

**Приложение для определения
названия фильма по
видеофрагменту в формате
YouTube Shorts**

Введение

С ростом популярности коротких видеофрагментов в социальных сетях, пользователи сталкиваются с проблемой неспособности быстро и легко определить название фильма, представленного в коротком видеоролике. Это создает потребность в эффективных инструментах для распознавания контента внутри видео и обеспечения более удобного пользовательского опыта.

Платформы типа TikTok, Instagram Reels и YouTube Shorts стали магнитами для аудитории по всему миру, предлагая быстрое употребление развлекательного и образовательного контента. В сфере социальных медиа, формат коротких видео стал одним из самых растущих по популярности трендов последних лет [1]. На рис. 1 показано сравнение популярности TikTok с другими соцсетями.

Top 10 most used social networks in the U.S.

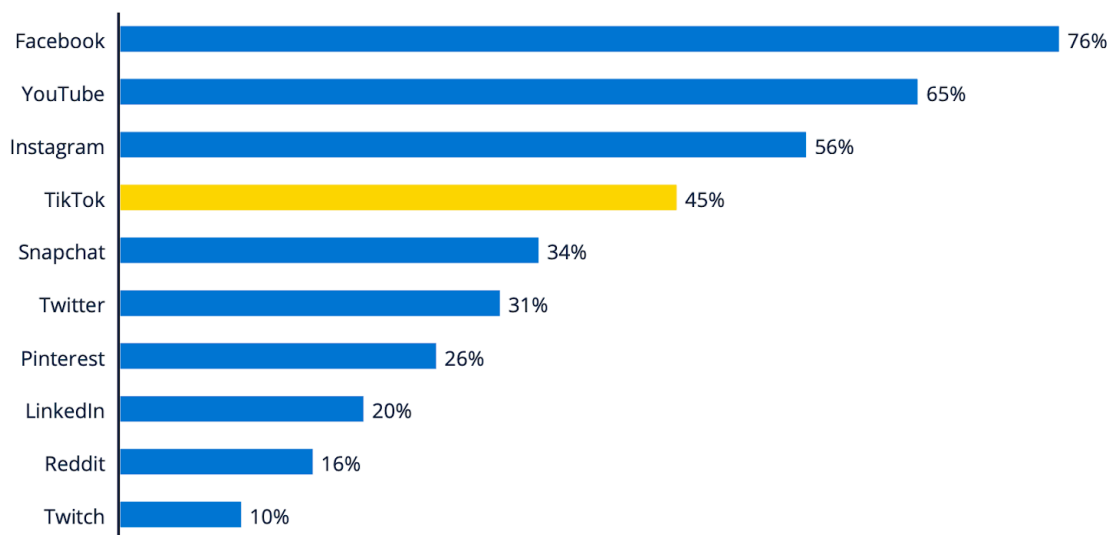


РИС. 1. САМЫЕ ПОПУЛЯРНЫЕ СОЦИАЛЬНЫЕ СЕТИ В США

На рис. 2 показан рост пользователей платформы TikTok за последние два года.

