

Генерация кратких обзоров статей на основе полного текста

Дата: 23 апреля 2024 г.

Автор проекта:
Шерстнёва Алёна

Содержание презентации

- Введение
- Сбор данных с сайта TechCrunch и сохранение в базу данных
- Использование GPT модели для создания кратких пересказов
- Обработка и исследование данных
- Документация
- Визуализация данных
- Мониторинг и обновление данных
- Анализ результатов
- Заключение

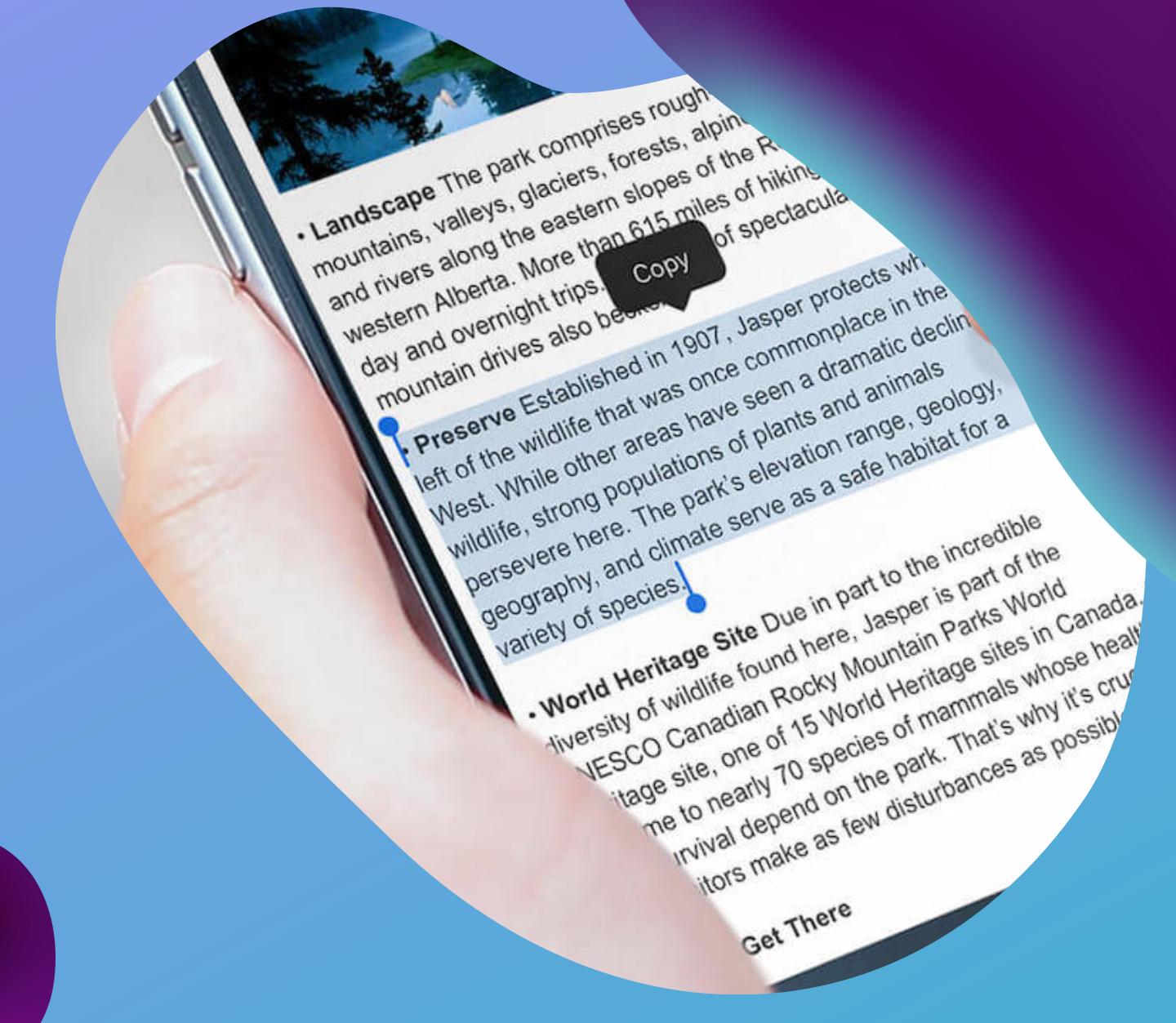
ОПИСАНИЕ ПРОЕКТА

Проект направлен на разработку системы автоматизации создания кратких пересказов статей с использованием GPT модели. Это решение поможет облегчить и оптимизировать работу с большим объемом текстовых данных, таких как статьи, новостные публикации, и другие текстовые материалы.

Введение

ЦЕЛЬ ПРОЕКТА

Основной целью проекта является создание эффективного инструмента, который позволит облегчить процесс получения кратких обзоров содержания статей. Это может быть полезно для аналитических отчетов, подготовки новостных сводок и других задач. Я постаралась улучшить процесс обработки информации, делая его более удобным, быстрым и эффективным.



Введение

ЗАДАЧИ ПРОЕКТА

- Разработка системы сбора данных: Создан скрипт для сбора статей с выбранного веб-сайта, используя библиотеки для работы с HTTP запросами и парсинга HTML кода.
- Интеграция с GPT моделью: Настроено взаимодействие с GPT моделью для генерации кратких пересказов текстов статей.
- Создание базы данных: Разработана база данных для хранения и управления собранными данными, включая исходные статьи и их краткие пересказы.
- Визуализация результатов: Создан бот в Telegram для постинга кратких статей в группу NewsSummariesHub.
- Автоматизация процесса: Оптимизирован и автоматизирован процесс сбора и обработки данных, а также постинга в группу в Telegram .

