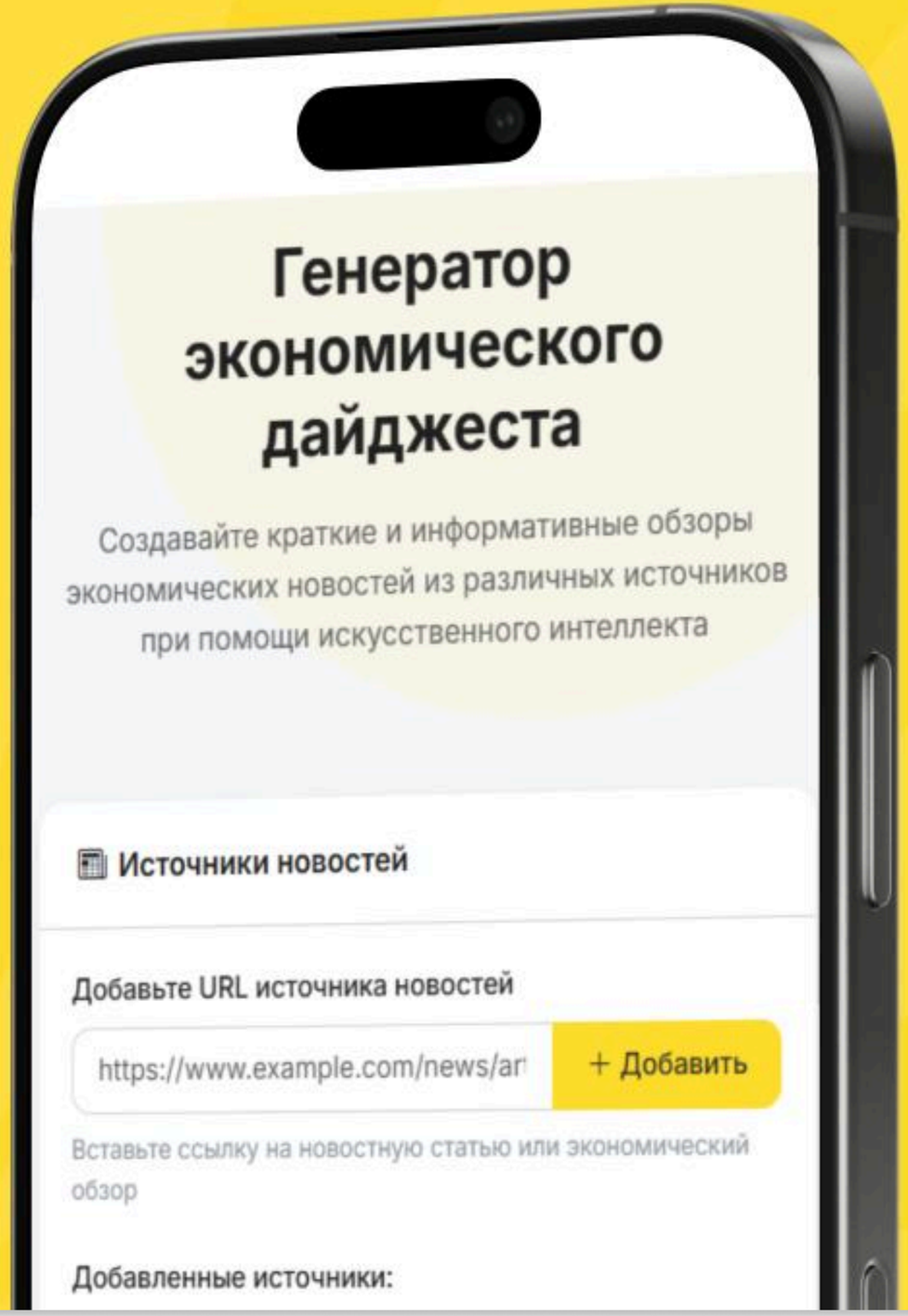
T BANK

ФуфелДайджест

Бот-саммаризатор инвест-новостей

*Признан иноагентом на территории РФ

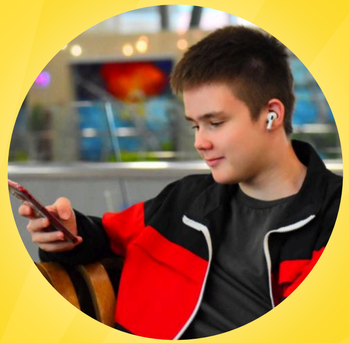


Команда

Алгоритм суммаризации, обработка изображений в тгк, стилизация текст, Web-App

Парсинг тг-каналов, исключение рекламы из постов

