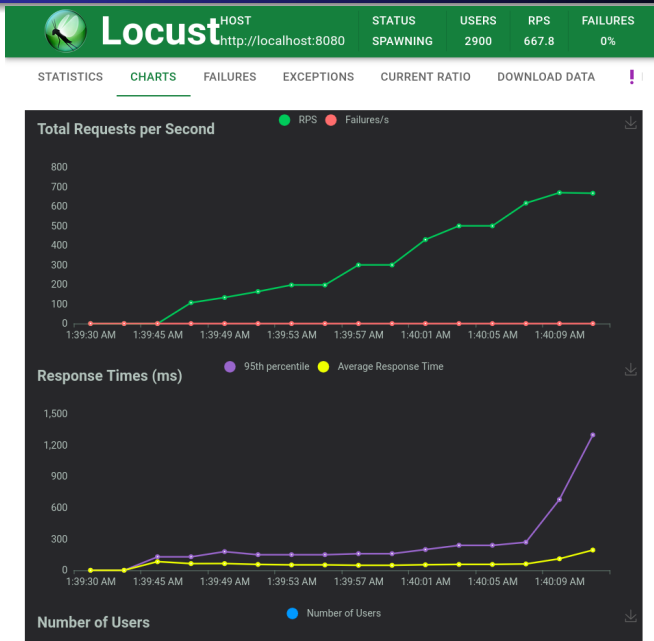
FAILURES
0%

STATISTICSCHARTSFAILURES EXCEPTIONSCURRENT RATIODOWNLOAD DATA

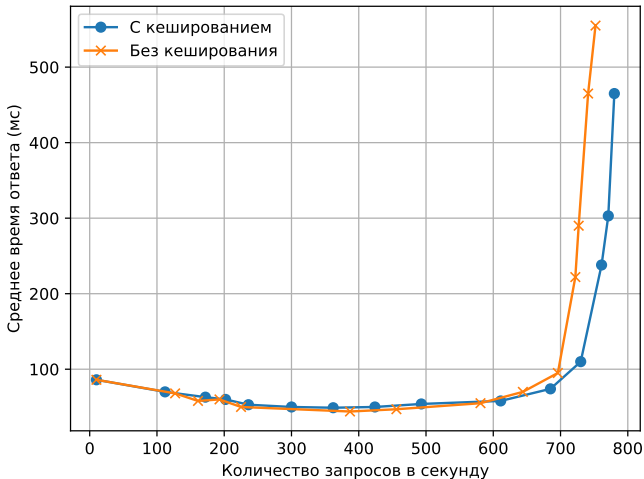


Type	Name	# Requests	95%ile (ms)	Average (ms)	Current RPS
GET	/d/003b2330-2ec2-4dfe-be84-484672b314c0	5	240	79.97	0
GET	/d/0121a78a-f2db-48bc-ab57-0b0a4b8eee24	2	850	426.47	0.1
GET	/d/0161a71d-ce1a-49f8-8a10-fdb65ff85065	3	1700	583.64	0
GET	/d/0170a098-c9cb-4e18-b8d5-9d50f26a2ad9	3	530	197.39	0.1
GET	/d/01ce6e6c-73c7-4eaf-832d-251ad31109ad	2	48	25.33	0
GET	/d/01d47687-ada1-472d-8ba0-4953de341cca	3	1600	680.7	0.1

Проведение исследования, интерфейс Locust

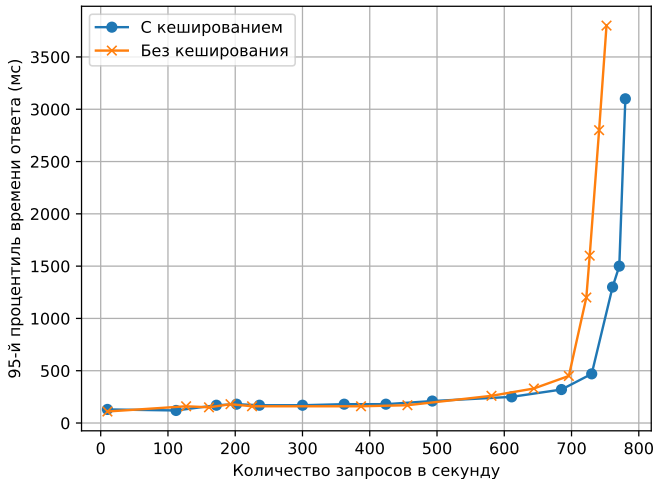


Результаты исследования



При 752 запросах в секунду использование кеша позволило ускорить среднее время ответа более, чем в 3 раза.

Результаты исследования



При 750 запросах в секунду 95 процентов запросов к приложению с кешем обрабатываются в 3.45 раза быстрее.

Поставленная цель: Разработка базы данных для автоматизации рабочего места разметчиков параллельного корпуса технических текстов была достигнута.

Для достижения цели были выполнены поставленные задачи:

- Проведен анализ предметной области корпусов текстов;
- Спроектирована и разработана база данных, описаны ее сущности, ограничения целостности, ролевая модель на уровне базы данных и используемые триггеры;
- Разработано приложение для доступа к базе данных;
- Исследована зависимость времени ответа от количества запросов в секунду.

В результате исследования было выяснено, что использование кеша позволяет ускорить среднее время ответа более, чем в 3 раза при большом (752) количестве запросов в секунду.

Далее на основе реализованных базы данных и приложения к базе данных можно сделать, например, следующее:

- Добавить поддержку загрузки PDF-документов;
- Интегрировать приложение с различными автоматическими выравнителями и разметчиками;
- Разработать приложение — социальную сеть для разметчиков текстов, в котором разметчики могут оперативно делиться разметками и выполнять задания, повышая свой рейтинг.