Домашнее задание №2

Курс «Цифровой архив»

Горбачев Николай

Сохранение коллекции веб-сайтов, отобранных по общей тематике: чарты популярных русскоязычных подкастов на подкаст-платформах и агрегаторах аналитики

1. **Что это такое?**

Коллекция популярных русскоязычных подкастов – набор ресурсов в открытом доступе, агрегирующих информацию о топ-популярных подкастов в данный момент, данные о которых собраны с различных ресурсов (с российских и зарубежных подкаст платформ, а также сервисов сбора интернет-аналитики)

В коллекцию входят:
- Чарт подкастов на Яндекс-музыке
- Стартовая страница подкастов Google (за неимением онлайн-чарта в Google подкастах)
- Чарт топ-подкастов castbox.fm по региону Россия
- Аналитика/чарт топ-подкастов Apple music с портала Chartable (чарты Apple music доступны в открытом доступе но только в приложении на macos/ios)
- Чарт топ-подкастов Deezer.com по подвыборке русскоязычных подкастов (можно заметить что выборка определена неточно и большинство подкастов в чарте не на русском языке)
- Стартовая страница zvuk.com (за неимением онлай-чарта в открытом доступе)

Предполагается что коллекцию популярных русскоязычных подкастов можно использовать для анализа динамики рынка подкастинга для русскоязычной аудитории: динамика тем и тональности (сентимента) популярных в целевой аудитории подкастов, анализ рейтингов, метаданных, звуковых метрик, сетевой анализ студий подкастов, тем и персоналий.

2. **Почему вы сохраняете этот объект?**

Мотивация для сохранения коллекции популярных русскоязычных подкастов могла бы состоять из несколько ключевых аспектов:

1. Сохранение информации и анализа рынка может иметь множество применений как в академической среде так и в индустрии, тем не менее многие данные о подобной аналитике скрыты (напр. отсутствие русскоязычного чарта в Spotify, отсутствие чарта на платформе Zvuk), тк предположительно большие

стриминг-сервисы не заинтересованы в сборе данных по разным соображениям. Притом исторические (не о реальном времени) данные фактически отсутствуют и исследовать динамику рынка фактически невозможно не имея архива.

2. Ученые, исследователи и аналитики могут быть заинтересованы в изучении эволюции рынка подкаст-индустрии с течением времени. Архив коллекции русскоязычных подкастов может предоставить ценные данные для сравнительных исследований рынка.
3. В силу специфики бизнес-среды для платформ в разных странах, а также в силу социальных и политических причин чарты подкастов на различных платформах разные (также как и набор подкастов на площадках) – предполагается что имея общий инструмент аналитики в исследованиях можно получить более сбалансированное представление о запросах аудитории и предложении на рынке.
4. Стремительный темп развития digital-среды а также резкие изменения в социального и политического климата заставляет сомневаться в долговечности открытого доступа к тем или иным данным - как самих подкастов на платформах так и данных о них в открытом доступе

3. Зачем он был создан и кто будет использовать его в будущем

Идея создания подобной коллекции состоит в том, что, учитывая развитие индустрии подкастов на русском языке за последние несколько лет, а также учитывая набор тем обсуждаемых в них (как бытовых, так и социальных) – подкасты сами по себе представляются важнейшим источником данных для цифровых социальных исследований (Social Data Science). Предполагается, что подобную коллекцию могли бы использовать исследователи в самых различных направлениях

- Исследователи в социальных науках (используя данные о динамике популярных тем и тональности в аудитории, сетевой анализ)
- DH-исследователи, рассматривая динамику популярных подкастов как культурных объектов
- Исследователей digital-среды рассматривая развитие подкаст-индустрии как таковой
- Компьютерные лингвисты, рассматривая данные из карточек-описания подкастов и звуковых данных из выпусков (напр. на предмет новой подкаст-стилистики письменной речи между публицистическим и разговорным стилем)

Кроме того, представляется, что подобный ресурс мог бы быть полезным внутри самой индустрии:

- Исследователям-маркетологам для изучения динамики популярности подкастов в зависимости от тем, от описаний, от превью и других признаков
- Продюсерам для изучения практик записи и сведения подкастов среди подвыборки из самых популярных из них

- Авторам для изучения спроса, предложения и специфики рынка в данный момент

Наконец, сами слушатели могли бы использовать такой ресурс для знакомства с новыми подкастами

4. **Что вы можете узнать о рисках для данного типа объекта?**

Среди рисков сопряженных с созданием архива коллекции русскоязычных подкастов, можно было бы выделить несколько основных:

