



Сибирский федеральный университет

часть данных скрыта в целях конфиденциальности



Разработка бота с функцией модерации голосовых чатов

Тема проекта | выпускной квалификационной работы



@3ndetz

Студент

Выпускник бакалавриата



Р. М. Е.

Научный руководитель

Старший преподаватель



А. О. А.

Консультант

Доцент, доктор технических наук

Содержание выступления



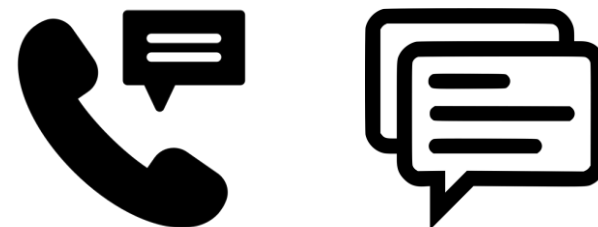
*Частичное внедрение дополнительной функции для определения оскорбительного поведения

Проблема и актуальность

- Мессенджеры — ключевой инструмент современной коммуникации
- Популярность голосовых чатов, особенно в мессенджере Discord
- **Небезопасная** среда общения для детей
- **Неэффективность** ручных методов модерации
- **Недостаточная** научная проработанность области → **отсутствие** выбора средств автоматизации

Объект и предмет

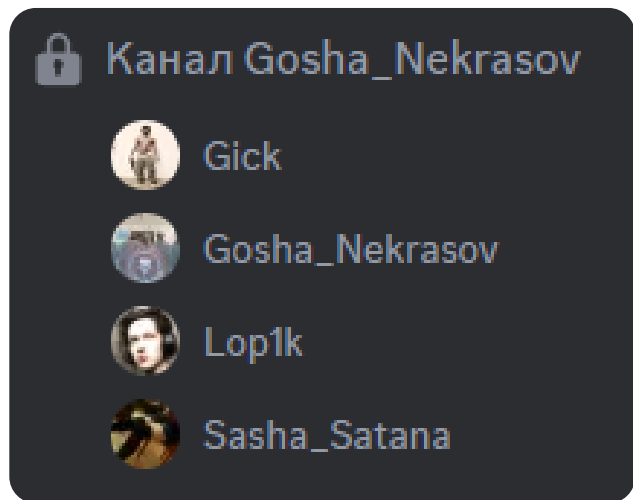
- **Объект** — автоматизированные системы администрирования и модерации для мессенджеров
- **Предмет** — бот с функцией модерации голосовых чатов



Коммуникация в мессенджерах может происходить разными способами

Цель работы

Разработать бота с функцией модерации голосовых чатов для мессенджера Discord



Пример голосового канала в мессенджере **Discord**

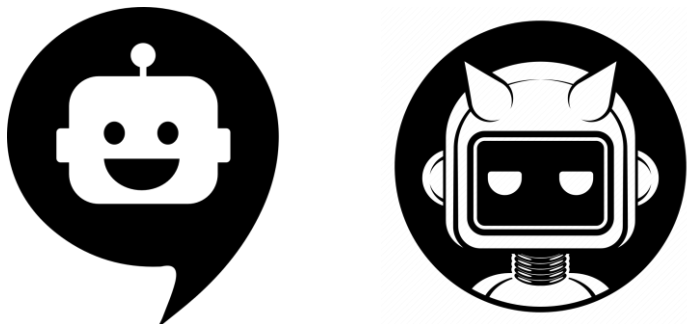
Задачи

- **Изучить** основные понятия, связанные с ботами, чатами в мессенджерах и их модерацией
- **Выявить особенности разработки** ботов для голосовых чатов
- **Спроектировать** архитектуру бота с функцией модерации голосовых чатов Discord
- **Разработать** и протестировать бота на основе выбранных библиотек и технологий

Определение бота

- **Бот** — программное приложение, выполняющее однообразные действия в интернете

Существуют и другие определения



Существуют как полезные, так и вредоносные боты

Классификация ботов

Вредоносные:

- Спам-боты
- Боты для кражи данных
- DDoS-боты
- Ботнеты

По способу взаимодействия:

- Автономные
- Командно-управляемые
- Визуально-интерактивные

По платформе:

- В социальных сетях
- На веб-платформах
- В моб. приложениях
- В мессенджерах

По функционалу:

- Информаторы
- Ассистенты
- Игровые
- Административные

Чат и мессенджер

- **Чат** — область виртуального пространства, используемая людьми для общения
- **Мессенджер** — коммуникационная система, использующая интернет

Правила общения в мессенджерах

- **Необходимость** в правилах **зависит** от числа участников чата
- Правила чатов не должны **противоречить** правилам платформы

Какими бывают чаты

По способу взаимодействия:

- Текстовые
- **Голосовые**
- Голосовые с возможностью передачи видео

По уровню безопасности:

- Защищенные
- Незащищённые

По числу пользователей:

- Личные (2 участника)
- **Групповые** (3 и более)