Создание презентации, оформление гита, дизайн Web-App

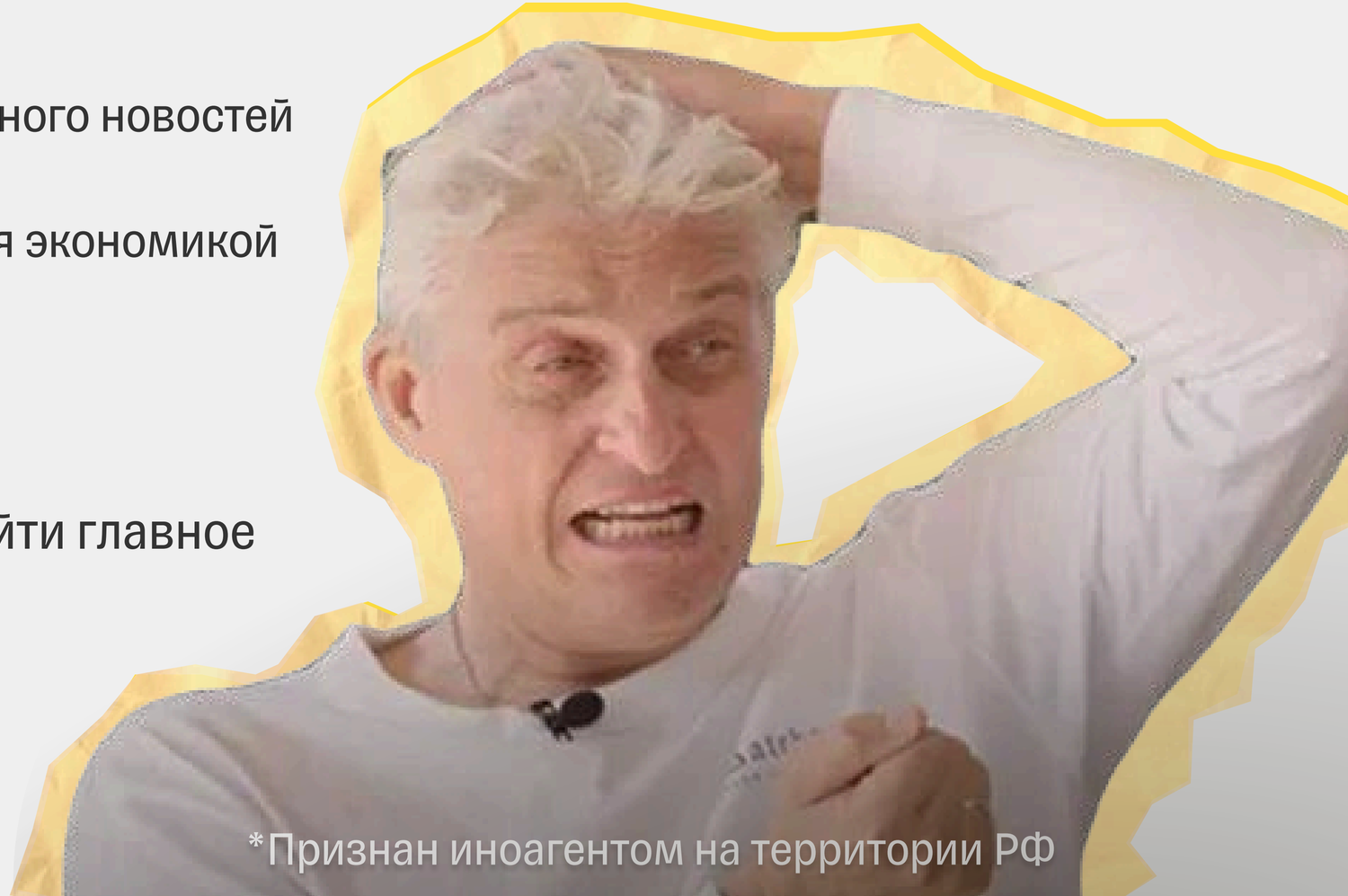


Разработка тг бота, работа с БД, Web-App

Проблема

Каждый день в Telegram выходит очень много новостей про экономику и инвестиции.
Инвесторы, аналитики и интересующиеся экономикой пользователи — не успевают всё читать