Введение

ПРЕИМУЩЕСТВА ПРОЕКТА

- Автоматизация процесса: Уменьшение времени и ресурсов, затрачиваемых на создание кратких обзоров статей.
- Высокая точность: GPT модель обеспечивает высокое качество кратких пересказов, соответствующих содержанию оригинальных текстов.
- Оптимизация: Упрощение работы с текстовыми данными

Введение

ИСПОЛЬЗОВАНИЕ БИБЛИОТЕК

Для сбора данных с сайта использованы две основные библиотеки:

- Requests: Это библиотека Python, которая позволяет отправлять HTTP-запросы к веб-сайтам. Мы используем эту библиотеку для отправки GET-запроса к странице сайта и получения HTML-кода страницы.
- BeautifulSoup: Это библиотека Python для парсинга HTML и XML документов. Мы используем BeautifulSoup для анализа HTML-кода страницы и извлечения необходимой информации, такой как заголовки статей и URL.

Сбор данных с сайта TechCrunch

Сохранение данных в базу данных

**ПОСЛЕ СБОРА ДАННЫХ О СТАТЬЯХ, СОХРАНЯЕМ ИХ В БАЗУ
ДАННЫХ SQLITE. ДЛЯ ЭТОГО ВЫПОЛНЯЕМ СЛЕДУЮЩИЕ ШАГИ:**

СОЗДАНИЕ ТАБЛИЦЫ

Создаем таблицу в базе данных для хранения информации о статьях. Эта таблица включает в себя следующие столбцы:

- id: Идентификатор статьи (PRIMARY KEY).
- title: Заголовок статьи.
- complete_text: Полный текст статьи.
- url: URL адрес статьи.

ВСТАВКА ДАННЫХ

После извлечения информации о статьях мы вставляем ее в созданную таблицу базы данных. Для каждой статьи мы сохраняем заголовок, полный текст и URL

Сбор данных с сайта TechCrunch и сохранение в базу данных

```

1 # Импортируем необходимые библиотеки
2 import requests
3 from bs4 import BeautifulSoup
4 import sqlite3

```

ФРАГМЕНТ КОДА:

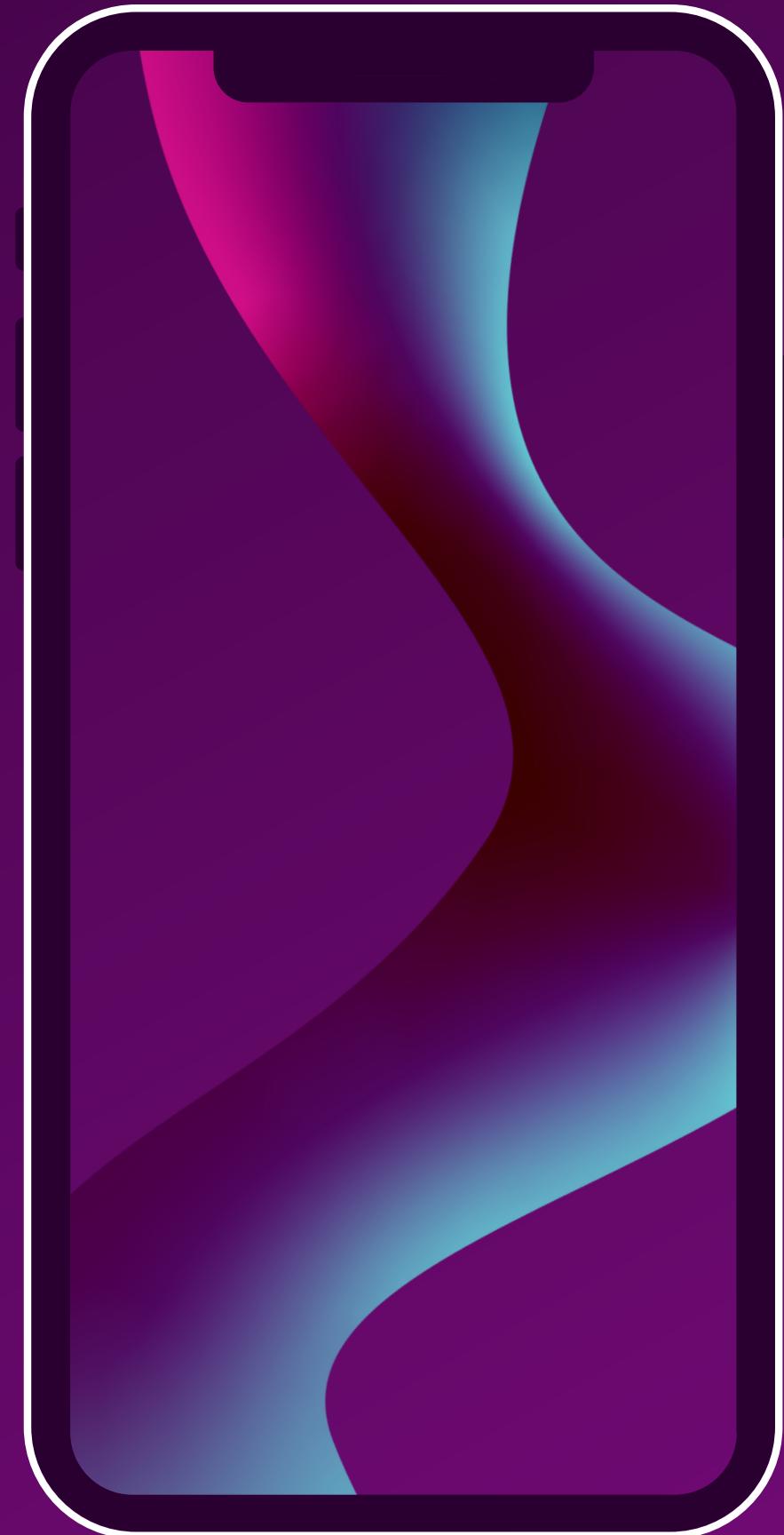
```

9 url = 'https://techcrunch.com/'
10
11 # Отправляем GET-запрос к странице
12 response = requests.get(url)
13
14 # Проверяем успешность запроса и парсим HTML-код страницы
15 if response.status_code == 200:
16     soup = BeautifulSoup(response.content, 'html.parser')
17     text_generation = soup.find_all('a', class_='post-block__title__link')
18
19 # Создаем подключение к БД
20 conn = sqlite3.connect('text_generation.db')
21 cursor = conn.cursor()
22
23 # Создаем таблицу для хранения статей
24 cursor.execute('''CREATE TABLE IF NOT EXISTS text_generation
25             (id INTEGER PRIMARY KEY AUTOINCREMENT,
26              title TEXT,
27              complete_text TEXT,
28              url TEXT)''')
29
30
31 # Вставка данных статьи в таблицу
32 cursor.execute("INSERT INTO text_generation (title, complete_text, url) VALUES (?, ?, ?)",
33                 (article_title, article_text, article_url))
34 conn.commit()
35 print('Title:', article_title)
36 print('Text:', article_text)
37 print('URL:', article_url)

