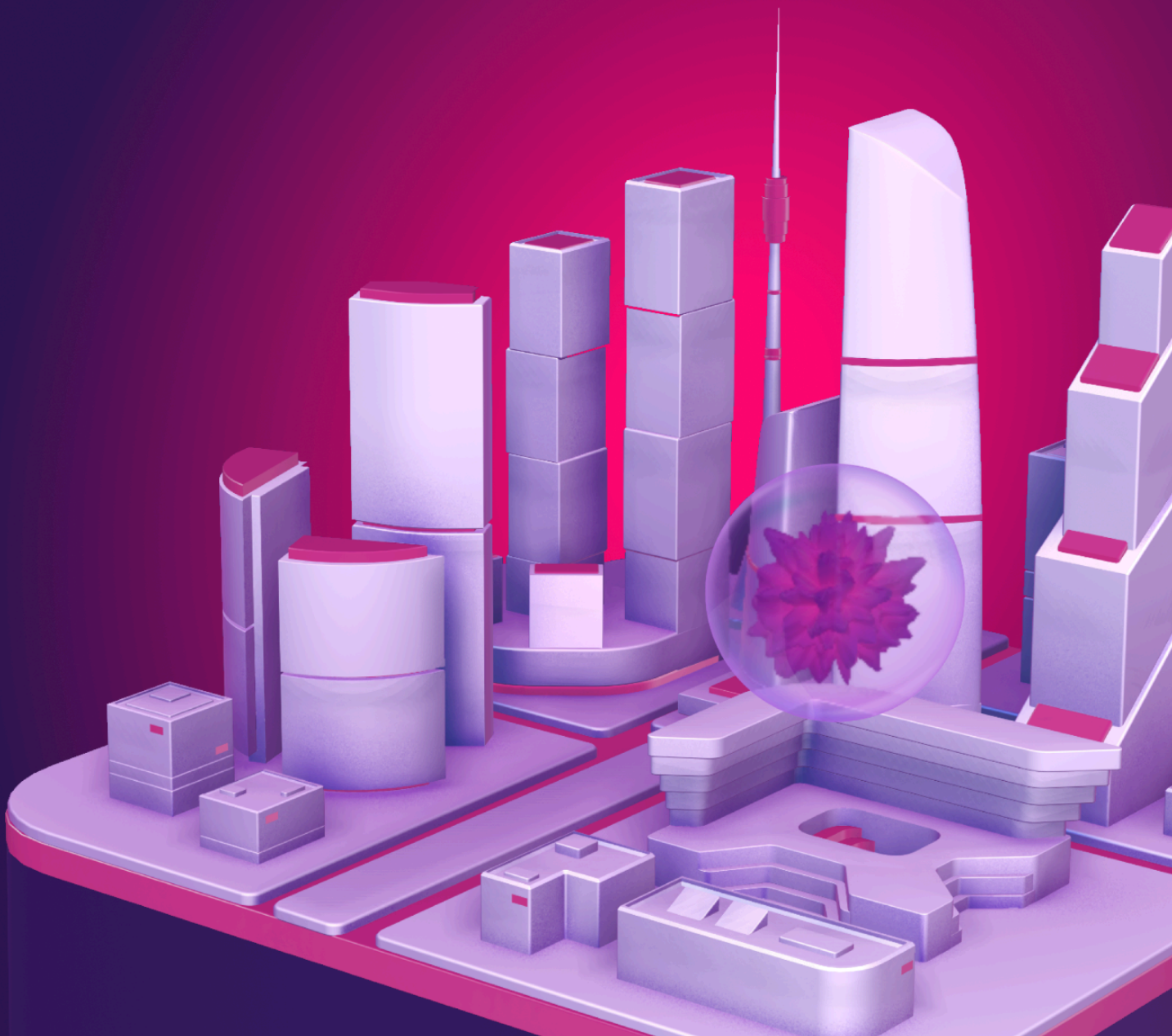
ЛИДЕРЫ ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Yappy

ТЕХНИЧЕСКОЕ ЗАДАНИЕ

ЗАДАЧА 15

Сервис текстового поиска по медиаконтенту



ТЕХНИЧЕСКОЕ ЗАДАНИЕ

Сервис текстового поиска по медиаконтенту

1.Актуальность задачи

Функция поиска имеет актуальность в современном мире из-за большого объема информации, доступной в интернете. Пользователи ищут информацию о товарах, услугах, новостях и других темах, и функция поиска позволяет им быстро и удобно находить нужные данные. Кроме того, функция поиска используется на многих веб-сайтах и приложениях для повышения удобства пользования и улучшения пользовательского опыта.