Много времени уходит на то, чтобы найти главное и важное среди кучи текста



*Признан иноагентом на территории РФ

Анализ области

	@summary_of_channels_bot	ФуфелДайджест	@searchforchatsbot
Описание	Бот для создания саммари из сообщений в каналах. Пользователь оставляет ссылку на публичный канал.	Бот-саммаризатор автоматически присылает в чат/канал краткую сводку топ-N экономических новостей из заданных источников с нужной периодичностью.	Бот помогает собрать много каналов и чатов из нужной тематики. Работает через встроенный поиск в телеграм
Специализация под экономические новости	✗	✓	✗
Краткий анализ и прогноз	✗	✓	✗
Персанолизация саммари	✗	✓	✗

Цель

Разработать тг-бота, который добавляется в чат и выдает суммаризацию постов в инвест-каналах

Задачи

1. Выбор технологического стека
2. Разработка саммаризации, персонализация по стилям
3. Реализовать парсер Telegram-каналов
5. Ранжирование новостей
4. Создание тг Web-App приложения
6. Тестирование системы
7. Обертка системы в докер

*Признан иноагентом на территории РФ

Функциональные требования

- Настройка количества новостей, выдаваемых ботом
- Парсинг каналов, данных пользователем
- Стилизация саммари
- Отсеивание рекламы
- Хранение информации в БД
- Обертка в Docker

План

08 апреля **09**

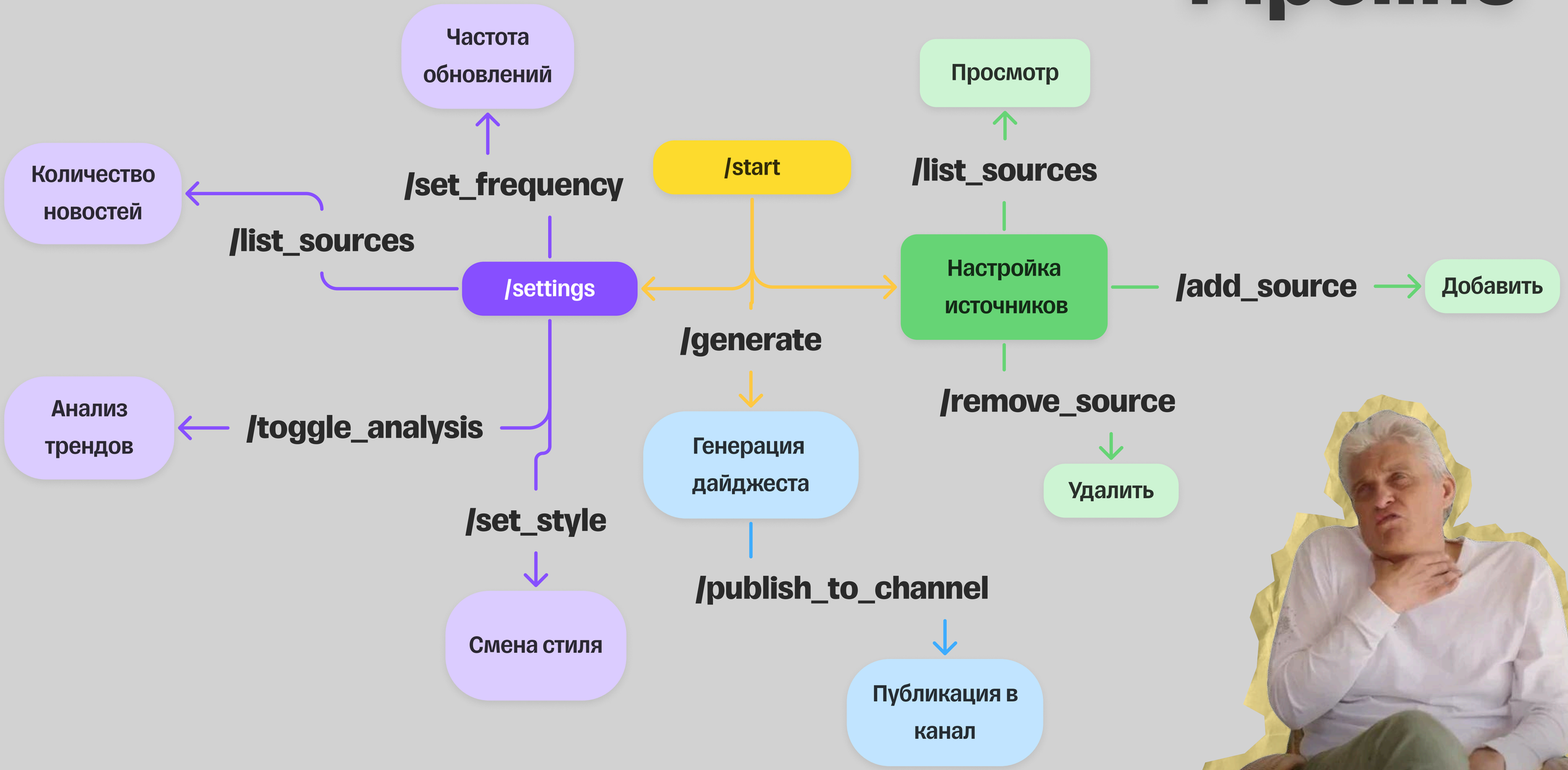
1. Анализ требований и постановка задачи
2. Выбор технологий
3. Реализация системы парсинга
4. Реализация алгоритма саммаризации и ранжирования контента
5. Разработка Telegram-бота

