

# VK-GPT

Делали:

Коваленко Артём Владимирович, 10 класс, Красноярский Край, Лидер команды, Программист

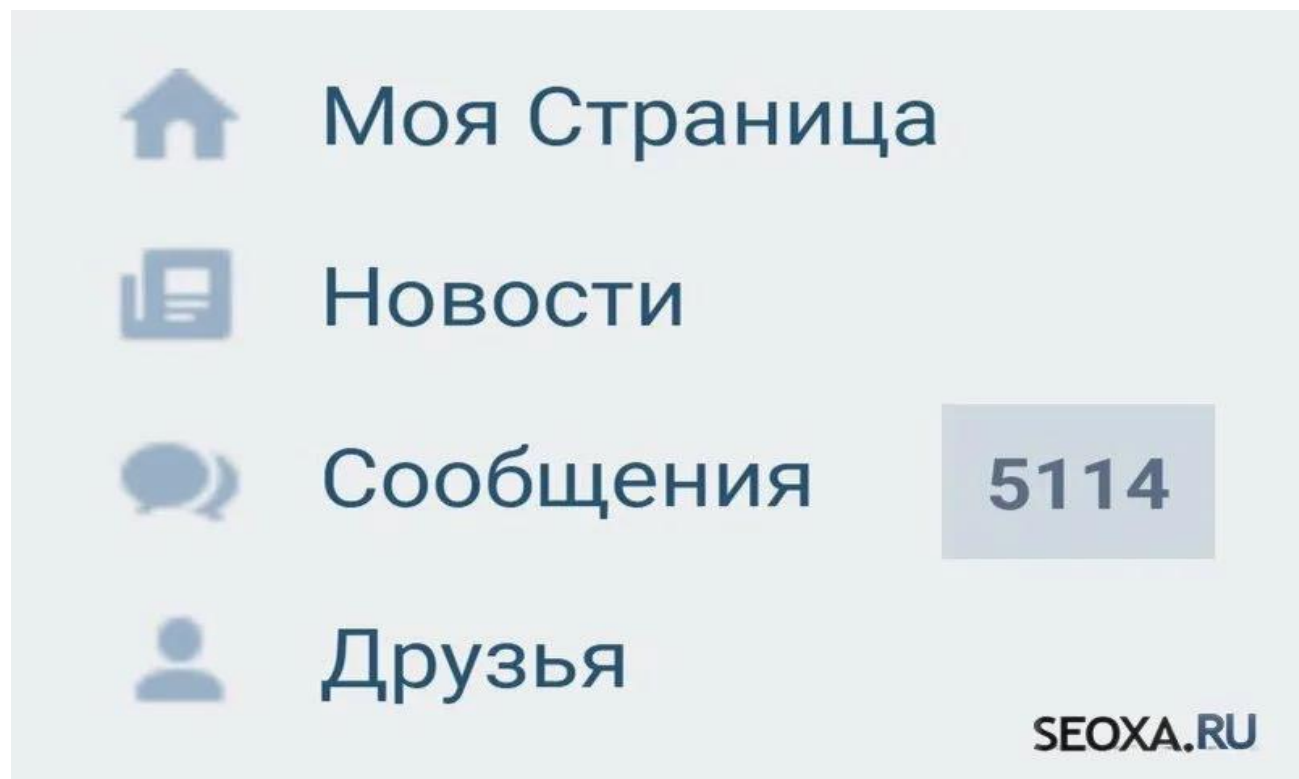
Балкин Георгий Олегович, 10 класс, Красноярский Край, Дизайнер, Программист

Савостьянов Георгий Юрьевич , 10 класс, Красноярский Край, Координатор, Программист

Тимаков Илья Викторович , 10 класс, Красноярский Край, Программист

# Проблема

Добрый день, уважаемые жюри! Сегодня я хотел бы обратить ваше внимание на актуальную проблему, связанную с большим количеством сообщений в мессенджерах. Современные коммуникационные средства позволяют нам легко общаться и обмениваться информацией, но иногда большое количество поступающих сообщений может вызывать множество проблем у пользователя. Давайте рассмотрим некоторые из них:



# Проблема

- 1) Информационный шум: Большое количество сообщений может привести к информационной перегрузке, что затрудняет отслеживание и нахождение нужной информации. Это может привести к пропуску важных сообщений или недопониманию.
- 2) Отвлекающий фактор: Постоянно поступающая информация в новых сообщениях может отвлекать от текущих задач и снижать производительность, когда нет сил и времени для прочтения всех сообщений разом.
- 3) Потеря контекста: Большое количество сообщений может затруднить следить за хронологией и контекстом беседы. Это может привести к недопониманию и потере важной информации.
- 4) Утомление и стресс: Большое количество поступающей информации может привести к перегрузке мозга, быстро утомляя организм и затрудняя запоминание полученной информации.
- 5) Проблема занятости: Не каждый человек может себе позволить читать тысячи сообщений в течение множества часов для получения ценной информации, необходимой для работы.

# Методология

Существуют различные подходы к решению задачи суммаризации сообщений. Вот некоторые из них:

## 1. Извлечение ключевых фраз:

Этот подход основан на выделении наиболее информативных фраз или предложений из текста и их объединении в краткую сводку. Обычно используются методы анализа частотности слов, статистические методы и алгоритмы машинного обучения для определения ключевых фраз. Этот подход прост в реализации и обладает хорошей интерпретируемостью, но может упускать некоторую важную информацию, не учитывая контекст и связи между предложениями.

## 2. Генеративные модели:

Генеративные модели используют нейронные сети для генерации краткой сводки. Они обучаются на больших наборах данных, где каждая пара входных данных и сводки является обучающим примером. Такие модели могут учитывать контекст и связи между предложениями, но требуют больших вычислительных ресурсов и объемных наборов данных для обучения.

## 3. Абстрактные модели:

Абстрактные модели генерируют сводку, используя абстрактное представление текста, отображая его в новую форму, которая содержит основную информацию. Этот подход позволяет генерировать более естественные и краткие сводки, сохраняя контекст и связи между предложениями. Однако, такие модели сложнее в реализации и требуют больше вычислительных ресурсов.

# Сравнительный анализ

Сравнительный анализ этих решений:

1. Извлечение ключевых фраз: Преимущества: простота в реализации, хорошая интерпретируемость.

Недостатки: упускание важной информации, отсутствие учета контекста и связей между предложениями.

2. Генеративные модели:

- Преимущества: учет контекста и связей между предложениями, способность генерировать сводки с высокой информативностью.

- Недостатки: требуют больших вычислительных ресурсов, объемных наборов данных для обучения.

3. Абстрактные модели:

- Преимущества: генерация естественных и кратких сводок, сохранение контекста и связей между предложениями.

- Недостатки: сложность в реализации, требование больших вычислительных ресурсов.

Выбор конкретного подхода зависит от требований и ограничений задачи суммаризации сообщений. Если важна простота и интерпретируемость, то извлечение ключевых фраз может быть предпочтительным. Если требуется учет контекста и связей между предложениями, генеративные или абстрактные модели могут быть более подходящими. Однако, необходимо учитывать доступные вычислительные ресурсы и объемы данных для обучения моделей.

# Описание предлагаемого решения.

Для решения данной проблемы наша команда решила использовать генеративную модель gpt3.5-turbo-16k . Как описывалось ранее, генеративные модели отлично справляются с задачами суммаризации текста, при этом требуя меньшее количество ресурсов, чем абстрактные модели.



# Сравнительный анализ имеющихся решений

Сравнительный анализ модели ru-gpt3.5-turbo-16k с другими моделями, решающими задачу суммаризации:

## 1. ru-gpt3.5-turbo-16k vs. ru-gpt3-large:

- Обе модели основаны на генеративных трансформерах и способны генерировать сводки текста.
- Преимущества ru-gpt3.5-turbo-16k: ru-gpt3.5-turbo-16k является более компактной и экономичной моделью по сравнению с ru-gpt3-large, но при этом сохраняет хорошее качество генерации текста.
- Преимущества ru-gpt3-large: ru-gpt3-large имеет больший размер и может обладать некоторыми преимуществами в качестве генерации сводок, особенно при работе с большими объемами данных.

