цифровой прорыв

сезон: ИИ

КЕЙС

Генерация тегов для видео

RUTUBE

Министерство экономического развития Российской Федерации

РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

Кейсодержатель

RUTUBE

01 Сфера деятельности

Видеохостинг

02 Краткое описание кейса

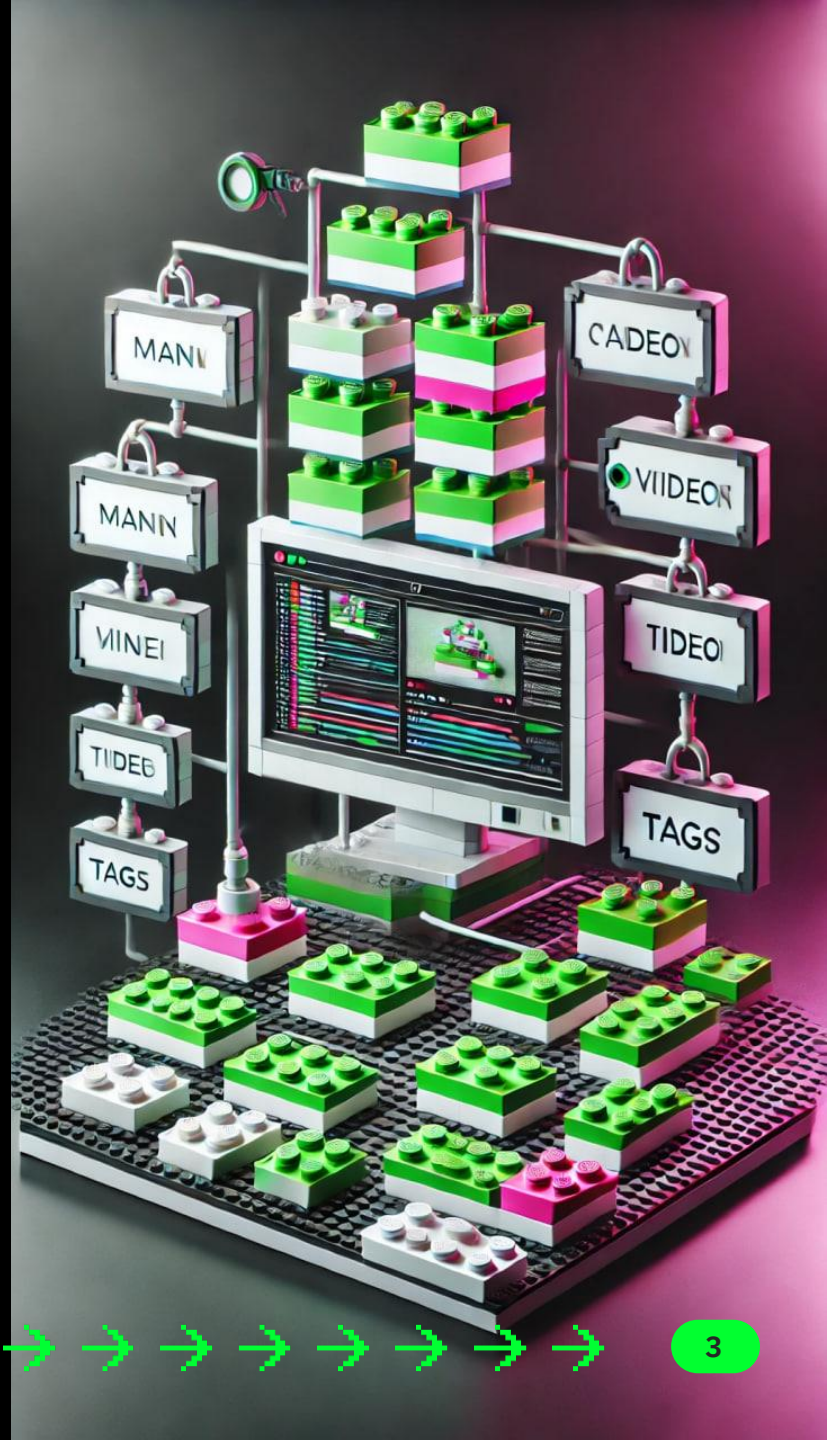
На основе доступного контента присвоить к видео теги из заранее известного иерархического списка тегов

Сайт организации

https://rutube.ru/

Постановка задачи

Необходимо создать систему тегирования видео на основе видео контента, названия и описания видео. Тегирование происходит по универсальному списку тегов для web платформ, широко затрагивающему различные тематики и подтематики. В решении участников протегированное видео может иметь тег родительской категории и тег подкатегории, соответствующий родительской. Видео может содержать несколько тегов из различных тематик.