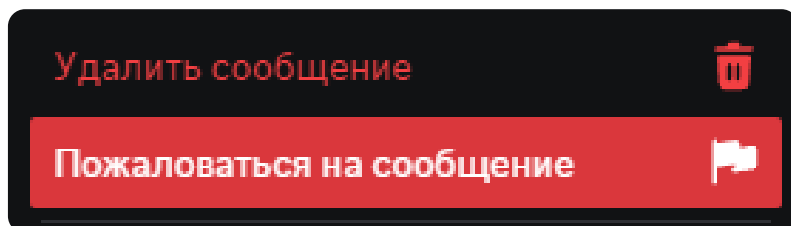
По способу доступа:

- Открытые
- Закрытые



Модерация чатов в мессенджерах

- За соблюдением правил чата следит его **администратор**
- Для контроля соблюдения правил администратор может как проверять чат вручную, так и использовать ботов
- Существует множество ботов для модерации **текстовых** чатов, **но не голосовых**
- За соблюдением правил мессенджера следит **администрация платформы**
- Для контроля соблюдения правил платформы типично использование **системы жалоб**



Виды недопустимого поведения

Каждый чат может иметь свои правила. Под **недопустимым поведением** каждый администратор может понимать разный набор признаков:

- Использование **нецензурных слов** и выражений (*легче отследить*)
- Оскорбительное поведение с использованием **нецензурной брани** (*легче отследить*)
- Оскорбительное поведение **без** использования **нецензурных слов** (*сложнее отследить*)



Выбор подходящей системы распознавания речи

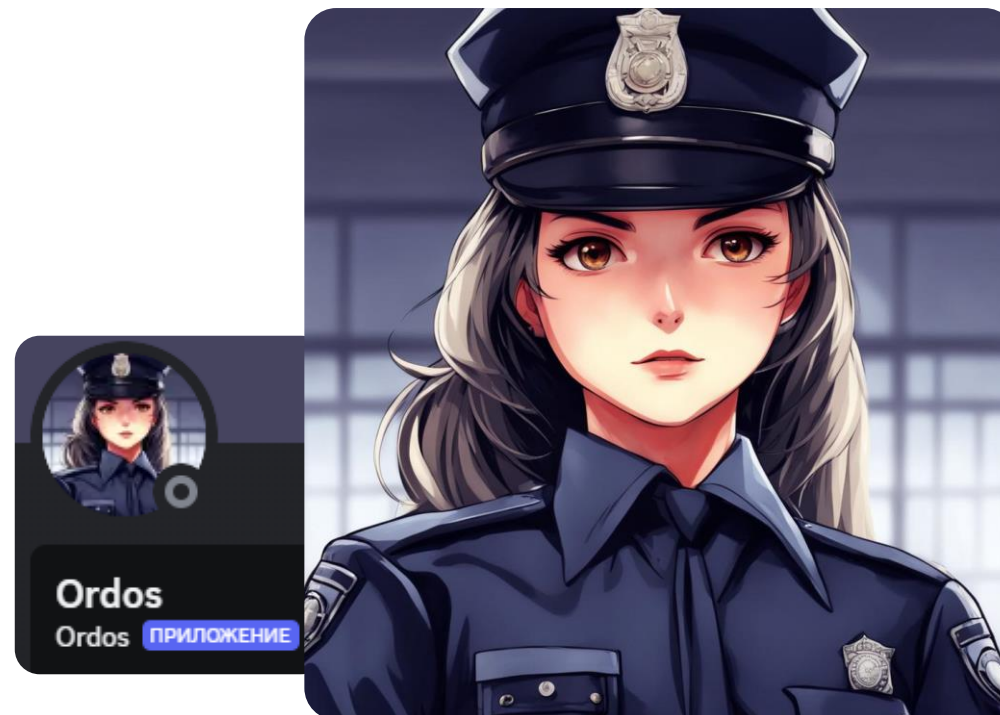
Назв. системы → Критерии	OpenAI Whisper Large	Salute Citrinet	Vosk Small	Vosk Big	Nvidia NeMo RNNT	Wav2Vec, FunASR
Качество	++	+	-	++	++	-
Скорость работы	-	++	++	-	++	-
Простота установки	++	-	++	++	-	+
Технические требования (убывание)	-	++	++	+	++	-

Разработка бота: план работы

1. **Формирование идеи**
2. **Проектирование**
3. **Написание программы**
4. **Тестирование**
5. **Развёртывание**

Формирование идеи

- Придумано название бота — Ordos, что схоже в написании с латинским словом «Ordo» — порядок.
- Концепция бота: распознавание речи в голосовых чатах, проверка на наличие нецензурных слов
- Создан профиль приложения бота в Discord