- Точность и актуальность данных, хранящихся в архиве, могут представлять собой риск. Легко пронаблюдать, насколько отличаются чарты друг о друга на разных платформах в силу разным причин – воспользовавшись коллекций из нескольких ресурсов исследователь может сделать выборку более сбалансированной но опираться на приведенные данные можно только с оговоркой на заведомую смещенность данных
- Быстрое изменение цифровой среды – фактически невозможность гарантировать поддержание работы подобного ресурса долгое время
- Необходимость в постоянном обновлении архива, что в свою очередь требует ресурсов

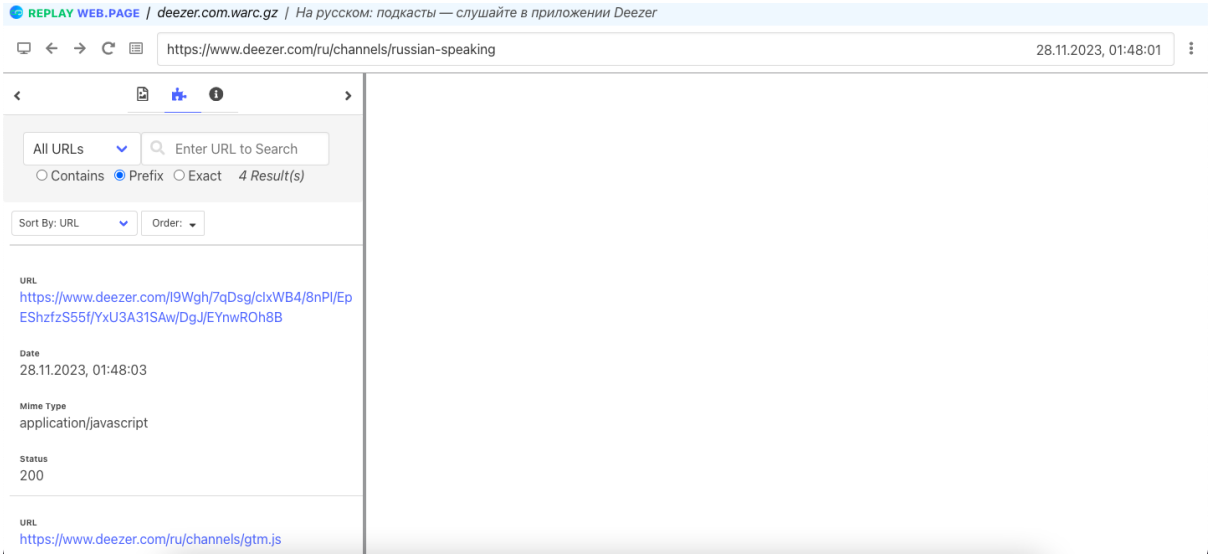
5. **Ход работы**
    a. Для сохранения описанных выше ресурсов мы воспользовались скриптом массовой загрузки из нескольких ресурсов через wpull
    b. Из-за специфики ресурсов обнаружилось несколько затруднений в загрузке из первоисточников
        i. стартовая страница Google-подкастов https://podcasts.google.com/ не имеет особого субдомена для выбора языка – возможна выдача другая чем в браузере системы
        ii. страница чарта на castbox.fm подгружается не через субдомен а в javascript -> с сайта не загрузился ни один документ (страница чарта https://castbox.fm/categories/0?utm_campaign=ex_share_ch&utm_medium=exlink&country=ru)
    c. Через ~ полтора часа работы скрипта прерываем процесс вручную
    d. Обнаружилось что архивы по Google подкастам и Zvuk.com загрузили более 250мб данных, остальные архивы подгрузили не больше 50мб данных, Deezer загрузил меньше 0.5мб
    e. Перезапускаем архивацию для чарта Deezer - https://www.deezer.com/ru/channels/russian-speaking а также для чарта castbox, на этот раз без фильтра по региону:

https://castbox.fm/categories/0, выбираем limit уровней рекурсии 5 для ограничения зоны сайта для архивации