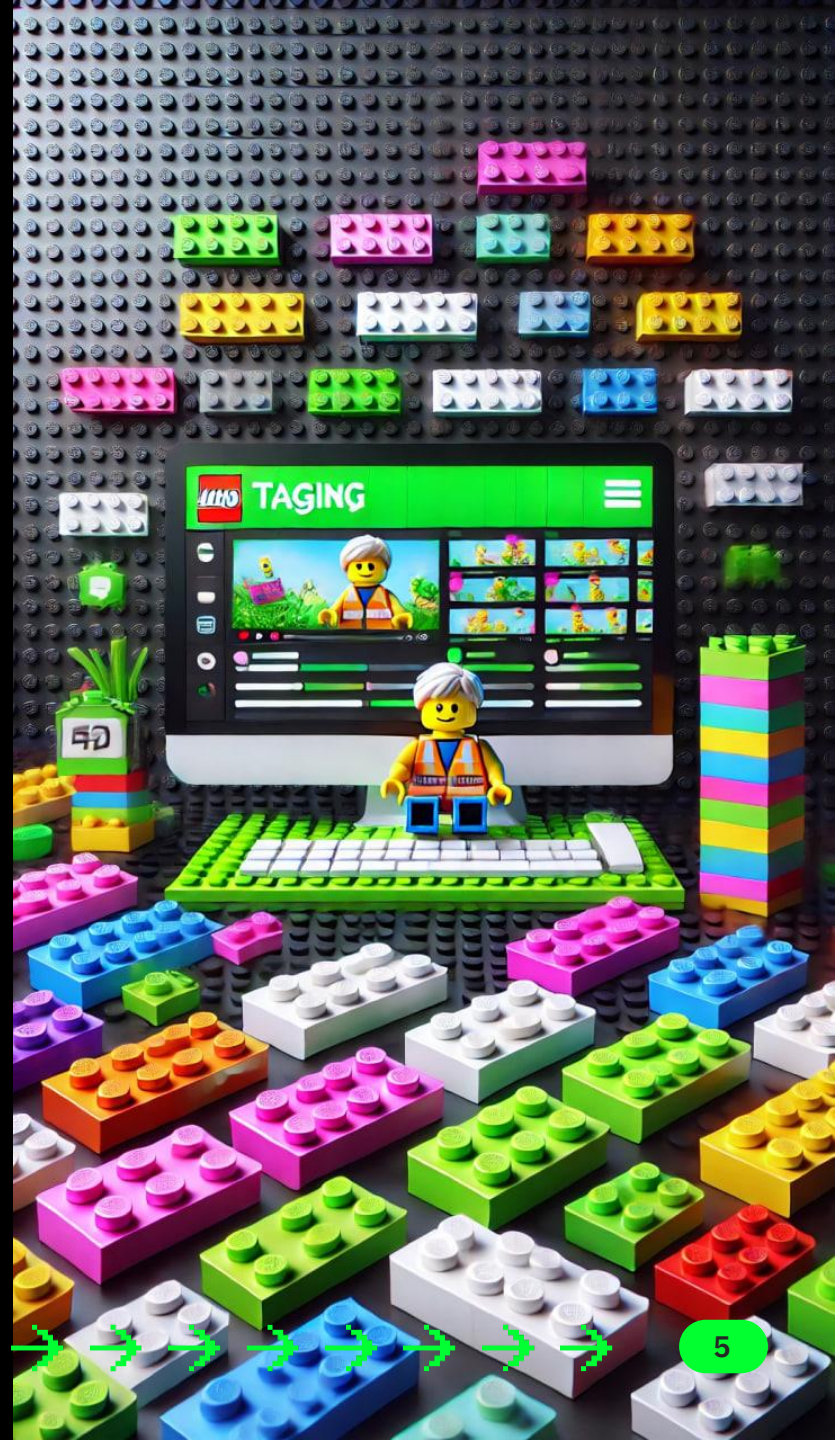
Проблематика

На платформу RUTUBE ежедневно заливаются сотни тысяч видео, большая часть которых - это ugc контент, то есть видео от обычных пользователей. Часть контента - это популярные шоу, передачи, каналы, другой лицензионный контент. Чтобы упорядочить весь этот контент, необходимо создать систему тегирования видео, чтобы разделять их по категориям и подкатегориям. Причем система должна быть достаточно гибкой к обновлению списка тегов, широко и разнообразно покрывать контент. Такая система также улучшит рекомендательную систему, так как возможно будет рекомендовать контент из любимой категории пользователя, например.

Решение

Прототип системы, создающей теги для видеоконтента. На вход приходят видео (или видео id), название и описание видео. На выходе - список вероятных категорий (и подкатегорий), предсказанных в соответствии с заранее известным иерархическим списком тегов.

Необходимо иметь репозиторий с понятным и качественным кодом модели, оформленным README. Реализовать прототип возможно в виде веб сервиса, демонстрирующего способности алгоритма (в рамках критериев интерфейс веб-сервиса не оценивается).



Стек технологий, рекомендуемых к использованию

01

Язык программирования - Python
Библиотеки для использования - torch, pandas, tensorflow, любые доступные в opensource
Обязательные условия - решение должно работать без доступа к интернету

Необходимые данные, дополнения, пояснения, уточнения

02

В названиях и описаниях к видео уже может содержаться некоторая информация о видеоконтенте, но часто информация не совсем релевантна или неполна. Поэтому важно научиться работать с видеорядом, извлекать важные или часто повторяющиеся сущности, которые можно соотнести со списком тегов. Можно пробовать суммаризировать видео или извлекать сущности из полного видеоряда. Немаловажную роль может играть аудиодорожка, которую можно преобразовать в текст и извлечь теги оттуда, но не забывайте учесть, что часть видео могут не содержать речи совсем или речь может не соответствовать видеоряду. Список тегов - это также текстовые данные, которые можно преобразовать в вектора или при желании - разметить под свои нужды.