Профиль бота в Discord

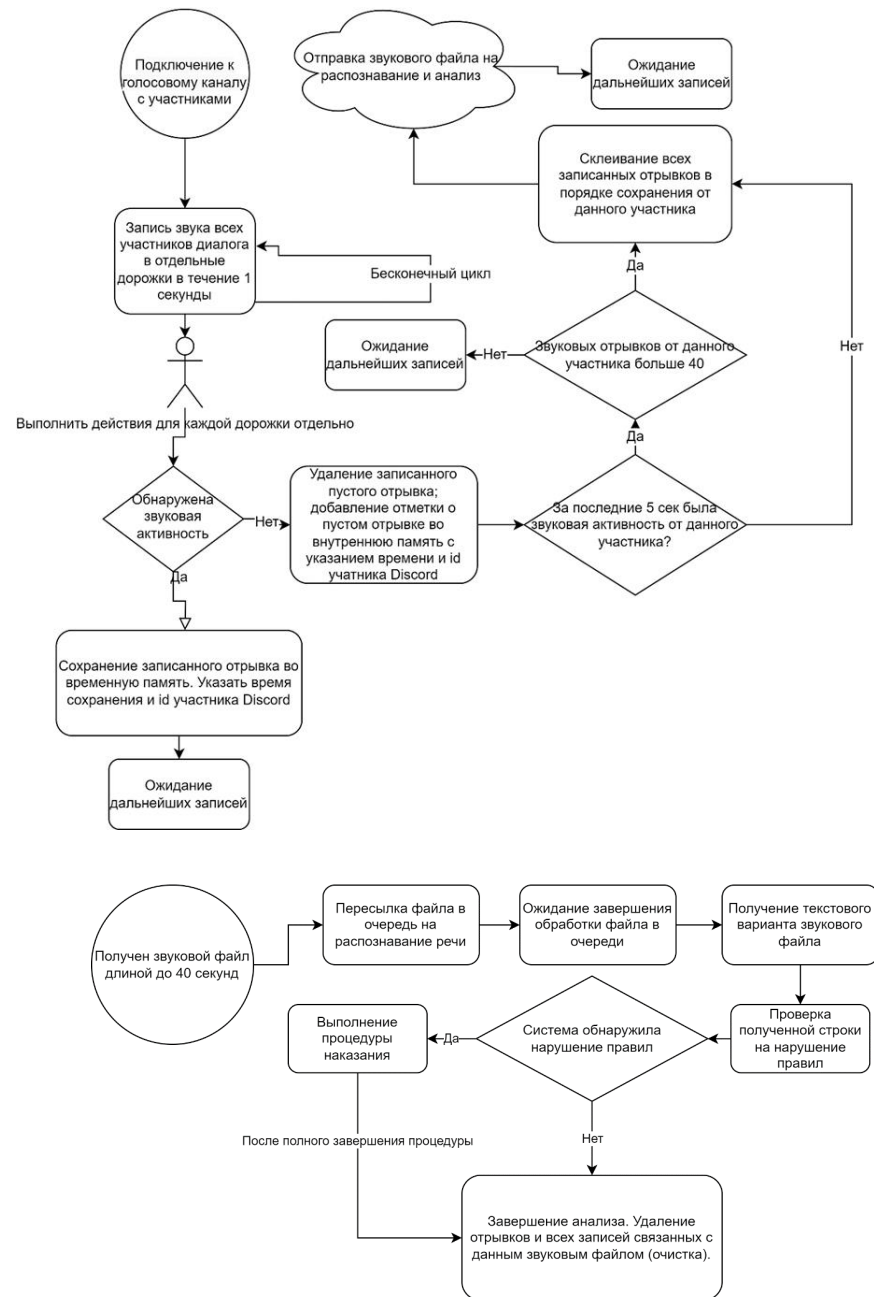
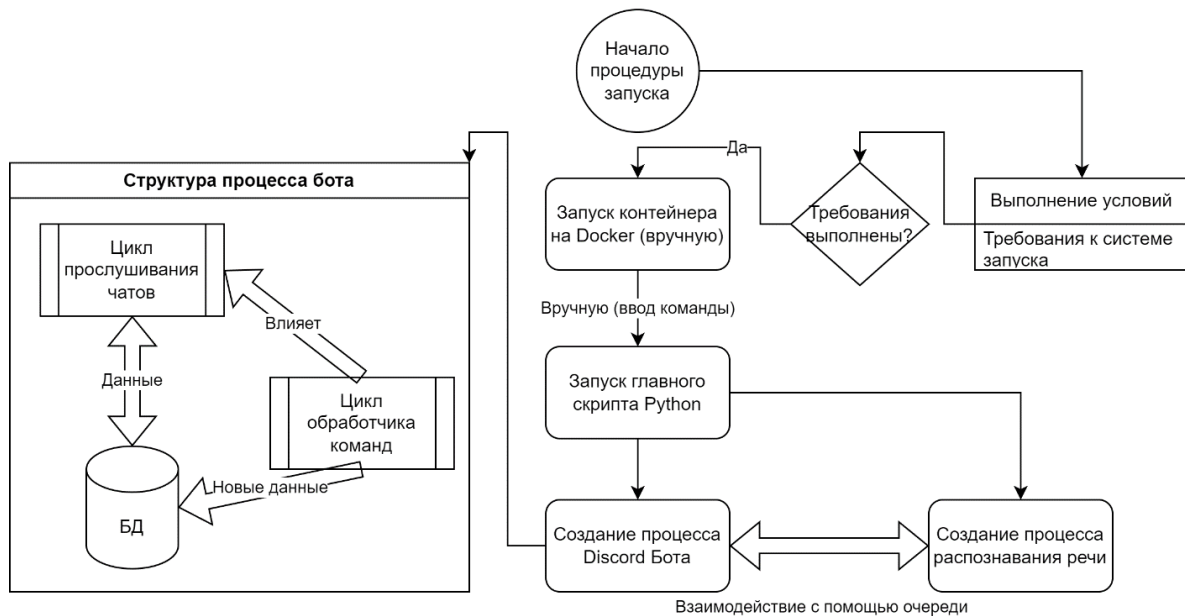
Иллюстрация создавалась при помощи нейросети