1. Стилизация саммари
2. Обработка изображений в постах
3. Улучшение пользовательского опыта
4. Создание Web-App приложения
5. Создание докера

Стек технологий

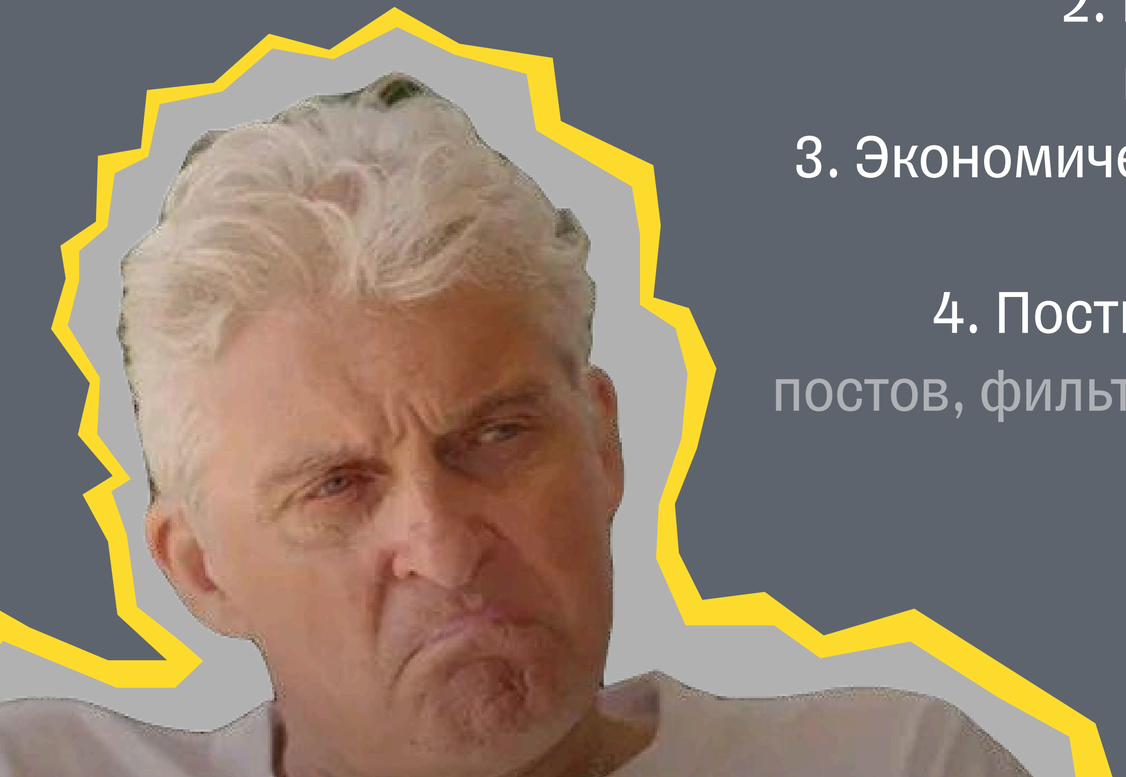
1. **Asyncio + aiohttp** для быстрого парсинга множества Telegram-каналов одновременно
2. **MongoDB** для гибкого хранения пользовательских источников и настроек
3. **Mistral API** для семантического анализа новостей (эмбеддинги, похожие посты, тематическая фильтрация) без локальных ML-моделей
4. **Docker** для удобства запуска нашего решения