Оценка

- Для оценки решений применяется метод экспертных оценок и автоматизированные средства оценивания.
- Жюри состоит из отраслевых и технических членов жюри.
- На основании описанных далее характеристик, жюри выставляет оценки.
- Возможность скачивания тестового датасета с паролем открывается за 12 часов до стоп-кода.
  Возможность отправки сабмитов и пароль открываются за 4 часа до стоп-кода.
  Интервал успешных отправок: 20 минут.
- Итоговая оценка определяется как сумма баллов всех членов жюри, суммируемая с оценкой автоматизированной системы, нормализованной в 25% от итоговой оценки.

Технический член жюри оценивает решение по следующим критериям:

01

Документация и комментарии к коду

Шкала: 0-2-4-6

02

Обоснованность выбранного метода (описание подходов к решению, их обоснование и релевантность задаче)

Шкала: 0-1-2-3

03

Решение использует фичи с видео

Шкала 0-2-4

04

Прозрачность решения

Шкала 0-1-2

05

Выступление команды (умение презентовать результаты своей работы, строить логичный, понятный и интересный рассказ для презентации результатов своей работы)

Шкала 0-1-2-3

Автоматизированные средства оценивания точности работы предложенных участниками алгоритмов (решений) выставляют оценку в диапазоне 0-1, где 1 равно 100% точности работы решения.

Итоговая оценка определяется как сумма баллов всех членов жюри, суммируемая с оценкой автоматизированной системы, нормализованной в 25% от итоговой оценки.

Метрика: IoU

Отраслевой член жюри оценивает решение по следующим критериям:

01

Качество иерархического тегирования

Шкала 0-2-4

02

Качество работы решения в различных тематиках

Шкала 0-2-4

03

Скорость работы решения

Шкала 0-1-2

04

Выступление команды (умение презентовать результаты своей работы, строить логичный, понятный и интересный рассказ для презентации результатов своей работы)

Шкала 0-1-2-3

цифровой прорыв

сезон: ИИ

Министерство
экономического развития
Российской Федерации

РОССИЯ –
СТРАНА
ВОЗМОЖНОСТЕЙ