```

Использование GPT-модели для создания кратких пересказов

GPT (Generative Pre-trained Transformer) - это серия языковых моделей, разработанных компанией OpenAI. Они основаны на архитектуре трансформеров и обучаются на огромных объемах текстовых данных с использованием метода обучения без учителя.



Особенности GPT моделей

ОСОБЕННОСТЬ 1

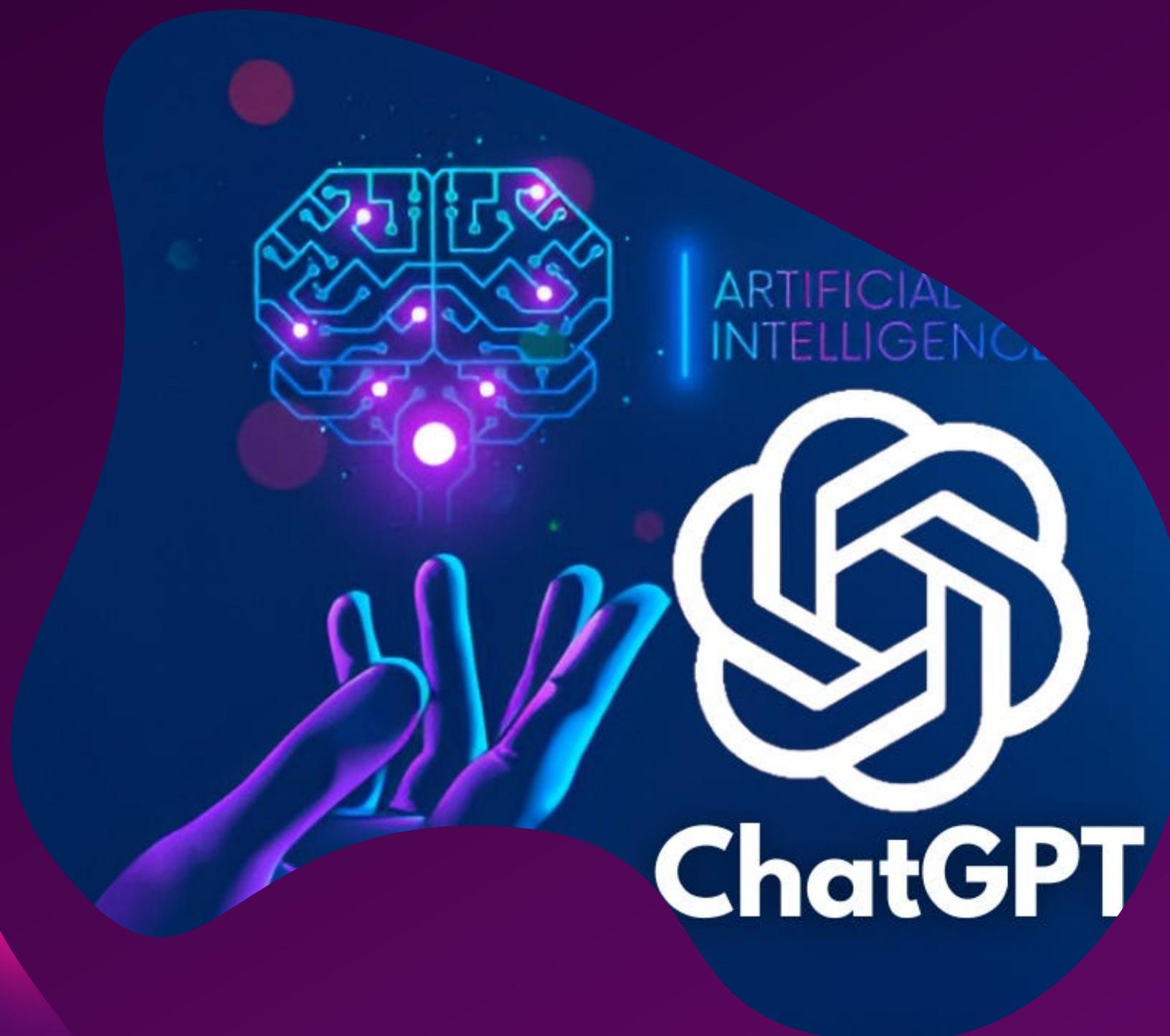
Эти модели обладают уникальной способностью генерации текста, что означает, что они могут создавать новые тексты на основе входных данных. GPT модели могут генерировать продолжения текста, отвечать на вопросы, создавать статьи, рецензии, поэзию и многое другое.

ОСОБЕННОСТЬ 2

В основе работы GPT моделей лежит механизм самообучения: они анализируют большие объемы текстов и строят внутреннее представление о структуре языка и связях между словами. Затем они используют это знание для генерации текстов, которые соответствуют заданному контексту.