Pipeline


Частота обновлений
/set_frequency
Количество новостей
/list_sources
/settings
/start
/generate
Генерация дайджеста
/publish_to_channel
Публикация в канал
Анализ трендов
/toggle_analysis
/set_style
Смена стиля
Просмотр
/list_sources
Настройка источников
/add_source
Добавить
/remove_source
Удалить

Парсинг и фильтрация постов

1. Сбор постов из **тг** каналов
2. Назначение весов для каналов: на основе количества подписчиков, просмотров, времени, наличию ссылок
3. Экономическая релевантность: Проверка полученных новостей на соответствие экономических тем
4. Посты с каналов проходят фильтры (объединение похожих постов, фильтрация рекламы, присваивание экономической темы)

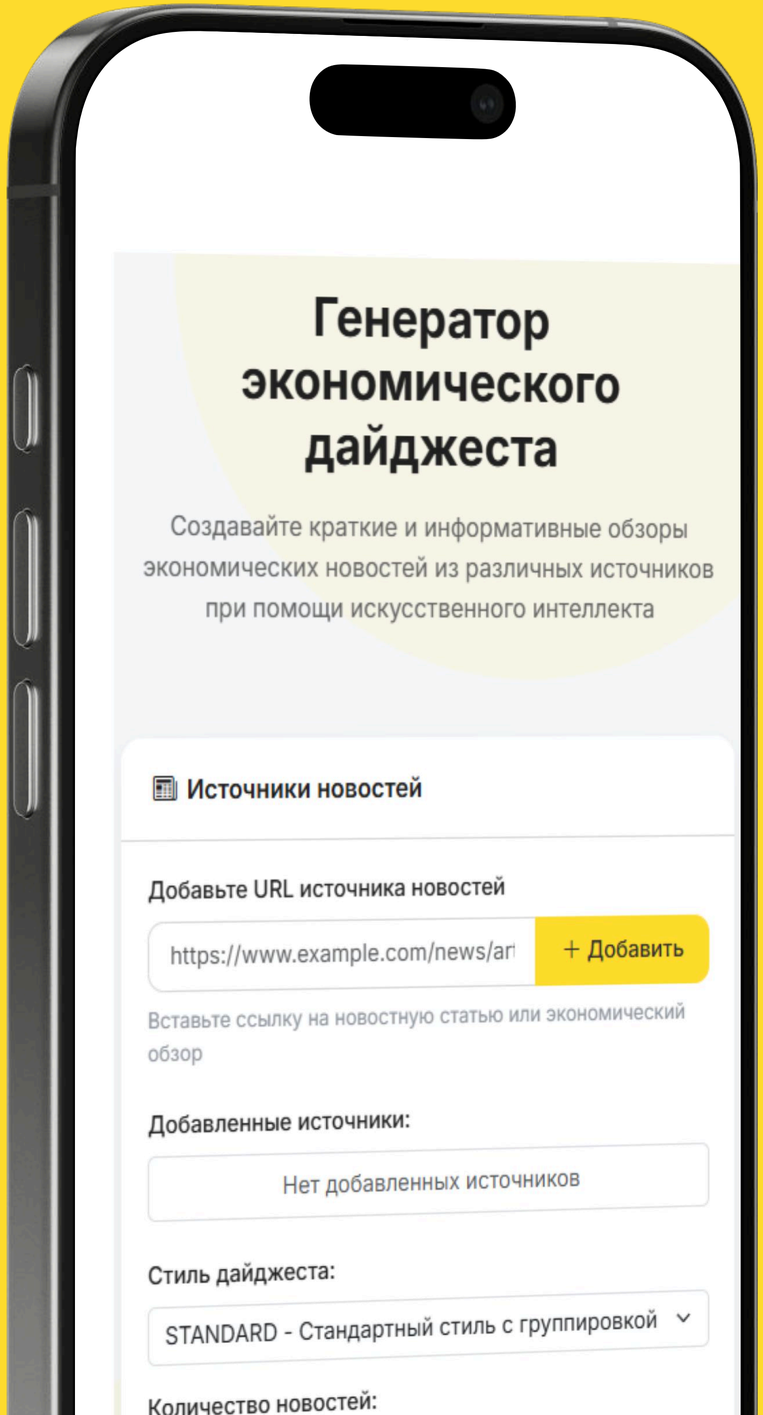


Стили саммаризации

- Стандартный
- Аналитический
- Компактный
- Карточки
- Социальный
- Медиа

Выбор embedding-моделей

	jina-embeddings-v3	**Mistral-Embed**	ru-en-RoSBERTa	sbert_large_nlu_ru
Mean (Task)	61.33	**67.63**	58.69	38.89
Mean (TaskType)	57.57	**64.25**	55.52	40.19
Classification	58.93	**74.30**	60.6	57.61
Clustering	45.17	**58.17**	47.34	46.34
Multilabel Classification	47.42	**50.01**	44.69	35.84



Web-App

Провели улучшение пользовательского опыта, добавив визуальный пользовательский интерфейс, упрощающий восприятие информации