“Yappy” представляет собой социальную сеть коротких вертикальных видео. Авторы загружают видео на платформу, обогащая их описанием. Задача поиска контента по запросу является актуальной для социальной сети “Yappy”. Качество и скорость выдачи контента по поисковым запросам является неотъемлемой функцией приложения. С помощью поиска пользователи могут находить интересный для них контент, открывать неизведанные ранее категории контента, вдохновляться новыми идеями, смотреть актуальные ролики из разных категорий.

Текущая версия приложения уже имеет функцию поиска. Сейчас поиск осуществляется за счет поиска подобных слов в описании под видео, которые оставляет пользователь. Описание включает в себя простой текст и хештеги. Хештег - это текст с символом “#” перед ним, который кратко описывает основной смысл видео, например: #еда, #рецепты, #вкусно, #ням-ням. Очень часто авторы контента оставляют некорректное описание и хештеги под видео. Это влияет на поисковую выдачу. Необходимо включить в поиск больше признаков, особенно тех, которые можно достать непосредственно из видео, например, используя модели машинного и глубокого обучения.

Ваше решение должно усовершенствовать текущую систему поиска, сделать выдачу контента более качественной. Это увеличит популярность этой функции среди пользователей социальной сети, и, в свою очередь, увеличит общее времяпрепровождение пользователей в приложении.

Целевая аудитория сервиса - молодые люди возраста от 14 до 35 лет.

2. Описание задачи

Участникам предлагается решить следующий кейс - создать сервис “Умного поиска” для социальной сети Yappy. Социальная сеть “Yappy” - это платформа коротких вертикальных видео длительностью до 1 минуты,

загружаемых пользователями-авторами. Каждое видео обогащается описанием, которое составил сам автор.

Пользователь в поисковой строке приложения вводит некоторый текст. Необходимо для его текста отдать наиболее релевантные видео из всего их многообразия. На текущий момент поисковая система учитывает только текст из описания под видео. Заметим, что для успешной работы поисковой системы необходим некоторый индекс. Под индексом понимается специальным образом подготовленные признаки для каждой сущности - видео, по которым осуществляется поиск. Вы можете разработать любые признаки и использовать любые инструменты для хранения этой информации, за исключением проприетарных инструментов с закрытыми лицензиями. Допускается использование инструментов только с открытым исходным кодом.

Фактически задача разбивается на две подзадачи:

1. Обработка нового видео. Обогащение его дополнительными признаками с помощью специальных алгоритмов, моделей машинного обучения. Добавление этой информации в индекс - хранилище.
2. Непосредственно поиск. Выдача наиболее релевантного контента на текстовый запрос от пользователя. При этом надо учитывать, что пользователь может вводить несколько букв, слова, словосочетания, предложения, любые комбинации символов.

Предоставленные для решения задачи данные состоят из видеофайла формата mp4 и описания видео в текстовом формате. Система поиска может учитывать разные сущности, такие как: описание пользователя под видео, речь в самом видео, текст в самом видео, описание видео, полученное какой-либо системой или нейронной сетью. Генерация таких признаков может осуществляться с помощью любых open-source или самописных решений/инструментов.

Лучший поисковый сервис может быть интегрирован в социальную сеть “Yappy”.

3.Возможный пользовательский путь

Пользователь заходит в социальную сеть “Yappy”, переходит на вкладку поиска и вбивает свой запрос в поисковую строку. Это может быть набор букв, слово, словосочетание или предложение, набор бессвязных символов. Цель поисковой системы отдать ему наиболее релевантные видео из всего многообразия видео, загруженных на платформу. При этом на поисковую выдачу пользователь может налагать фильтры, сортировки. Например, сортировка выдачи по дате добавления, сначала популярное и т.п.

Для пользователя наиболее важным параметрами поиска в приложении является: качество выдачи контента и скорость работы системы.

4.Требования к решению

Решение должно представлять собой сервис, реализующий поисковую систему. Код для создания сервиса может быть написан на любом языке программирования, однако наиболее предпочтительные варианты это Python или Go. При разработке системы поиска разрешается использовать открытые библиотеки, инструменты, алгоритмы, модели машинного обучения. Использование закрытых сервисов, осуществляющих работу по подписке или оказывающих платные услуги, а также работающих по закрытой для коммерческого вида деятельности лицензии, запрещено.

К обязательным условиям итогового решения относятся:

1) Наличие API системы поиска с двумя основными ручками - методами публичного интерфейса. Одна ручка принимает на вход видео с описанием, а точнее ссылку на видео и описание. Извлекает признаки с помощью какого-либо набора алгоритмов. Кладет всю необходимую информацию в индекс-хранилище. Другая ручка принимает на вход поисковый запрос в виде текстовой строки и выдает список видео в ответе.
2) Скорость работы системы. На прием видео и индексацию выделяется не более 5 минут. На выдачу поискового запроса не более 500 миллисекунд (0,5 секунды).
3) Снабдить решение кратким описанием логики работы системы, основных алгоритмов и использованных инструментов.
4) Развернуть свой сервис для осуществления быстрых поисковых запросов к нему. Под сервисом подразумевается поднятый бэкенд с активными двумя ручками.