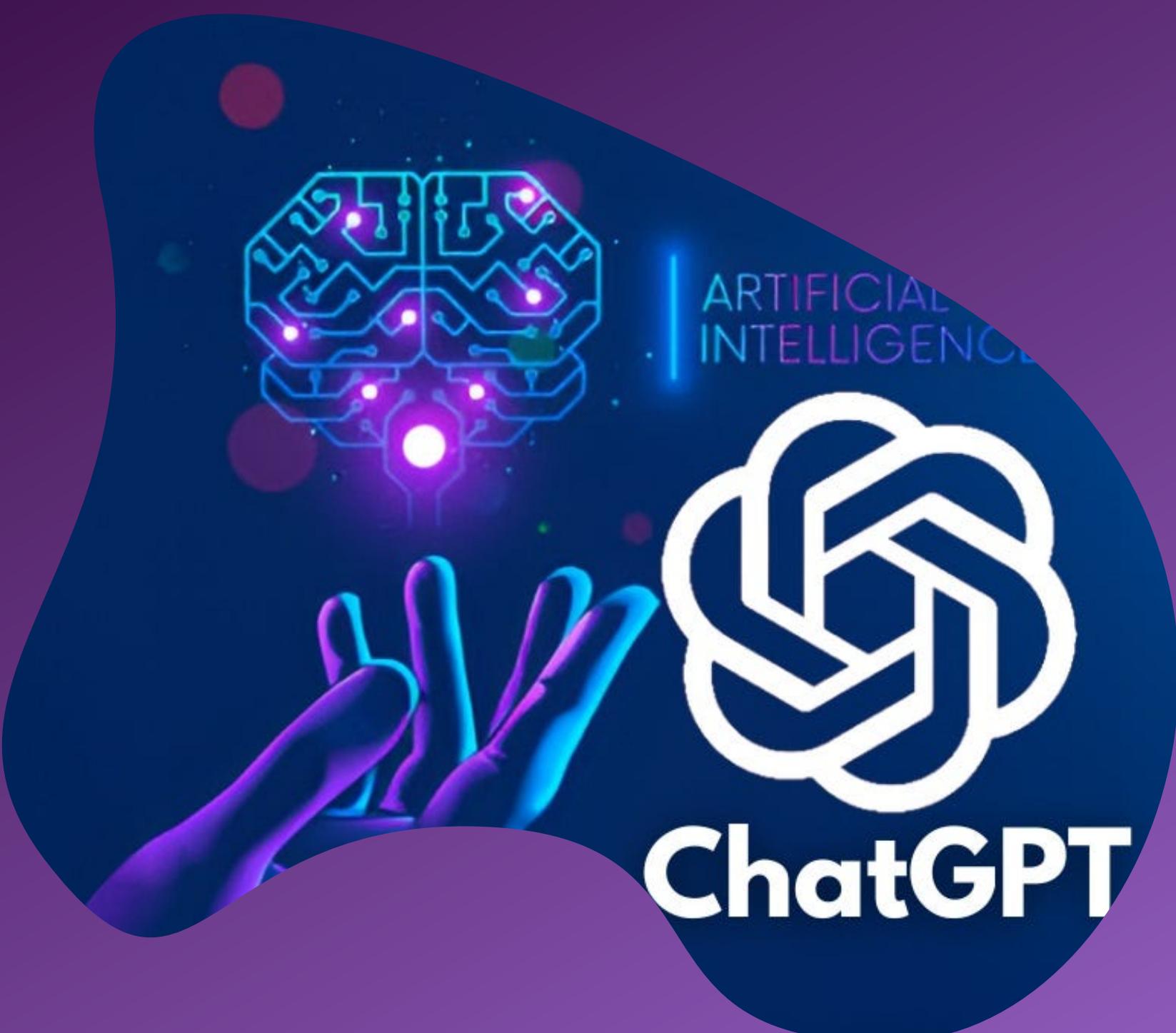
ОСОБЕННОСТЬ 3

GPT модели часто используются для автоматической генерации текстов, создания контента, обработки естественного языка, анализа данных и других задач, где требуется работа с текстовой информацией.



ОБЗОР API CHAT GPT:

- Для создания кратких пересказов используется API chat GPT, которое предоставляет доступ к GPT модели. Это API позволяет обращаться к мощным языковым моделям и генерировать тексты на основе введенного контекста.



ОБРАЩЕНИЕ К МОДЕЛИ:

- Отправка текстов статей в API: Отправка полных текстов статей в API chat GPT для обработки. Каждый текст статьи служит входным контекстом для модели.
- Генерация кратких пересказов: Модель анализирует введенный текст и генерирует краткий пересказ на основе содержания статьи. Этот пересказ модель переводит на русский язык в качестве краткого обзора статьи на русском языке.

СОХРАНЕНИЕ РЕЗУЛЬТАТОВ:

- Сохранение пересказов в базе данных: Полученные краткие пересказы сохраняются в базе данных. Для каждой статьи сохраняется связанный с ней краткий пересказ.
- Связывание с исходными статьями: Каждый краткий пересказ связывается с соответствующей исходной статьей. Это позволяет легко отслеживать и сопоставлять краткие обзоры с оригинальными текстами.