Kandinsky 3.0 по следующему запросу:

«Строгая девушка в полицейской форме крупным планом, стиль мультипликация»

Проектирование программной архитектуры бота — системы (превью)

- Система прослушивания голосовых чатов
- Система хранения данных
- Система анализа звуковых данных
- Связывающая система



Разработка и тестирование бота

- Подготовка: установка ПО
- Написание кода модулей
- Создание основной части программы
- Внутреннее тестирование

Написание программного кода модулей

- Система прослушивания голосовых чатов: **цикл прослушивания (Pycord AudioSink)**
- Система анализа звуковых данных: **модули распознавания речи, проверки текста речи, выдачи наказаний**
- Система хранения данных: **модуль** сохранения и загрузки из **приватных чатов** Discord

Связывающая система

- Связка модулей в рамках «тела» бота (в т.ч. связь с **Docker** через *Python SyncManager*)
- Вспомогательный механизм выдачи **ролей** Discord
- Вспомогательный механизм подтверждения **правил**
- Обработчики **команд** (команда помощи, изм. списка и др.)



Используемое ПО:

PyCharm, Docker, Nvidia NeMo

Идея

- Функция для определения **оскорбительного поведения** без использования нецензурных слов

Реализация

- Установка Python-пакета **Toxicity** для выявления токсичности в русскоязычных текстах с использованием модели ИИ
- Встройка метода **predict** в модуль проверки текста речи
- Высокопороговое срабатывание (при вероятности предсказания выше 98,5%)

Особенности

- Необходимость высокопороговой реализации для **минимизации** риска случаев **случайной блокировки** из-за особенностей ИИ модели.
- **Альфа-версия.** Необходимость последующей доработки функции и дополнительного тестирования для полноценного внедрения. *Модель ИИ — тяжёлый инструмент, а внедрение происходило уже после окончания проектирования.*

Accuracy	Recall	Precision	F1
90.6%	83.86%	87.24%	85.52%

Метрики выбранной модели ИИ для
определения токсичности

Скриншоты: настройка бота

Команда помощи

1. Подключение к голосовому каналу

чат-ordos

ТЕКСТОВЫЕ КАНАЛЫ

общее

игры

музыка

ГОЛОСОВЫЕ КАНАЛЫ

Комната отдыха



Разработчик



Разработчик

@Ordos помощь



Ordos ПРИЛОЖЕНИЕ

Данный бот создан с целью автоматической модерации голосовых чатов и пресечения использования запрещенной лексики на вашем сервере Discord. Список запрещенных слов можно изменить, добавить собственные. На данный момент бот может модерировать не более 1 голосового чата на сервере.

Доступные команды:

help	Вызвать окно помощи.
включить	Включить автоматическую модерацию голосового канала.
документация	Максимально объемное описание команд этого бота.
настроить	Настройте голосовой канал!
отключить	Выключить автоматическую модерацию голосового канала.
поведение	Включить/отключить проверку речи на оскорбительное поведен...
режим	Изменить режим (лояльный или стандарт).
список	Изменить список плохих слов для модерации.

Все команды вводятся с помощью @Ordos название_команды [аргументы] (если они есть, без квадратных скобок). Введите команду @Ordos помощь название_команды для получения большей информации о команде.



Разработчик

@Ordos настроить

2. Вызов команды настройки



Ordos ПРИЛОЖЕНИЕ

Голосовой чат настроен – теперь модерироваться будет голосовой чат с названием Комната отдыха.