Timeline of social media users using TikTok

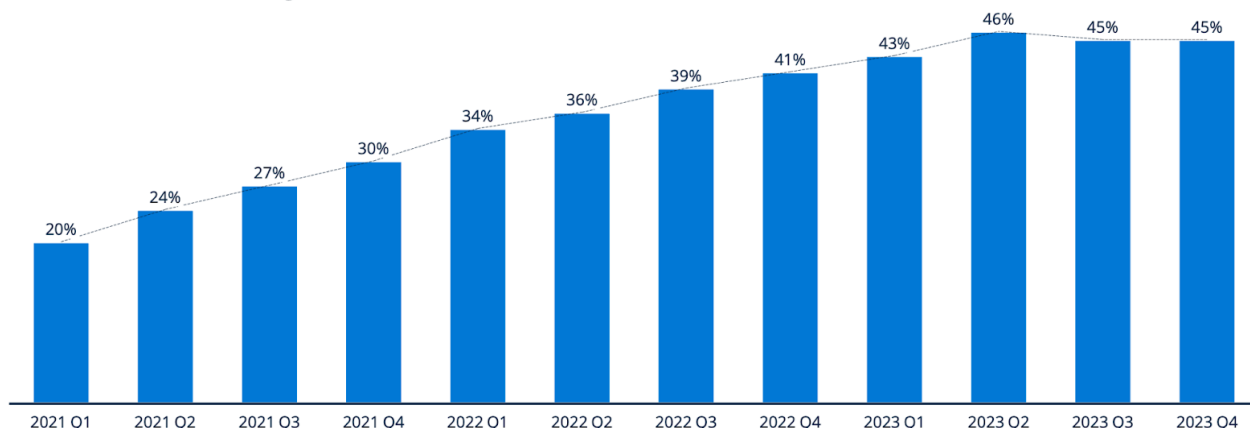


РИС. 2. ЧИСЛО ПОЛЬЗОВАТЕЛЕЙ ПЛАТФОРМЫ ТИКТОК

С ростом популярности платформ для коротких видеороликов растёт и популярность коротких отрывков из фильмов. Кинопоиск в своей статье [2] делится опасением, что популярность TikTok даже несёт угрозу стриминговым сервисам, ведь всё больше пользователей смотрят отрывки, а иногда и целые эпизоды сериалов и фильмов в формате коротких видеороликов.

По статистике TikTok Creative Center [3] на их платформе опубликовано более 500 миллиардов коротких видеороликов с хештегом film. Из них 93 миллиарда - за последние три года. Тренд публикаций показан на рис. 3.

Interest over time

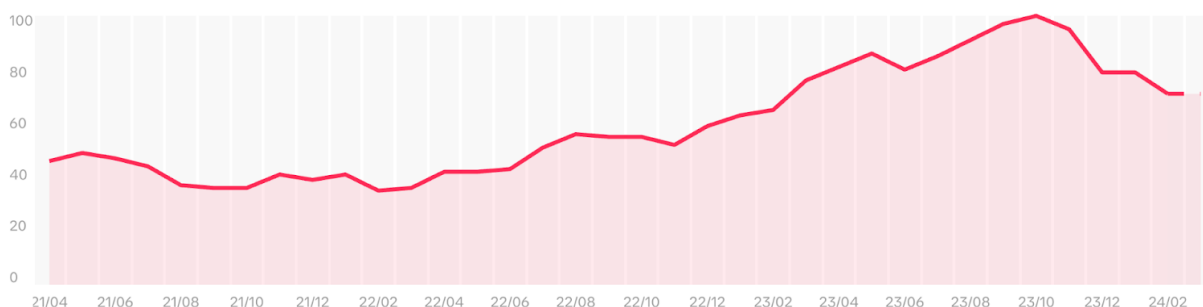


РИС. 3. ТРЕНД ПУБЛИКАЦИИ РОЛИКОВ С ХЕШТЕГОМ #FILM В ТИКТОК

Сфера стриминга и киноиндустрия в целом не отстаёт от тренда коротких видеороликов. Например, один из самых популярных стриминговых сервисов Netflix запустил функцию Fast Laughs - аналог TikTok с отрывками из фильмов и сериалов [4].

Проанализировав данные популярного интернет ресурса Quora [5], можно заметить, что с каждым годом эта тема всё актуальнее (рис. 4):