ИСПОЛЬЗОВАНИЕ GPT МОДЕЛИ ДЛЯ СОЗДАНИЯ КРАТКИХ ПЕРЕСКАЗОВ

ФРАГМЕНТ КОДА:

```
7  from openai import OpenAI
8
9
10 # Создаем таблицу для хранения кратких пересказов статей
11 cursor.execute('''CREATE TABLE IF NOT EXISTS short_table
12     (id INTEGER PRIMARY KEY AUTOINCREMENT,
13      short_text TEXT,
14      article_id INTEGER,
15      FOREIGN KEY(article_id) REFERENCES text_generation(id))''')
16
17 # Установка API ключа OpenAI
18 api_key = 'sk-'
19 client = OpenAI(api_key=api_key)
20
21 # Ограничение на количество обрабатываемых статей
22 max_articles = 20
23 articles_processed = 0
24
25
26 # Вызываем API для создания краткого пересказа
27 completion = client.chat.completions.create(
28     model="gpt-3.5-turbo",
29     messages=[
30         {"role": "system", "content": "Ты переводчик и пересказчик."},
31         {"role": "user", "content": f"Сделай краткий пересказ текста и
32         | переведи краткий пересказ на русский язык:\n\n{article_text}"}
33     ]
34 )
35 short_text = completion.choices[0].message
36 short_text = str(short_text)
```

short_text.replace('ChatCompletionMessage(content='', '')'): Эта строка заменяет подстроку 'ChatCompletionMessage(content=' в переменной short_text на пустую строку. Она используется для удаления лишней информации, которая может быть возвращена в начале пересказа.

```
short_text = short_text.replace('ChatCompletionMessage(content='', '')')
short_text = re.sub(r',? role=\\"assistant\\", function_call=None, tool_calls=None', '', short_text)
short_text = short_text.replace('Краткий пересказ:', '')
short_text = short_text.replace('Пересказ:', '')
short_text = short_text.replace('\\n', '\n')
```

re.sub(r',? role=\\"assistant\\", function_call=None, tool_calls=None', '', short_text):

Эта строка использует регулярное выражение для замены подстроки ',? role=\\"assistant\\", function_call=None, tool_calls=None' в переменной short_text на пустую строку. Это снова удаляет ненужную информацию из текста.

short_text.replace('Краткий пересказ:', ''): Эта строка заменяет подстроку 'Краткий пересказ:' в переменной short_text на пустую строку. Это часто используется для удаления заголовков или меток, которые могут быть добавлены к тексту.

short_text.replace('Пересказ:', ''): Эта строка заменяет подстроку 'Пересказ:' в переменной short_text на пустую строку. Это также удаляет метки или заголовки из текста.

short_text.replace('\\n', '\n'): Эта строка заменяет символы перевода строки '\\n', которые могут быть представлены как '\n', в переменной short_text на реальные символы новой строки \n. Это используется для правильного форматирования текста.

Обработка данных

PYTHON

Обработка данных

SQLite

...

```
DELETE FROM text_generation
WHERE id NOT IN (
SELECT MIN(id)
FROM text_generation
GROUP BY title, complete_text
);
```

УДАЛЕНИЕ ДУБЛИКАТОВ

```
CREATE TABLE IF NOT EXISTS final_table(
id INTEGER PRIMARY KEY AUTOINCREMENT,
full_text TEXT,
short_text TEXT,
url TEXT,
FOREIGN KEY (id) REFERENCES text_generation(id)
);
```

СОЗДАНИЕ СВОДНОЙ
ТАБЛИЦЫ final_table

```
INSERT INTO final_table(full_text, short_text, url)
SELECT text_generation.complete_text, short_table.short_text , t
FROM text_generation
JOIN short_table ON text_generation.id = short_table.article_id;
```

ВСТАВКА ДАННЫХ В
СВОДНУЮ ТАБЛИЦУ

```
DELETE FROM final_table WHERE full_text IS NULL OR full_text = '';
```

УДАЛЕНИЕ ПУСТЫХ
ЗНАЧЕНИЙ

Сравнение количества символов
в полном и кратком текстах

Исследование данных

```
SELECT id,
       LENGTH(full_text) AS full_text_symbols,
       LENGTH(short_text) AS short_text_symbols
  FROM final_table;
```

123 id	123 full_text_symbols	123 short_text_symbols
546 ↗	3,380	766
547 ↗	3,653	1,009
548 ↗	5,508	562
549 ↗	7,882	956
550 ↗	4,016	1,070
551 ↗	1,708	627
552 ↗	5,986	472
553 ↗	2,051	472
554 ↗	3,653	1,322
555 ↗	3,389	506
556 ↗	3,887	620

Исследование данных

Средняя длина текста статей

```
SELECT ROUND(AVG(LENGTH(full_text)), 2)
AS avg_full,
ROUND(AVG(LENGTH(short_text)), 2)
AS avg_short
FROM final_table ft;
```

123 avg_full	123 avg_short
5,199.74	1,244.26

Процент общей длины коротких пересказов к
общей длине полных текстов

```
SELECT SUM(LENGTH(short_text))
AS total_short_text_length,
SUM(LENGTH(full_text))
AS total_full_text_length,
ROUND((SUM(LENGTH(short_text)) * 100.0 / SUM(LENGTH(full_text))),1)
AS percent
FROM final_table;
```

123 total_short_text_length	123 total_full_text_length	123 percent
476,254	1,904,117	25