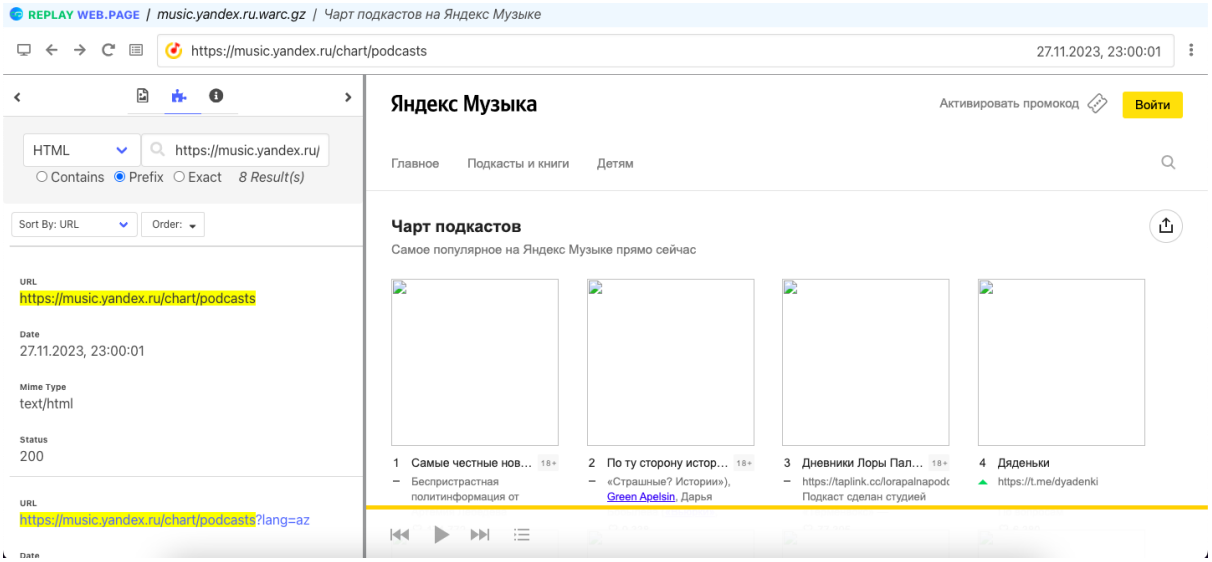
6. Результаты

Итоговые результаты разнятся по качеству и количеству от ресурса к ресурсу, проведем проверку с помощью сервиса https://replayweb.page/

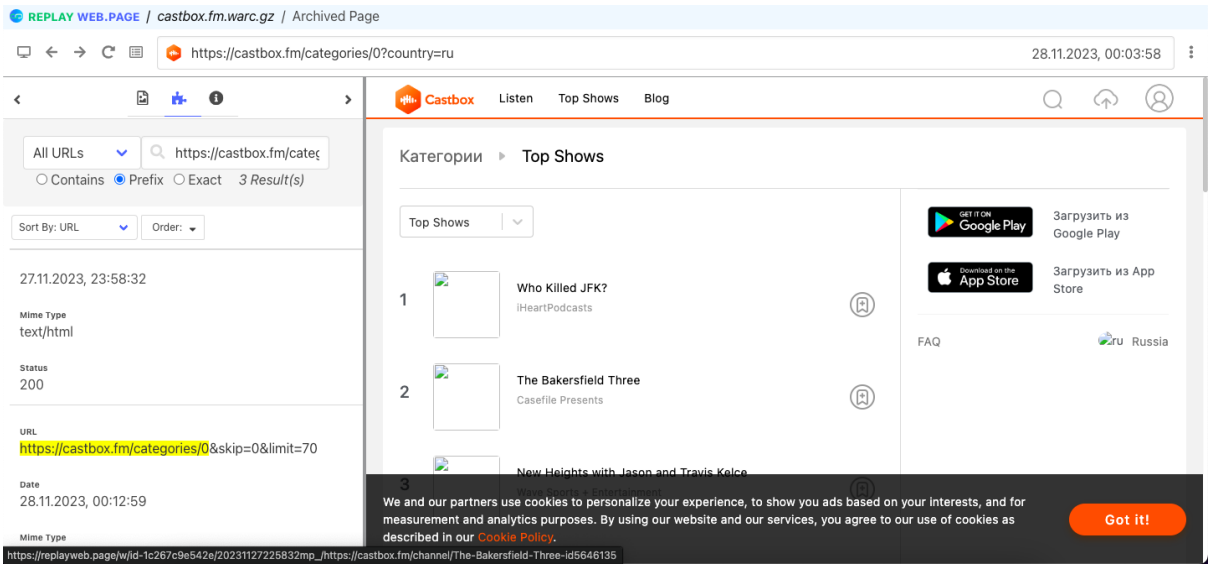
1. Чарт Deezer не удалось загрузить ни в одну из двух попыток – предположительно сервис блокирует скачивание wpull или запрос wpull составлен не вполне подходящим образом для ресурса