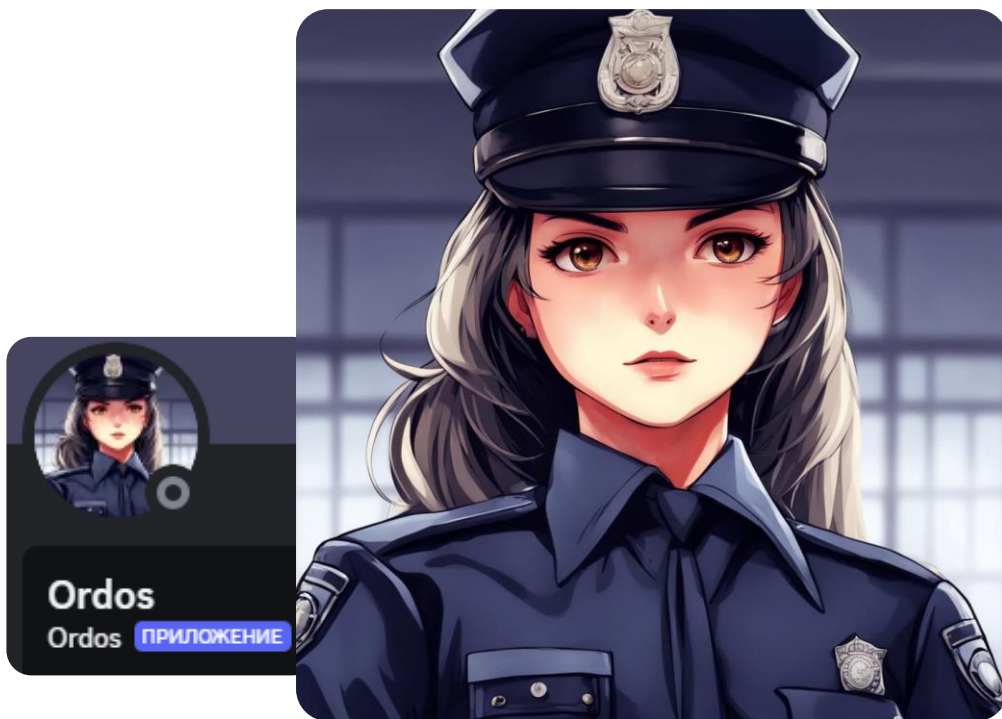
Отключить/включить модерацию вы можете при помощи команд «отключить» или «включить» соответственно.

3. Настройка завершена!

Голосовая связь под
Комната отдыха / Ordos se...

Подведение итогов

Выполнена поставленная цель —
разработан бот с функцией модерации
голосовых чатов для мессенджера Discord



Особенности

- Модульная, распределяемая архитектура приложения
- Распознавание речи качественными инструментами, использующими методы ИИ
- Функция для определения **токсичного поведения**, использующая методы ИИ

Перспективы

- Масштабирование бота под более серьёзную нагрузку
- Адаптация бота для других платформ и языков
- Возможно создание системы предсказания нарушений (при анализе звука напрямую: тональность, эмоциональность и т.д.)