ДОКУМЕНТАЦИЯ

Словарь данных

text_generation - таблица полных текстов статей

Table name	Column name	Data type	Business-description	Example	Access level
text_generation	id	INTEGER	Уникальный идентификатор статьи	3	Публичный
text_generation	title	TEXT	Заголовок статьи	Quibi redux? Short drama apps saw record revenue in Q1 2024	Публичный
text_generation	complete_text	TEXT	Полный текст статьи	Was Quibi just ahead of its time? Quibi founder Jeffrey Katzenberg ultimately blamed the COVID-19 pandemic for the failure of his short-form video app, but maybe it was just too soon etc.	Публичный
text_generation	url	TEXT	Uniform Resource Locator Адрес статьи	https://techcrunch.com/2024/04/11/quibi-redux-short-drama-apps-saw-record-revenue-in-q1-2024/	Публичный

short_table - таблица кратких пересказов статей

Table name	Column name	Data type	Business-description	Example	Access level
short_table	id	INTEGER	Уникальный идентификатор статьи	3	Ограниченный
short_table	short_text	TEXT	Краткий пересказ статьи на русском языке	В начале 2024 года такие приложения привнесли рекордные \$146 миллионов в мировых расходах потребителей, что является огромным ростом по сравнению с предыдущим годом.	Ограниченный
short_table	article_id	INTEGER	Служит для связи с полным текстом статьи <u>complete_text</u> из таблицы <u>text_generation</u>	256	Ограниченный

final_table - сводная таблица

Table name	Column name	Data type	Business-description	Example	Access level
final_table	id	INTEGER	Уникальный идентификатор статьи	3	Публичный
final_table	full_text	TEXT	Краткий пересказ статьи на русском языке	Was Quibi just ahead of its time? Quibi founder Jeffrey Katzenberg ultimately blamed the COVID-19 pandemic for the failure of his short-form video app, but maybe it was just too soon etc	Публичный
final_table	short_text	TEXT	Служит для связи с полным текстом статьи <u>complete_text</u> из таблицы <u>text_generation</u>	В начале 2024 года такие приложения привнесли рекордные \$146 миллионов в мировых расходах потребителей, что является огромным ростом по сравнению с предыдущим годом.	Ограниченный
final_table	url	TEXT	Uniform Resource Locator Адрес статьи	https://techcrunch.com/2024/04/11/quibi-redux-short-drama-apps-saw-record-revenue-in-q1-2024/	Публичный

ДОКУМЕНТАЦИЯ

Source				Target				
Table name	Column name	Data type	Logic of transformation	Table name	Column name	Data type	Data example	Comment
text_generation	id	INTEGER	-	final_table	-	INTEGER	7	
text_generation	title	TEXT	-	final_table	-	-	-	-
text_generation	complete_text	TEXT	<pre>INSERT INTO final_table(full_text, short_text, url) SELECT text_generation.complete_text, short_table.short_text , text_generation.url FROM text_generation JOIN short_table ON text_generation.id = short_table.article_id ;</pre>	final_table	full_text	TEXT	Was Quibi just ahead of its time? Quibi founder Jeffrey Katzenberg ultimately blamed the COVID-19 pandemic for the failure of his short-form video app, but maybe it was just too soon etc	Вставка данных из таблицы text_generation в таблицу final_table
text_generation	url	TEXT	<pre>DELETE FROM final_table WHERE id NOT IN (SELECT MIN(id) FROM final_table GROUP BY url);</pre>	final_table	url	TEXT	https://techcrunch.com/2024/04/11/quibi-reduces-short-drama-apps-saw-record-revenue-in-q1-2024/	Удаление дубликатов

Source				Target				
Table name	Column name	Data type	Logic of transformation	Table name	Column name	Data type	Data example	Comment
short_table	id	INTEGER	-	-	-	-	-	-
short_table	short_text	TEXT	<pre>INSERT INTO final_table(full_text, short_text, url) SELECT text_generation.complete_text, short_table.short_text , text_generation.url FROM text_generation JOIN short_table ON text_generation.id = short_table.article_id ;</pre>	final_table	short_text	TEXT	В начале 2024 года такие приложения принесли рекордные \$146 миллионов в мировых расходах потребителей, что является огромным ростом по сравнению с предыдущим годом.	Вставка данных из таблицы short_table в таблицу final_table
short_table	article_id	TEXT	-	-	-	-	-	-

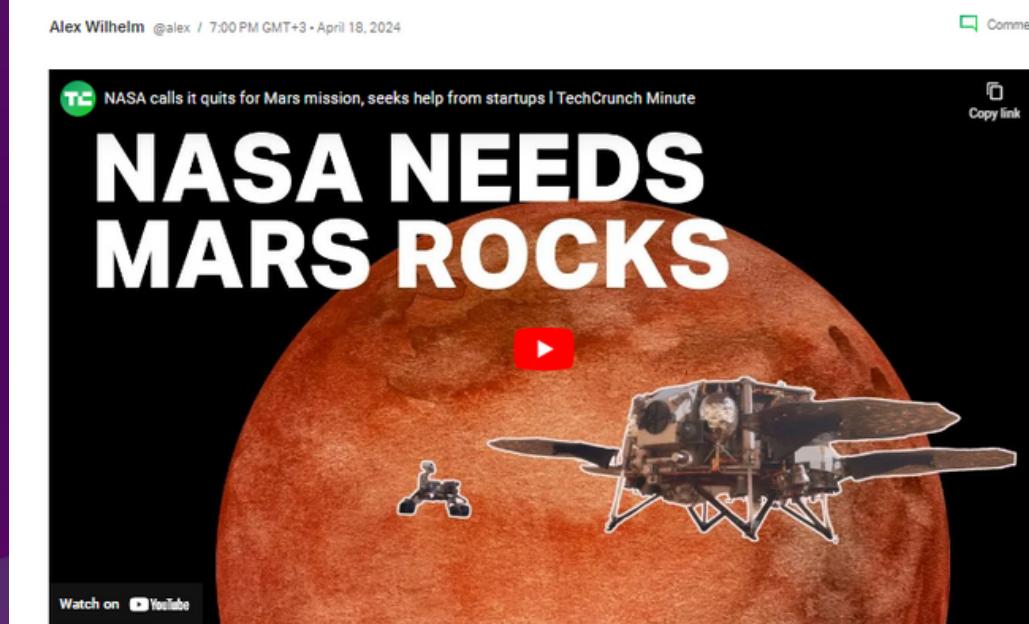
Source To Target Mapping

Визуализация данных

Выполнена в Tableau Public

Входные данные:

Статья с сайта TechCrunch на английском языке
<https://techcrunch.com/2024/04/18/techcrunch-minute-nasa-needs-your-help-to-bring-rocks-back-from-mars/>



NASA's decision to scrap its \$11 billion, 15-year mission to Mars to bring back samples could create a startup feeding frenzy. TechCrunch reports. Describing its plans as too slow, and too expensive, NASA is going back to the drawing board, with an eye on getting the space industry to help. Sure, you might worry that NASA can't manage its own mission on a timeline and budget that it deems acceptable, but the chance for a deluge of dollars to engulf the startups working on making space more accessible could prove a massive boon.