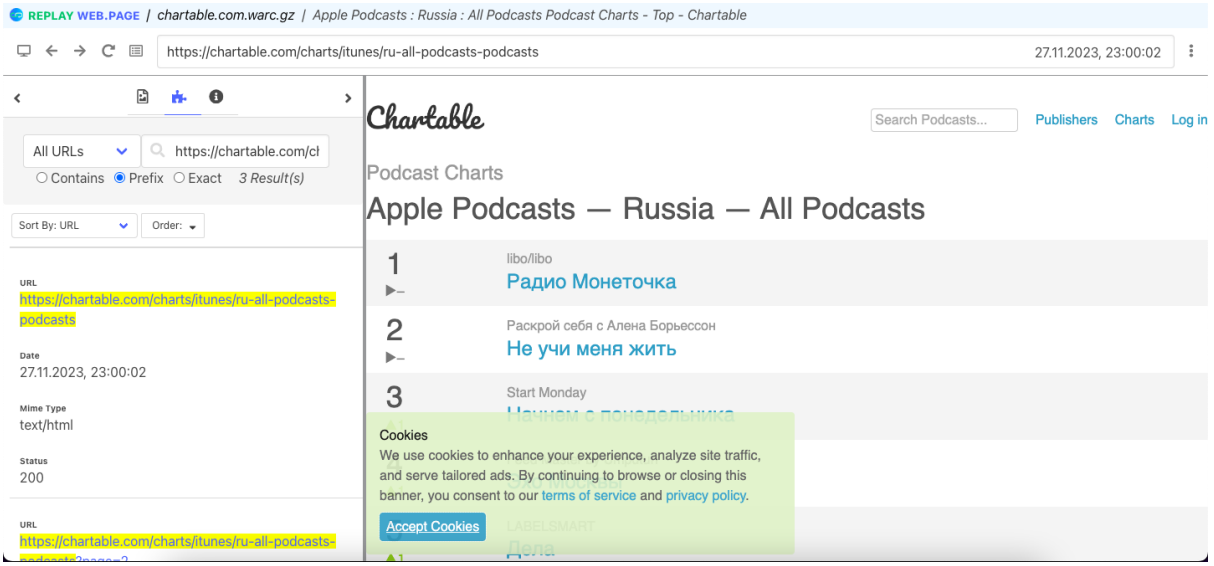
2. Подкасты на Яндекс музыки архивировались относительно успешно: к сожалению, рекурсия ушла в служебные страницы и перейти по ссылкам на превью подкастов не удается (а также не прогрузились картинки), но доступны названия, авторы, позиция в чарте и количества лайков



3. Castbox архивировался менее удачно: тк выбор региона/языка определяется через javascript, оказалось wpull успешно архивировал международный топ, но дальнейшие шаги рекурсии посвятились служебным страницам

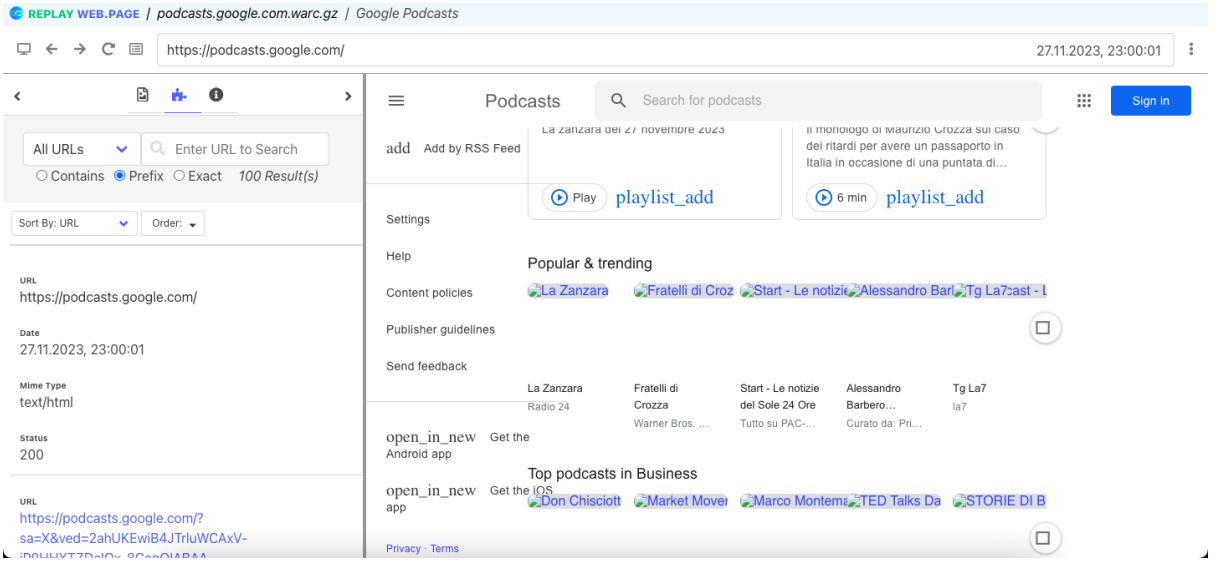


4. данные Apple подкастов с сайта Chartable успешно архивировались! сохранили названия, место в топе и динамику позиций. Снова, рекурсия ушла в служебные страницы и перейти к описаниям подкастов не удается