Конец основной части. Далее представлены дополнительные материалы

Благодарю за внимание

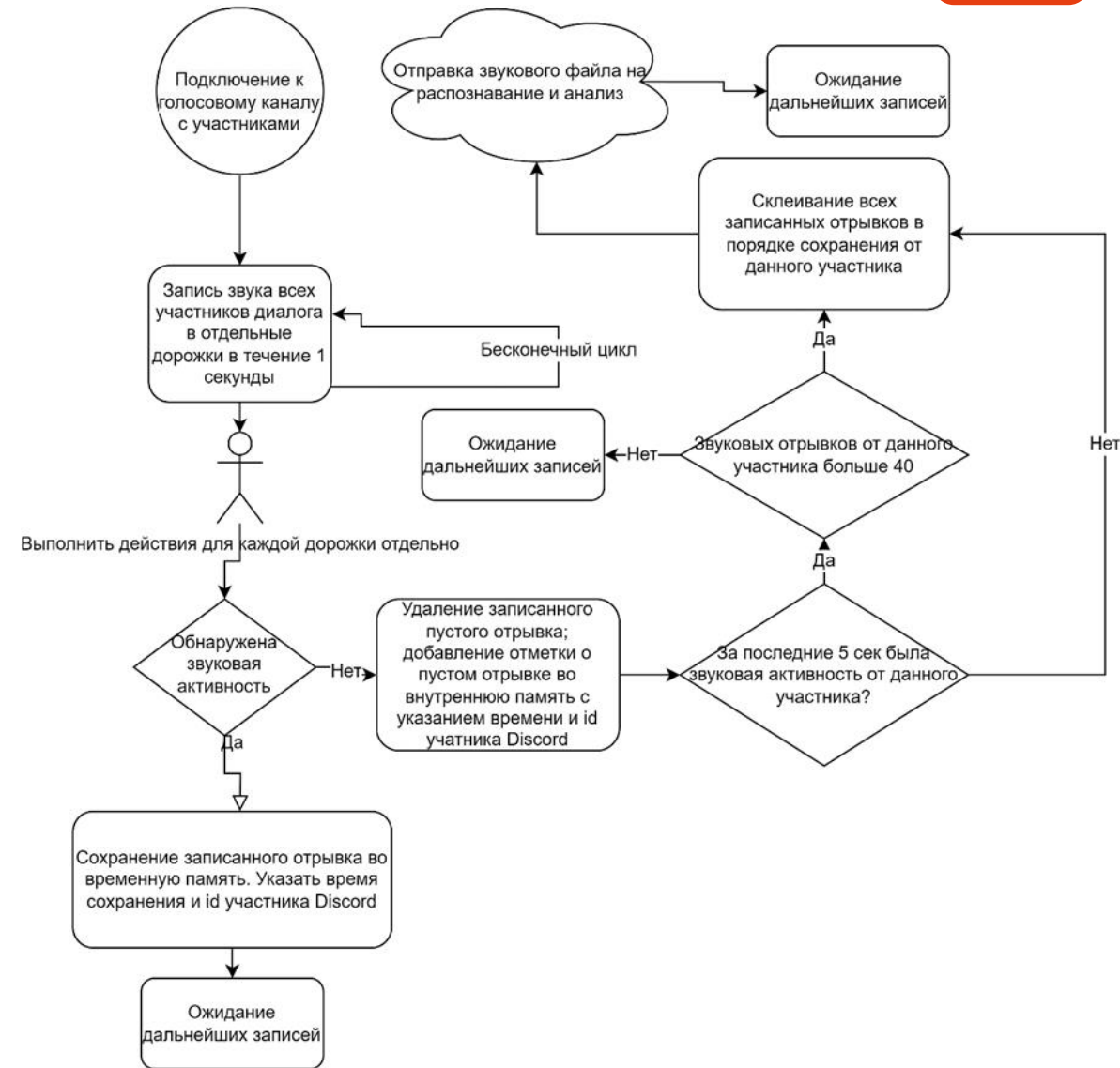


Проектирование программной архитектуры бота — системы:

- Система прослушивания голосовых чатов
- Система хранения данных
- Система анализа звуковых данных
- Связывающая система

Система прослушивания голосовых чатов

- Отвечает за прослушивание ботом всех модерлируемых голосовых чатов
- Функционирует постоянно
- Анализирует данные от каждого пользователя по отдельности
- Взаимодействие с Discord через **Py discord**



Условная схема системы прослушивания
голосовых чатов

Проектирование программной архитектуры бота — системы:

- Система прослушивания голосовых чатов
- Система хранения данных
- Система анализа звуковых данных
- Связывающая система

Система анализа звуковых данных

- Отвечает за распознавание речи
- Проверяет речь на соответствие правилам чата
- Исполняет процедуру наказания

Распознавание речи

- Фреймворк Nvidia NeMo на Docker



Условная схема обработки звуковых данных

Проектирование программной архитектуры бота — системы:

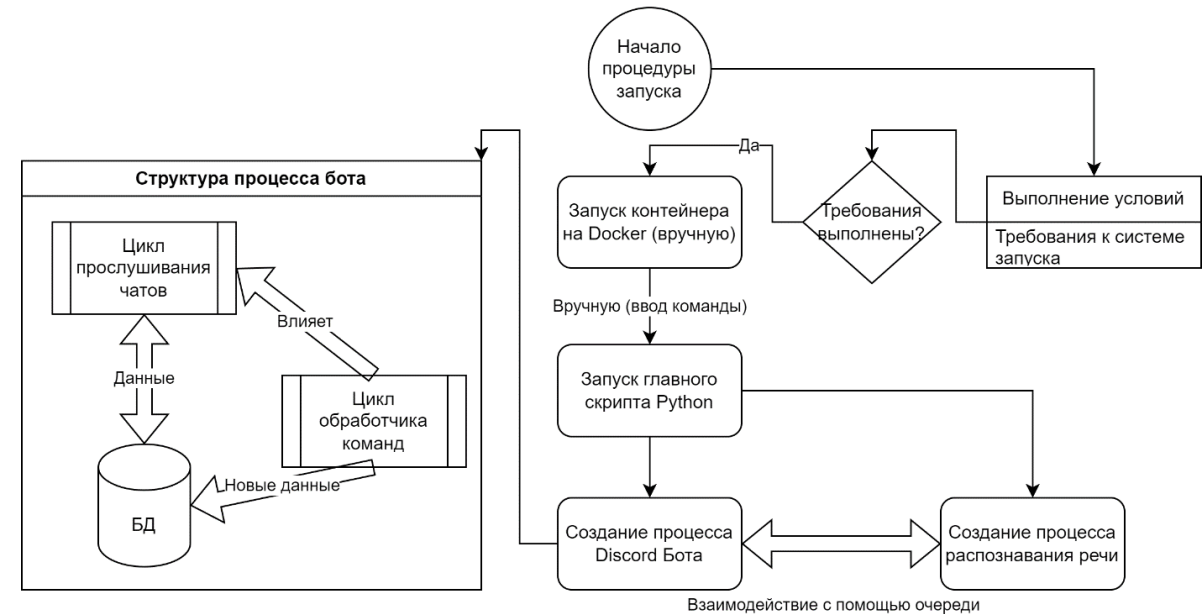
- Система прослушивания голосовых чатов
- Система хранения данных
- Система анализа звуковых данных
- **Связывающая система**

Система хранения данных

- Решено хранить настройки бота в приватных чатах сообществ Discord, без базы данных
- Прозрочно и просто