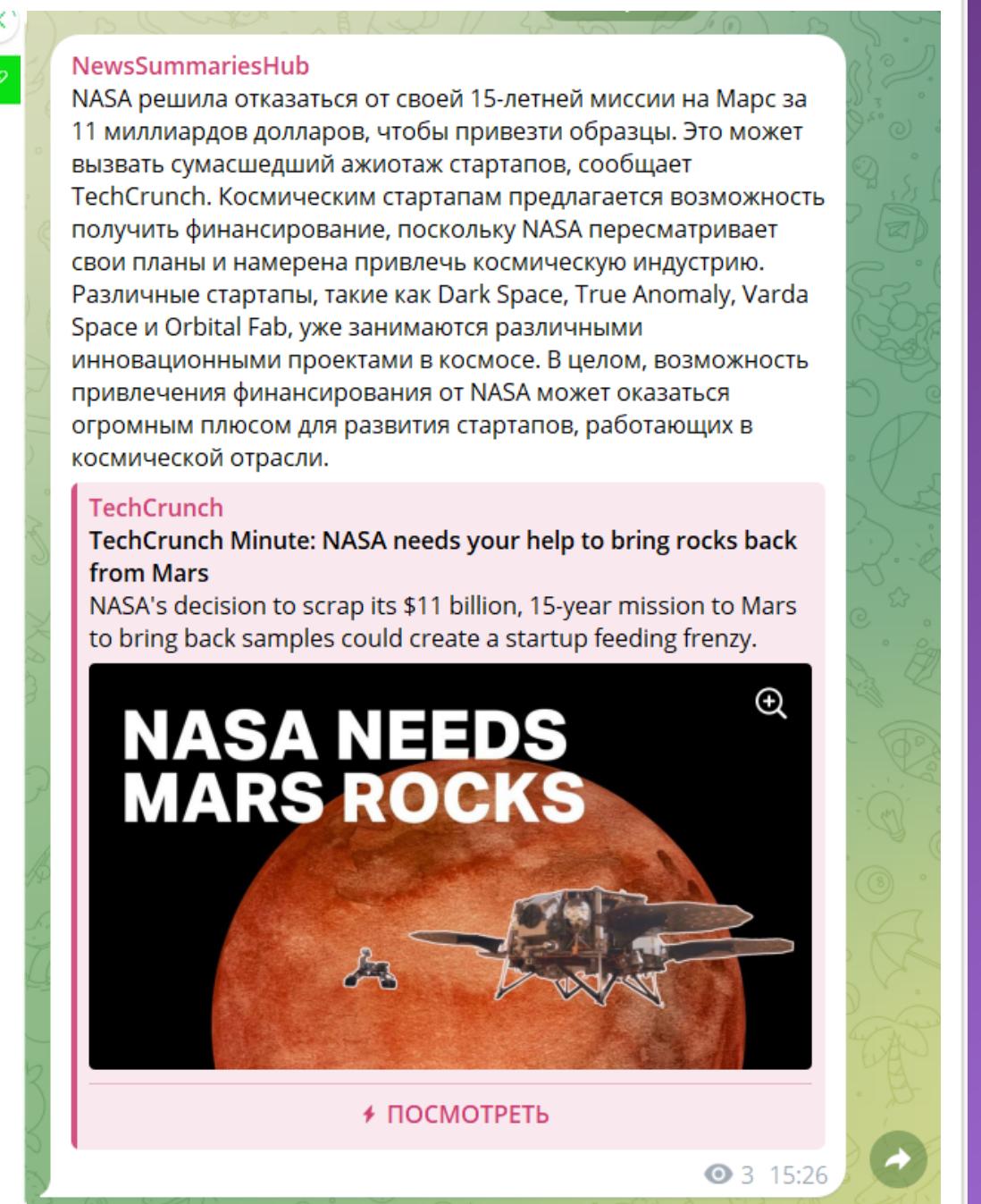
Startups are not all social media apps, enterprise software and NFT-based online games. There are a good number focused on the bits-and-atoms side of the technology fence, even if the idea of building advanced hardware without a software element is all but unthinkable. Ergo, hardware startups are really working both sides of the digital divide at the same time.

But space startups are not worried about it. Looking at recent TechCrunch space headlines, we can see that Dark Space is working on a way to clear space debris; True Anomaly's working on landing on the moon; Varda Space's work to manufacture drugs in space and bring them back to Earth seems to work, so it raised \$90 million more; Orbital Fab wants to refuel satellites; the list goes on and on.