Инструменты

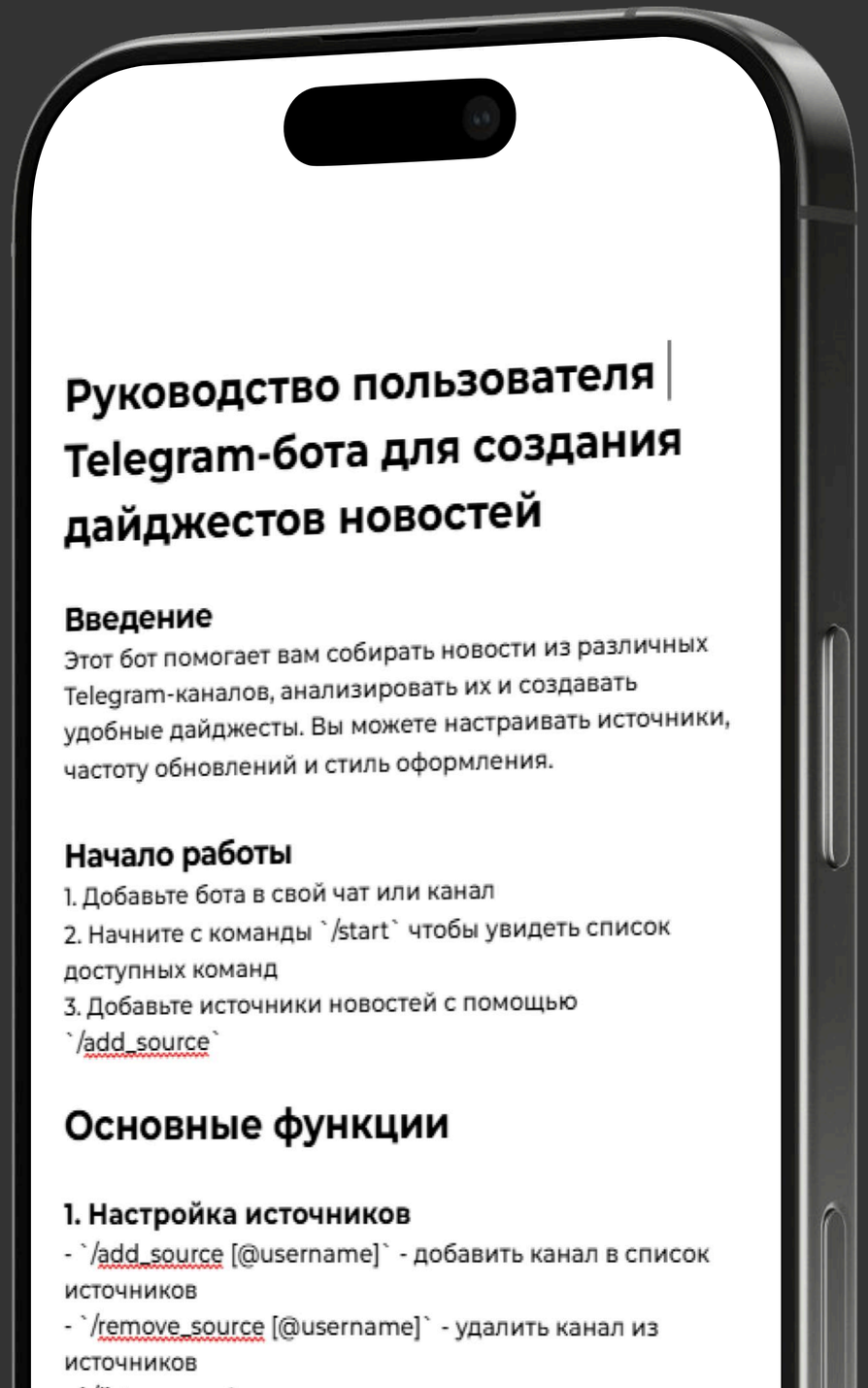
Frontend - HTML, CSS, JS

Backend - Flask, Python

Описание сценария работы с приложением

Подробное описание каждой функции сервиса в сопровождении визуальными изображениями

Вся информация для удобства собрана в Гугл-документе



Руководство пользователя Telegram-бота для создания дайджестов новостей

Введение

Этот бот помогает вам собирать новости из различных Telegram-каналов, анализировать их и создавать удобные дайджесты. Вы можете настраивать источники, частоту обновлений и стиль оформления.

Начало работы

1. Добавьте бота в свой чат или канал
2. Начните с команды `/start` чтобы увидеть список доступных команд
3. Добавьте источники новостей с помощью `/add_source`

Основные функции

1. Настройка источников

- `/add_source [@username]` - добавить канал в список источников
- `/remove_source [@username]` - удалить канал из источников