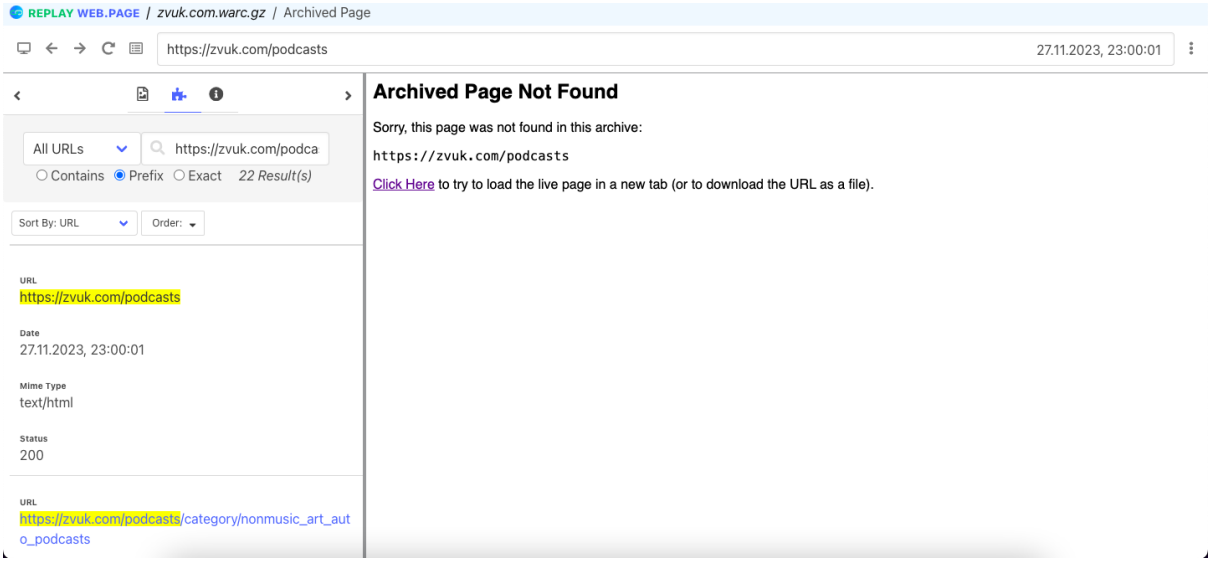


5. данные Google-подкастов архивировались менее удачно: тк регион/язык стартовой страницы определяется автоматически через js, выяснилось что в

архивированной версии произвольно сохранилась итальянская версия сайта



6. Чарт Zvuk.com не удалось архивировать (replay-web.page отображает множество страниц но при переходе на любую из них возникает ошибка как будто страницы не существует)



7. Итоги работы

При проведении работы выяснилось несколько технический особенностей, которые могут вызвать трудности при создании архива:

- Сайты отображение которых регулируется javascript а не поддоменами архивируются с большим трудом - возможно для таких сайтов wpull стоит дополнять другими инструментами или вовсе использовать другой инструмент
- Часто архивация занимает продолжительное время и ресурсы – под создание архива следует выделять как место на диске так и продолжительное время на выполнение

- С крайне высокой вероятностью рекуррентный обход сайта начнет посещать множество служебных страниц прежде чем приступит к основному содержанию страницы. Возможно стоит уточнять запрос wpull либо использовать дополнительные ресурсы для уточнения поискового запроса для подобных случаев

8. Вывод

В этой работе мы создавали коллекцию популярных русскоязычных подкастов – набор ресурсов в открытом доступе, агрегирующих информацию о топ-популярных подкастов в данный момент. Мы установили, что подобная коллекция могла бы послужить ценным источником знаний для специалистов по social data science и digital humanities в академии и индустрии.

В ходе работы оказалось что создание коллекции из нескольких сложных и многоуровневых ресурсов сопряжена с рядом технических особенностей и трудностей. В конечном счете в архиве нам удалось собрать удовлетворительные данные с двух источников (яндекс музыка и apple подкасты), в остальных случаях данные оказались неполными или процесс архивирования сопровождался техническими сбоями.

Это может быть связано как с технической сложностью выбранных ресурсов, так и с несовместимостью wpull с отдельными объектами архивации.