Связывающая система

- «Тело» бота, выполнение команд
- Взаимодействие с другими системами



Условная схема общей программной структуры бота


Детектор запрещённых слов

```
def detect_badwords(text: str, badwords: list = None) -> dict:
    if badwords is None:
        badwords = base_badwords_list
    words = str_to_words(text)
    result = {"found": 0.0}
    for word in words:
        if word in badwords:
            result["found"] = 1.0
            result["word"] = word
            break

    return result
```

Детектор токсичного поведения

```
def toxic_analyze(text: str, result: dict = None) -> dict:
    try:
        if result is None:
            result = {}
        result["predict"] = toxicDetector.predict([text])[0]
        # predict analyze, front
        if result["predict"] > 0.985:
            result["found"] = 1.0 # перезаписываем результат
            result["word"] = "Недопустимое поведение"
    except BaseException as err:
        print('DEBUG ошибка работы ToxicDetector, err=', err)
    return result
```



Ordos

ПРИЛОЖЕНИЕ



06.05.2024 19:45


!

Для того, чтобы использовать модулируемый голосовой чат "Комната отдыха" на сервере "Ordos сервер", вам нужно принять следующее соглашение об обработке вашей речи ботом:


Текст соглашения об обработке персональных данных

Ваши голосовые данные будут обработаны для распознавания речи и автоматической проверки её на соответствие правилам этого сервера. Система хранит данные разговоров не дольше минуты, затем они будут удалены. Принимая данное соглашение работы с ботом, вы подтверждаете свое согласие на обработку персональных данных вашего голоса из этого голосового канала. Данные ваших разговоров не будут переданы третьим лицам; они хранятся не более минуты, после чего подлежат удалению.

Если вы согласны и принимаете это соглашение, поставьте, пожалуйста, на это сообщение реакцию зеленой галочки , (реакция уже стоит, просто жмакните её). После этого вы сразу сможете присоединиться к чату 

 2

3. Реакция для подтверждения

Спасибо! Вы подтвердили соглашение и допускаетесь к чату "Комната отдыха" сервера "Ordos сервер"! 