"What is the movie where..." Quora questions count

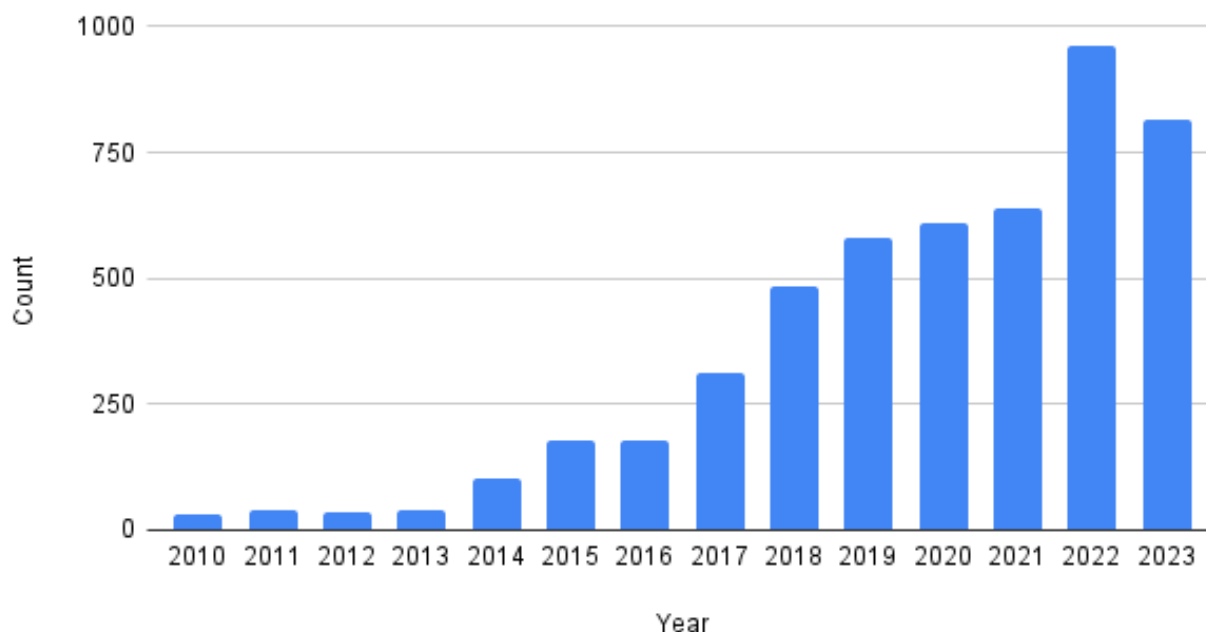


РИС. 4. РОСТ КОЛИЧЕСТВА ВОПРОСОВ, В КОТОРЫХ ПОЛЬЗОВАТЕЛИ ПЫТАЮТСЯ ОПРЕДЕЛИТЬ НАЗВАНИЕ ФИЛЬМА

Исходя из вышеописанного, мы делаем вывод, что индустрия коротких видеороликов и тренд публикации отрывков из фильмов в интернете пользуются значительной популярностью у пользователей интернета. Значит, наше приложение имеет большой потенциал к успеху.

Цель работы – разработка специализированного приложения, способного автоматически определять название фильма по короткому видеофрагменту, такому как YouTube Shorts.

Задачи:

- Сбор и анализ данных по текущей ситуации на рынке
- Изучение существующих методов распознавания кадров видеоконтента
- Разработка алгоритма распознавания фильма по видеофрагменту
- Интеграция с API и сервисами YouTube
- Разработка пользовательского интерфейса приложения
- Тестирование и оптимизация

Объект – создание алгоритма, распознающего фильмы по видеофрагменту в формате YouTube Shorts.

Предмет - существующие методы для обработки и распознавания фото, видео и аудио контента.

Существующие методы распознавания

Пользователи в интернете часто задаются вопросом: «Как найти название фильма по видео отрывку?»

Такой же вопрос задал Vivek Ranjan на популярном вопросном-ответном сайте Quora [6]. В основном пользователи посоветовали ему использовать реверсивный поиск изображений. Для этого необходимо сделать скриншот из видео и воспользоваться одним из сервисов для поиска по изображениям, например Google Images или Yandex Images.

Рассмотрим подробнее технологию реверсивного поиска изображений.

Реверсивный поиск изображений

Один из примеров публикаций, рассматривающих реверсивный поиск изображений, — это статья "Поиск изображений" автора qwertyforce, опубликованная на портале Хабр [7]. Qwertyforce обсуждает различные аспекты технологии, включая алгоритмы извлечения признаков и методы индексации данных, что может быть полезным для глубокого понимания темы.

Реверсивный поиск изображений — это технология, позволяющая находить информацию по изображению, загружаемому пользователем. Вместо того чтобы искать при помощи текстовых запросов, пользователи могут использовать уже существующее изображение для поиска схожих результатов в интернете.

Основные принципы работы реверсивного поиска изображений включают:

1. **Извлечение признаков:** Программное обеспечение анализирует загружаемое изображение и извлекает из него ключевые определяющие элементы (узоры, текстуру, формы, цвета и другие).
2. **Преобразование в уникальные отпечатки (хеш-коды):** Эти признаки преобразуются в числовой код, или "хеш", который уникально идентифицирует изображение.

3. **Сравнение с базой данных:** Полученный хеш сравнивается с хешами в базе данных поисковой системы. Это сравнение помогает определить, есть ли в базе похожие изображения.
4. **Ранжирование и представление результатов:** Найденные совпадения ранжируются по уровню схожести и презентуются пользователю в виде списка изображений и их источников.

Примером сервиса, предоставляющего такую функцию, является Yandex Images.

Автоматизация использования Yandex Images Search

Yandex Images, как и многие другие сервисы реверсивного поиска изображений не предоставляет публичного API.

Чтобы автоматизировать поиск по изображениям в Yandex, можно воспользоваться подходом, представленным пользователем ark-key на Stack Overflow [8]. Чтобы получить ссылку на страницу с результатами поиска нужно сформировать ссылку https://yandex.com/images/search/?url={image_url}&rpt=imageview, подставив в неё публичную ссылку на изображение.