So, the NASA money might have a bunch of startup-sized buckets to drip into, and I am here for it. Yes, I am a gigantic science-fiction dweeb, but I am still nothing short of dizzy with hype for our future as a species in space. To that end, if any startup that works with NASA on the Mars rock mission needs a human to send up there to check on the dials and such, I'm your guy. Hit play, kids, because I'm

Итоговый результат:

Краткий пересказ статьи на русском языке, опубликованный ботом в Telegram-канале NewsSummariesHub





Визуализация данных

Мониторинг и обновление данных

Код осуществляет отправку последних 10 записей из БД в группу Telegram. Он подключается к базе данных SQLite, выбирает последние записи из таблицы final_table, формирует сообщение, объединяя короткий текст и URL каждой записи, и отправляет его в указанный чат с помощью Telegram бота.

```
def send_message(token, chat_id, text):
    url = f"https://api.telegram.org/bot{token}/sendMessage"
    params = {"chat_id": chat_id, "text": text}
    response = requests.get(url, params=params)
    return response.json()
```

```
# Данные о боте и группе
token = ""
chat_id = "@NewsSummariesHub"
```

```
# Получение последних 20 записей из таблицы final_table
cursor.execute("SELECT short_text, url FROM final_table ORDER BY id DESC LIMIT 20")
rows = cursor.fetchall()

# Отправка каждой записи в группу Telegram
for row in rows:
    short_text = row[0]
    url = row[1]
    message_text = f"{short_text}\n{url}"
    send_message(token, chat_id, message_text)
```

Полный текст статьи с сайта <https://techcrunch.com> - 1148 символов

NASA's decision to scrap its \$11 billion, 15-year mission to Mars to bring back samples could create a startup feeding frenzy, TechCrunch reports. Describing its plans as too slow, and too expensive, NASA is going back to the drawing board, with an eye on getting the space industry to help. Sure, you might worry that NASA can't manage its own mission on a timeline and budget that it deems acceptable, but the chance for a deluge of dollars to engulf the startups working on making space more accessible could prove a massive boon.

Startups are not all social media apps, enterprise software and NFT-based online games. There are a good number focused on the bits-and-atoms side of the technology fence, even if the idea of building advanced hardware without a software element is all but unthinkable. Ergo, hardware startups are really working both sides of the digital divide at the same time.

But space startups are not worried about it. Looking at recent TechCrunch space headlines, we can see that Dark Space is working on a way to clear space debris; True Anomaly's working on landing on the moon; Varda Space's work to manufacture drugs in space and bring them back to Earth seems to work, so it raised \$90 million more; Orbital Fab wants to refuel satellites; the list goes on and on.

So, the NASA money might have a bunch of startup-sized buckets to drip into, and I am here for it. Yes, I am a gigantic science-fiction dweeb, but I am still nothing short of dizzy with hype for our future as a species in space. To that end, if any startup that works with NASA on the Mars rock mission needs a human to send up there to check on the dials and such, I'm your guy. Hit play, let's have some fun!

Перевод полного текста статьи на русский язык - 1051 символ

Решение НАСА отказаться от своей 11-миллиардной миссии на 15 лет на Марс, чтобы привезти образцы, может вызвать настоящий ажиотаж в стартап-сообществе, сообщает TechCrunch. Описывая свои планы как слишком медленные и слишком дорогие, НАСА возвращается к чертежной доске с намерением привлечь к помощи индустрию космоса. Конечно, вы можете опасаться, что НАСА не в состоянии управлять своей собственной миссией в сроки и бюджет, которые она считает приемлемыми, но возможность того, что стартапы, работающие над сделкой космоса более доступным, могут получить огромную поддержку, может оказаться огромным плюсом.

Стартапы - это не только приложения для социальных сетей, корпоративное программное обеспечение и онлайн-игры на основе NFT. Существует большое количество компаний, сосредоточенных на стороне технологического забора из битов и атомов, даже если идея создания сложного оборудования без программной составляющей практически невозможна. Следовательно, стартапы в сфере аппаратного обеспечения действительно работают с обеими сторонами цифрового раздела одновременно.

Но стартапы, работающие в сфере космоса, об этом не беспокоятся. Просматривая недавние заголовки TechCrunch о космосе, мы видим, что Dark Space работает над способом очистки космических мусоров; True Anomaly работает над приземлением на Луну; работа Varda Space по производству лекарств в космосе и их возвращению на Землю, кажется, успешна, поэтому она привлекла еще 90 миллионов долларов; Orbital Fab хочет заправлять спутники; список продолжается.

Таким образом, деньги НАСА могут иметь ряд стартап-размерных ведер, чтобы капать в них, и я рад этому. Да, я фанат научной фантастики, но я все же не могу не ощущать восторга от нашего будущего как видов в космосе. В этом духе, если какой-либо стартап, работающий с НАСА над миссией на Марс, нуждается в человеке, чтобы отправить его туда, чтобы проверить дисплеи и прочее, я ваш человек. Нажмите play, давайте повеселимся!

Анализ результатов

Краткий пересказ статьи на русском языке - 484 символа

NASA решила отказаться от своей 15-летней миссии на Марс за 11 миллиардов долларов, чтобы привезти образцы. Это может вызвать сумасшедший ажиотаж стартапов, сообщает TechCrunch. Космическим стартапам предлагается возможность получить финансирование, поскольку NASA пересматривает свои планы и намерена привлечь космическую индустрию. Различные стартапы, такие как Dark Space, True Anomaly, Varda Space и Orbital Fab, уже занимаются различными инновационными проектами в космосе. В целом, возможность привлечения финансирования от NASA может оказаться огромным плюсом для развития стартапов, работающих в космической отрасли.

Анализ результатов

Выводы



Пересказ предоставляет основные факты и суть оригинального текста.



Пересказ короток и содержит ключевую информацию о событии, поддерживая интерес читателя к продолжению оригинальной статьи.



В пересказе уделено внимание основным аспектам: решение NASA отказаться от миссии на Марс, возможности для стартапов и конкретные примеры таких стартапов.

ЗАКЛЮЧЕНИЕ



- Разработана система автоматизации создания кратких пересказов статей с использованием GPT модели, это облегчает процесс создания кратких обзоров статей, что полезно для аналитических отчетов, подготовки новостных сводок и других задач
- Используются актуальные технологии для сбора, анализа и обработки данных, что позволяет существенно сократить время и ресурсы, затрачиваемые на обработку текстовой информации
- Результаты проекта имеют потенциал стать полезным инструментом как для пользователей, работающих с текстовой информацией

Презентация окончена

