Система сбора и хранения данных о просмотрах каналов (роликов) Youtube

# Цели и задачи

Цель системы состоит в регулярном сборе и хранение значимых метрик роликов и каналов Youtube с обеспечением доступа к этой информации с помощью систем визуализации (powerBI) или других систем.

Основной потребитель данных – аналитики, маркетинг.

# Технологический стек системы

Python 3.10, PostgreSQL, Selenium

(отладка и разработка jupyter notebook)

# Структура и краткое описание системы

Для хранения данных используется БД PostgreSQL. Она состоит из 4х таблиц. Текущее состояние «частично» нормализованная по 2ой форме. В таблицах хранится:

1. список youtube каналов (yt\_channels)
2. список и описание youtube роликов (yt\_reels)
3. принадлежность роликов к каналам (yt\_channels2reels)
4. статистика по роликам (yt\_reels\_stat).

Выгрузка данных о просмотрах осуществляется с сайта popsters.ru (платная подписка) скриптом get\_popsters\_sql.py, который получает с сайта xlsx файлы с данными по роликам на канале. Список каналов и ссылки на них скрипт получает из таблицы БД. Также скрипт генерит два файла - report\_urls.csv – файл, в котором описывается привязка xlsx файлов и каналов, по которым они содержат информаицю и файл failed\_yt.csv – список каналов, по которым статистика (по каким-либо причинам) не была получена (выгрузка данных не произошла).

Вторым скриптом – yt\_parser\_loader.py – производится парсинг файла report\_urls.csv и последовательная обработка xlsx файлов. В ходе которой производятся следующие операции:

1. получение списка роликов и обновление таблицы роликов
2. получение списка роликов на данном канале и обновление таблицы yt\_channels2reels.
3. Обновление таблицы со статистикой по роликам yt\_reels\_stat

После обработки всех xlsx файлов (всех данных собранных по каналам) производится обновление таблицы yt\_reels\_stat – дублирование последних данных статистики на данную дату сбора информации и разметка удаленных роликов.

# Структура системы и БД (ERD)



# Работа скриптов

## Скрипт get\_popsters\_sql.py

Требования: python 3.x, настройки для подключения к БД системы, установленный драйвер selenium, наличие учетной записи в браузере с логином в систему popsters.ru.

Функции в скрипте:

def get\_url(url, t\_delay) – функция выгрузки статистики с сайта popster.ru в xlsx файл. На входе url ссылка (техническая, c ID канала) на канал для выгрузки данных, t\_delay – временная задержка до закрытия страницы (ставить не менее 90 сек).

def main(filename): - функция формирования списка и формирования файла отчета. Аргумент filename – если не указан, данные берутся из SQL таблицы, если указан, то такие же данные берутся из указанного файла (сделано было для того, что если по каким-то ссылкам данные не выгружены – то в качестве аргумента передается файл failed\_yt.csv для дополнительных попыток и анализа причины ошибки).

## Скрипт yt\_parser\_loader.py

Требования: python 3.x, настройки для подключения к БД системы, файл отчета report\_urls.csv, xlsx файлы с выгруженной статистикой, файл ./ref/vocabulary.csv словаря категоризатора роликов

Функции в скрипте:

Функции парсинга названий

def check\_items\_cartoon(reel\_name): функция по названию ролика формирует список обнаруженных вхождений в словарь и возвращает список

def check\_len\_cartoon(collection\_list): просчет числа записей в списке обнаруженных вхождений и возвращает ‘none’ если запись одна или ‘сборник’, если записей несколько

def convert\_collect(collection\_list): преобразует список записей в одну строку, если записей нет ( none) – то возвращает none

def alphanum(element): «очистка» строки от спецсимволов и тп. Остаются буквы и символы []":.,//"

## Функции обновления таблиц БД

def proc\_reels\_to\_sql(reels\_to\_sql): Функция добавления роликов в таблицу yt\_reels, которых там нет. На входе dataframe с новыми роликами, которые функция добавляет в таблице yt\_reels. В функции происходит парсинг названия и соответствующая категоризация по результатам парсинга.

proc\_ch2reels\_to\_sql(ch2reels\_to\_sql, ch\_id, fdate): Функция добавления роликов в таблицу yt\_channels2reels. На входе dataframe с новыми роликами, ch\_id -идентификатор канала и дата, которая будет проставлено в поле 'yt\_reel\_date\_add – дата добавления ролика на канал.

proc\_ch2reels\_update(removed\_reels, ch\_id, fdate): Функция обновления таблицы yt\_channels2reels. Для роликов, которые перестали присутствовать на канале ставится дата «удаления» с канала в поле yt\_reel\_date\_removed. На входе dataframe с удаленными роликами, ch\_id – идентификатор канала, fdate – дата.

proc\_new\_reels\_stat(reels, ch\_id, fdate): Функция добавления статистики новых роликов в таблицу yt\_reels\_stat. Reels – dataframe c информацией о роликах, fdate – дата информации.

proc\_exist\_reels\_stat(reels, ch\_id, fdate): Функция добавления статистики роликов, которые уже есть в таблице yt\_reels\_stat. Reels – dataframe c информацией о роликах, fdate – дата информации.

process\_file(fpath, ch\_id): Функция парсинга XLSX файлов со статистикой и формирования данных для вызова функций обновления таблиц. Считываются данные из файла. После чего происходит последовательное сравнение с данными из таблиц yt\_reels (загрузка данных при обработке каждого файла XLSX), yt\_channels2reels (загрузка данных только по ch\_id при обработке каждого файла XLSX), yt\_reels\_stat (данные из таблицы загружаются один раз вначале скрипта).

Остальной код скрипта строится на парсинге файла report\_urls.csv, последовательном вызове обработчика файлов XLSX (данные по которым – путь и id канала берутся из файла report\_urls.csv). И в завершение происходит обновление таблицы yt\_reel\_stat – для роликов, по которым больше нет статистики (считаем что они удалены с youtube каналов) – статическая информация копируется из последних имеющихся данных, дифференциальная устанавливается равной нулю, устанавливается поле removed в значение 1.

# План развития системы

Июль 2022 года – запуск пробной версии системы

Декабрь 2022 года – Приведение системы к стандартам: нормализация БД, реализация новой информации по структуре youtube (например, что ролик может быть на нескольких каналах), сохранение статистики по удаленным роликам.

🡪

Март 2023 года – отказ от popsters.ru. Использование API youtube для получения данных, в том числе тех, что не выгружаются popsters.ru

Апрель 2023 года – внедрение NLP системы для категоризации роликов (парсинга текстовых описаний)

~~Май-июнь 2023 года – исследования информации доступной через API youtube для собственных каналов Студии. Формирование новой БД – уникальные зрители, частота просмотров и другая статистика (зависит от того что доступно через API)~~

III-IV квартал 2023 года – оперативная система получения доступной статистики (пользовательский интерфейс для получения статистики на момент запроса).