2. Отключение и оповещение

1. Попытка подключения к ГЧ

музыка

ГОЛОСОВЫЕ КАНАЛЫ

Комната отдыха

Разработчик

Комната для трансляций

РОЛИ

Верифицирован в ГС ботом

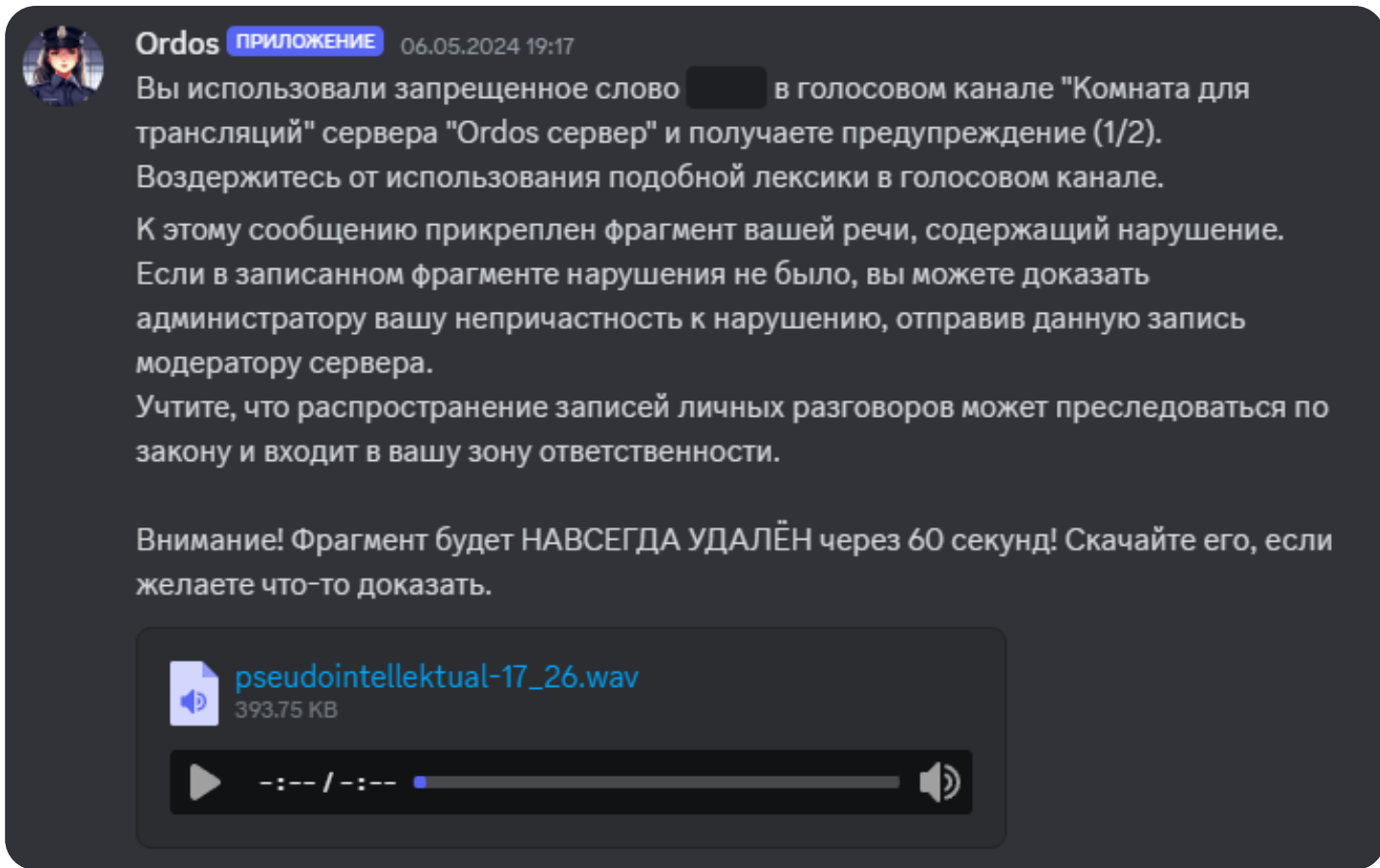
ЗАМЕТКА

Нажмите, чтобы добавить заметку

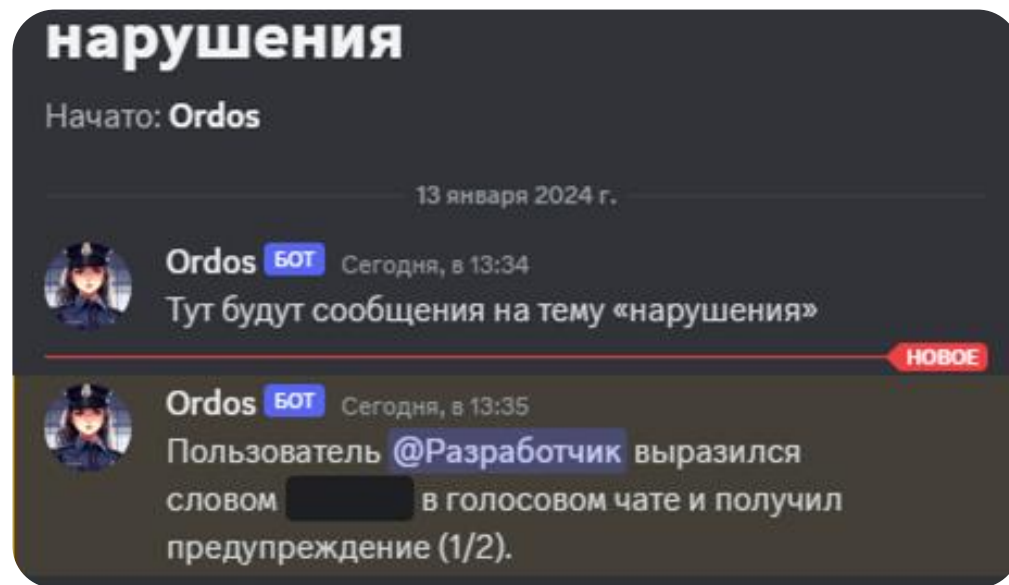
4. Выдача роли для допуска

Окно подтверждения правил

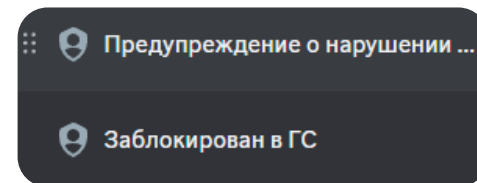
Дополнительные материалы



1. Оповещение нарушителя



2. Оповещение администрации



3. Выдача соответствующих «ролей» доступа

! Вы были заблокированы за нарушение в модулируемом голосовом чате "Комната отдыха" на сервере "Ordos сервер" и не можете присоединиться. Ожидайте снятия блокировки 😞

При попытке подключения к ГЧ

Дополнительные материалы

Разработчик Сегодня, в 13:48

@Ordos список свой крик, электростанция, катер

Ordos БОТ Сегодня, в 13:48

Успешно! Установлен новый список из 3 слов.

Разработчик Сегодня, в 13:50

@Ordos список добавить кружка

Ordos БОТ Сегодня, в 13:50

Успешно! Добавлено 1 новых слов.

Разработчик Сегодня, в 13:51

@Ordos список добавить кружка

Ordos БОТ Сегодня, в 13:51

Неправильно заданы слова: Новых слов для добавления нет.

Изменение списка запрещённых слов

@Ordos отключить

Ordos БОТ 11.01.2024 11:56

Модерация голосового канала отключена.

Разработчик 11.01.2024 11:57

@Ordos включить

Ordos БОТ 11.01.2024 11:57

Модерация голосового канала включена.

Включение модерации

Разработчик 11.01.2024 13:44

@Ordos режим лояльный

Ordos БОТ 11.01.2024 13:44

Режим лояльный установлен.

Настройка режима модерации