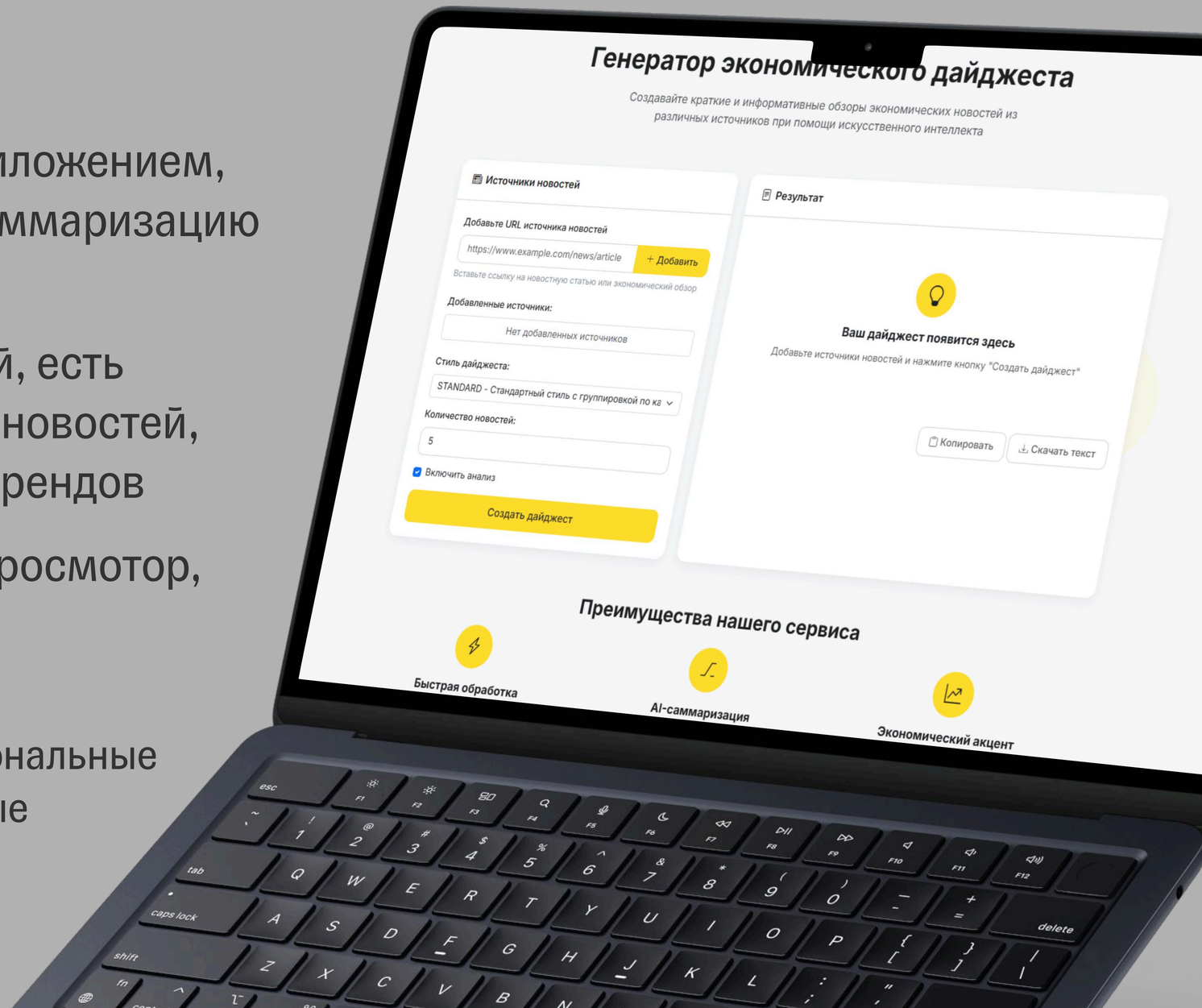
Результат

Мы разработали **тг бота с WebApp** приложением, который добавляется в чат и выдает саммаризацию постов в инвест-каналах.

Саммаризация имеет несколько стилей, есть вожможность настраивать количество новостей, частоту обновлений, вкл/выкл анализ трендов

Возможность настройки источников: просмотор, добавление и удаление.

Таким образом, мы выполнили все функциональные требования и реализовали все поставленные задачи



Спасибо за внимание!

Наш Git репозиторий

Немного о датасете

Russian News Dataset

Это датасет русскоязычных новостей, размеченных по категориям (например, "экономика", "политика", "спорт"). Источник - крупные СМИ(Их тгк)

Категория экономические новости содержит 50,816 rows.

Основные темы

Новости ЦБ РФ, курсы валют, фондовые рынки, налоги, криптовалюты, санкции, корпоративные отчеты

title float64	text string · lengths 1 4.11k	url string · lengths 40 53	sourceLink string · clas 1 value