## 2. ru-gpt3.5-turbo-16k vs. BART (Bidirectional and Auto-Regressive Transformers):

- BART является мощной моделью для суммаризации текста, использующей двунаправленные и авторегрессивные трансформеры.
- Преимущества ru-gpt3.5-turbo-16k: ru-gpt3.5-turbo-16k обучена на русскоязычных данных и специфична для русского языка, что может быть полезно для задач на русском языке.
- Преимущества BART: BART имеет широкую поддержку и доступны предобученные модели для нескольких языков, включая английский.
- Обе модели способны генерировать качественные сводки, учитывая контекст и связи между предложениями. Выбор между ними может зависеть от конкретных требований и доступных ресурсов.

# Сравнительный анализ

## 3. ru-gpt3.5-turbo-16k vs. Extractive Summarization (извлечение ключевых фраз):

- Extractive Summarization подходит для простоты и интерпретируемости, поскольку он основан на извлечении наиболее информативных фраз из текста.
- Преимущества ru-gpt3.5-turbo-16k: ru-gpt3.5-turbo-16k способна учитывать контекст и связи между предложениями, что может привести к более информативным сводкам.
- Преимущества Extractive Summarization: Extractive Summarization прост в реализации и позволяет явно видеть, какие фразы были извлечены в сводку.
- Оба подхода имеют свои преимущества и недостатки. Выбор между ними зависит от требований задачи, интерпретируемости и доступных ресурсов.

В целом, модель ru-gpt3.5-turbo-16k представляет собой компактную и экономичную модель, обладающую хорошим качеством генерации текста. Она специфична для русского языка и может быть полезна для задач на русском языке. Выбор между ru-gpt3.5-turbo-16k и другими моделями зависит от конкретных требований, доступных ресурсов и особенностей языка, так что мы выбрали именно эту модель за её эффективность, скорость и простоту использования.



# Высокоуровневый план реализации решения

1. Сбор данных: Необходимо собрать достаточное количество данных, которые будут использоваться для обучения модели суммаризации. Это могут быть различные тексты, сообщения или диалоги, которые содержат информацию, подлежащую суммаризации.
2. Предобработка данных: Перед обучением модели данные нужно предобработать. Этот шаг может включать удаление шума, очистку текста от лишних символов и структурирование данных в формат, пригодный для обучения модели.
3. Обучение и настройка модели: С использованием предобработанных данных нужно обучить модель суммаризации. Это может включать выбор и настройку алгоритма обучения, определение гиперпараметров и запуск процесса обучения на вычислительных ресурсах. Может потребоваться настройка гиперпараметров или проведение дополнительных итераций обучения для достижения желаемых результатов.
4. Интеграция в систему чатов: После успешного обучения и настройки модели, ее можно интегрировать в систему чатов. Это может включать разработку API или интерфейса, который позволит взаимодействовать с моделью суммаризации и получать сводки текста на основе входных сообщений.
5. Тестирование и оптимизация: После интеграции модели в систему необходимо провести тестирование для проверки ее работоспособности, производительности и качества сводок. При необходимости можно внести дополнительные оптимизации для улучшения результатов.
6. Развитие и обновление: Решение суммаризации сообщений требует постоянного развития и обновления. Необходимо следить за новыми исследованиями и методиками в области суммаризации текста и вносить соответствующие улучшения в модель и систему чатов.

Важно отметить, что конкретные шаги и подходы могут варьироваться в зависимости от требований и контекста задачи. План реализации должен быть адаптирован под конкретные потребности и ресурсы проекта.

# Описание отдельных шагов плана.

- 1) Изучена и проанализирована решаемая проблема.
- 2) Найдены модели, способные решить задачи сумаризации.
- 3) Создан генератор случайных диалогов.
- 4) Найден и протестирован способ удаления шума в сообщениях для лучшей суммаризации текста.
- 5) Проведено обучение моделей на заранее отобранных тренировочных диалогов ( 3 шт.)
- 6) Проведён сравнительный анализ обученных моделей, отобрана лучшая.
- 7) Создание приложения для генерации диалогов и их суммаризации.



# ИТОГИ

В результате выполнения проекта была достигнута цель - создание системы для суммаризации диалогов . Это решение может быть использовано для автоматической генерации сводок текста из диалоговых сообщений в различных областях, таких как чат-боты, анализ социальных медиа и другие приложения, где требуется обработка и суммаризация текстовых данных.