



Название проекта

Embedding based User Generated Content search

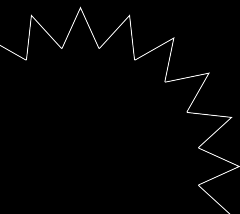
Команда проекта
ext



Проблема



**Сложно вычленить имена знаменитостей из
аудио- и видео ряда не зная заранее, кто
именно и в какой момент там будет
присутствовать**





Подход к решению

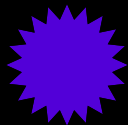
§1. Извлечем эмбединги из набора знаменитостей датасета VoxCeleb1

§2. Порежем тестовый клип на небольшие кусочки и сделаем для них векторное представление

§3. Найдем похожие по косинусному расстоянию между векторами



И, собственно, ожидаем Profit



Путь решения



Задачи

Сделано (оценка готовности)

Сервис для скачивания и хранения результатов локально

+*

Загрузчик результатов в облако

+*

Докеризация решения

+

Извлечение аудио и видео из mp4

+* (audio) / —(video)

Векторное представление для данных (в т.ч. векторов аудио/видео/текста) для тестового файла

— — +

Извлечение эмбеддингов из датасета со «звёздами» (VoxCeleb1)

—

Поиск схожих векторов

—

Рефакторинг кода и пр.

—

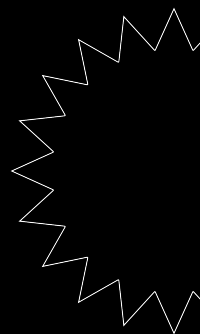
ИТОГО

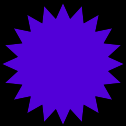
(+3.7 vs —6.3) из 10




* - страдает асинхронность

СБЕР ЗВУК






Альтернативные рассматривавшиеся решения



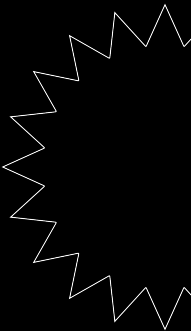
Подход к решению

Для каждого типа данных решать долго, поэтому попробовать построить векторные представления сразу над бинарными данными

Сделать акцент на работе с чанками исходного файла на предмет получения более универсальных векторных представлений, учитывающих историчность данных

- Добавить ASR и с помощью распознанного текста обогатить эмбединги
 - Добавить регулярных выражений над текстами для поиска именованных сущностей (количество “звездных” имен) ограничено
- 

Почему отказались?

- Плохо разработаны способы такого преобразования либо не удалось их быстро науглить
 - Нужно учиться подбирать для ядер сверток параметры (похоже, такие ядра должны быть сложнее чем используемые для изображений?..)
 - Нужна стандартизация данных (зависимость бинарных данных от форматов сжатия аудио и видео)
 - В итоге – оценка временных затрат – решать ещё дольше, чем стандартными методами
-
- Не успел
- 



Feedback себе и организаторам




Укоры себе

- Оказалось, что на решение простых задач, которыми не занимаешься в повседневной работе требуется значительно больше времени
- Нельзя недооценивать ценность наличия команды для решения задачи в сжатые сроки

Пожелания организаторам

Неожиданной оказалась подготовительная часть, не касающаяся самой задачи непосредственно – подготовка «инфраструктуры»/обвязки для обмена решением с организаторами

Наличие бойлерплейтов помогло бы больше внимания сосредоточить на самой задаче нахождения и фильтрации пользовательского контента



Прототип (видео)