null	Цены на нефть растут, американские военные отогнали наших от сирийский нефтяных…	https://t.me/nefte_baza/59?embed=1&mode=tme	https://t.me/
null	Истерия Лукашенко по поводу поиска альтернативных российским поставкам нефти,…	https://t.me/nefte_baza/69?embed=1&mode=tme	https://t.me/
null	Источник в силовых структурах: уголовное дело против экс-полпреда Президента Виктора Ишаев…	https://t.me/gazmyaso/1734?embed=1&mode=tme	https://t.me/
null	«ЛУКОЙЛ» подвергся жесткой критике знаменитого российского актёра и режиссёра…	https://t.me/nefte_baza/784?embed=1&mode=tme	https://t.me/
null	У Сечина забирают самолётик Сегодня США объявили о получении ордера на арест самолет…	https://t.me/papagaz/13278?embed=1&mode=tme	https://t.me/
null	Конечно выгодно... В 2022 году страны Старого Света переплатили за СПГ $1 трлн. Сущие…	https://t.me/papagaz/13281?embed=1&mode=tme	https://t.me/
null	"Роснефть", на дворе уже 2020-й, попробуйте в следующий раз чуть потоньше - а то сейчас…	https://t.me/nefte_baza/89?embed=1&mode=tme	https://t.me/