Будет плюсом:

1) Придумать подсказки для автозаполнения запроса пользователя. Пример. Пользователь вводит буквы “маш”. Ему в качестве подсказок выдается перечень из автодополнения: “Маша и медведь”, “машина”, “машина от Xiaomi”, “машинка для стрижки волос”. В качестве иллюстрации можно рассмотреть примеры выдачи при поисковом запросе в Google/Yandex.
2) Описать дальнейшее развитие системы. Даже если что-то не получилось сделать в рамках времени Хакатона, можно снабдить решение идеями по улучшению/доработке своей системы.
3) Описать ограничения текущей системы по количеству принимаемых запросов в секунду, скорости расчета/пересчета индекса и т.п.
4) Придумать инструмент исправления опечаток для пользователя. Например, пользователь вводит слово “машина”, при этом алгоритм

5) исправления опечаток понимает, что вводимое слово “машина” и осуществляет поиск по исправленному слову.

5.Требования к презентации

Презентация итогового решения присылается в формате pptx/pdf. К обязательным слайдам относится:

1) Информация о команде;
2) Описание преимуществ и недостатков выбранного решения;
3) Описание архитектуры и инструментов для решения;
4) Описание основных алгоритмов и моделей машинного обучения, задействованных при решении задачи;
5) Указание скорости работы системы для двух основных задач: добавление нового видео в индекс, непосредственный поиск.

6.Требования к презентации

Сопроводительная документация направляется в формате docx/pdf. Обязательными пунктами в документации является:

1) Описание структуры проекта;
2) Перечень инструментов/языков программирования/библиотек/open-source решений, использованных в работе;
3) Скорость работы системы в разрезе двух подзадач: вставка нового элемента в базу данных/индекс; ответы на текстовый запрос пользователя.
4) Идеи и гипотезы об улучшении/доработках системы, которые не успели сделать в рамках Хакатона.
5) Схема логики взаимодействия всех модулей внутри проекта.
6) Указание уникальности решения.

7.Источники данных

Организатор предоставит набор данных в 400.000 видеозаписей с описанием. Видео в датасете будут представлены как ссылка на общедоступное хранилище.

Пример данных:

```
{
    “video_link”: https://cdn-st.ritm.media/media/00/5d/fa5bcb3d40479d06d416d43a62ba/fhd.mp4,
    “description”: "А ваши коты, тоже не дают спать по ночам?
```

#топ #марафонконтентауappy #ямарафонец #кот"
}
```

Видео будут доступны на всем протяжении Хакатона, скачивать их необязательно.

## 8.Источники данных

Решение должно быть оформлено в единый репозиторий с кодом.

1) Ссылка на открытый репозиторий с кодом сервиса поиска
2) Ссылка на презентацию решения
3) Ссылка на развернутый прототип решения
4) Ссылка на сопроводительную документацию в формате (docx/pdf)

## 9.Критерии оценки

Решения участников будут оцениваться по следующим основным критериям:

1) Подход коллектива к решению задачи (полнота реализации поисковой системы, идея решения задачи, набор выбранных алгоритмов и моделей машинного обучения) - 10 баллов;
2) Техническая проработка решения:
   а) Скорость работы сервиса по двум основным задачам - 10 баллов;
   б) Работоспособность решения, поддерживаемость решения, качество кода (10 баллов);
   в) Инновационные идеи по улучшению опыта взаимодействия пользователя с поисковой системой приложения (оригинальные решения и предложения по усовершенствованию, решение доп. задач) - 8 баллов;
3) Соответствие решения поставленной задаче - 5 баллов.
4) Эффективность решения в рамках поставленной задачи. Будет оцениваться качество поисковой выдачи (при оценке поисковой выдаче будет вбиваться несколько фраз/слов. При этом определяется релевантность поисковой выдачи из топ-10 видео по каждой фразе. Каждому видео из выдачи ставятся очки качества: 0 - совсем не похоже, 1 - не похоже, 2 - похоже, 3 - очень похоже. Таким образом, по каждой фразе можно получить от 0 до 30 очков. Суммарная оценка будет выставляться как сумма очков по всем фразам. Например, 5 фраз с очками: 0 + 12 + 18 + 30 + 20 = 80 очков. Итоговая оценка: 80 очков. Далее очки будут переводиться в баллы согласно занятым местам каждой команды) - 10 баллов;
```

5) Выступление коллектива на питч-сессии (убедительность, информативность, лаконичные и аргументированные ответы, соответствие регламенту) - 10 баллов;