Далее можно воспользоваться технологией Web Scraping для извлечения результатов.

Web Scraping

Web Scraping — это процесс извлечения данных со страниц в интернете [9]. Данные, которые собираются с веб-страниц, могут использоваться для разнообразных целей, от аналитики до сбора контактной информации и цен.

Основные принципы Web Scraping включают:

1. **Запросы к веб-страницам:** Используя HTTP-запросы, инструменты для Web Scraping запрашивают содержимое веб-страниц, аналогично тому, как это делают браузеры.
2. **Анализ данных:** Полученное содержимое страницы анализируется для извлечения необходимой информации. Это может быть

выполнено с использованием таких методов, как регулярные выражения, HTML-парсеры и DOM-анализаторы.

3. **Извлечение данных:** После анализа данные извлекаются и сохраняются в структурированной форме (например, в CSV-файлах, базах данных).
4. **Автоматизация:** Web Scraping часто включает автоматизацию для периодического доступа к веб-страницам для обновления данных.
5. **Устранение проблем с доступом:** Веб-сайты могут применять различные меры для блокировки web scrapers, например, проверки на ботов, CAPTCHA или изменения структуры своих страниц. Разработчики скриптов для Web Scraping должны находить способы обхода таких препятствий.

Существует большое количество готовых решений для web scraping-a. Начиная от библиотек для большинства языков программирования, заканчивая полноценными SaaS приложениями. Например, Crawllee от компании Apify [10].

Глоссарий

Библиотека (Программная библиотека) - сборник ресурсов и программных инструкций, предоставляющих разработчикам стандартные функции, часто нужные для разработки программного обеспечения.

Микросервис - метод разработки программного обеспечения, при котором приложение разделено на множество небольших, независимых сервисов, каждый из которых выполняет одну функцию и взаимодействует с другими через легковесные механизмы, часто HTTP REST API.

API (Application Programming Interface) - набор определений и протоколов для построения и интеграции программного обеспечения, позволяющий взаимодействовать различным приложениям и микросервисам.

Blob Storage - сервис для хранения больших объемов неструктурированных данных, таких как текст, видео, аудио и изображения.

C# - современный, высокоуровневый язык программирования, разработанный Microsoft в рамках платформы .NET.

.NET - программная платформа от Microsoft, предоставляющая инструменты и библиотеки для разработки различных типов приложений, от веб до мобильных и настольных.

Web Scraping - процесс автоматизированного извлечения данных из веб-сайтов.

YouTube Shorts - короткие вертикальные видео, предлагаемые на платформе YouTube. Данный формат аналогичен TikTok или Instagram Reels и предназначен для быстрого потребления контента.

Список использованных источников

1. Social media: TikTok users in the United States // Statista (2024) - URL: <https://www.statista.com/study/72735/social-media-tiktok-users-in-the-united-states/> (дата обращения: 10.10.2023)
2. А TikTok-то тикает: как зумеры со своими короткими видео разрушают стриминги // Кинопоиск (2023) - URL: <https://www.kinopoisk.ru/media/article/4008419/> (дата обращения: 12.10.2023)
3. #Film Trending Hashtag on TikTok // TikTok Creative Center - URL: <https://ads.tiktok.com/business/creativecenter/hashtag/film/pc/en?period=7> (дата обращения: 11.10.2023)
4. Netflix запустил аналог TikTok с отрывками из фильмов и сериалов // 808 (2023) - URL: <https://808.media/netflix-zapustil-analog-tiktok/> (дата обращения: 13.10.2023)
5. A place to share knowledge and better understand the world // Quora - URL: <https://www.quora.com/> (дата обращения: 25.09.2023)
6. Can I find the name of a movie with a video clip? // Quora (2017) - URL: <https://www.quora.com/Can-I-find-the-name-of-a-movie-with-a-video-clip> (дата обращения: 15.12.2023)
7. Поиск изображений // Habr (2021) - URL: <https://habr.com/ru/articles/550528/> (дата обращения: 05.01.2024)
8. Reverse search an image in Yandex Images using Python // Stack Overflow (2020) - URL: <https://stackoverflow.com/questions/61978049/reverse-search-an-image-in-yandex-images-using-python> (дата обращения: 15.01.2024)
9. What is web scraping? // Apify (2023) - URL: <https://blog.apify.com/what-is-web-scraping/> (дата обращения: 30.01.2024)
10. Crawlee web scraping tutorial // Apify (2024) - URL: <https://blog.apify.com/crawlee-web-scraping-tutorial/> (дата обращения: 15.02.2024)