Datasheet

Мотивация

Наш проект в рамках курса “Сбор данных с Web-Scraping и API для социально-научных исследований” направлен на создание набора данных, который фокусируется на анализе российских музыкальных топов до и после введения специальной военной операции. Этот набор данных создаётся для изучения влияния социальных и политических изменений на музыкальные предпочтения и тенденции, отражая, как музыкальная индустрия адаптируется и реагирует на кризисные ситуации.

Цель проекта - идентифицировать изменения в музыкальных вкусах и предпочтениях, выявляя новые тенденции и паттерны, которые могли возникнуть в результате социальных и политических событий. Проект и его дальнейшее развитие может быть направлен на заполнение конкретного пробела в исследованиях, поскольку аналогичные данные не анализировались в данном контексте.

Композиция данных

Экземпляры в наборе данных: Наш набор данных включает музыкальные треки, их исполнителей, альбомы, длительность треков, год вхождения в топы, информацию о наличии ненормативного контента, ссылки на страницы с треками, тексты треков, дату выхода треков и ссылки на страницы с текстами треков.

Типы и содержание экземпляров: Набор данных содержит несколько типов экземпляров.
Числовые значения (numeric): Длительность треков, год вхождения в топы, номер в рейтинге, также в эту категорию можно включить булевые переменные: наличие ненормативного контента и принадлежность трека к альбому (1 - если трек из альбома, 0 - если это single).
Строковые значения (character): название трека, исполнитель, альбом (до обработки).
Ссылки: ссылки на страницы с треками из apple music (метаданные) и ссылки на страницы с текстами песен из genius.
Значение datetime: дата релиза трека.
Неструктурированные текстовые данные: тексты выбранных песен.

Объем набора данных: наш набор данных представляет из себя выборку треков за 2020-2023 год с платформы Apple Music, который включает в себя топ-100 по популярности в России треков за каждый год. Далее мы дополняем наши данные

дополнительными признаками, собираемыми с сайта Genius. Всего было собрано 400 экземпляров соответственно.

Репрезентативность: мы берем данные за 2 года до специальной военной операции и за 2 года после, таким образом, получается сформировать пропорциональную выборку. Данные за четыре года позволяют, насколько это возможно, оценить изменения в музыкальных предпочтениях до и после значимых события. Топ-100 чарты обычно включают треки различных жанров, что способствует широкому охвату музыкальных вкусов в регионе, а также отражают музыкальные предпочтения большого числа слушателей, что делает вашу выборку достаточно значимой для понимания общих тенденций.

Ошибки и пропуски: после формирования первого датасета, собранного с Apple Music, было выявлено 2 значения None в колонке “Длительность”, однако, это связано с тем, что на странице отсутствовала соответствующая информация. После дополнения датасета с сайта Genius пропуски возникли в наблюдениях по конкретным трекам, которых физически нет на сайте. В общем, не предполагается, что наличие небольшого количества пропусков значительно повлияет на результаты исследования, которое можно провести на наборе данных.

Зависимость от внешних ресурсов: набор данных зависит от данных с сайтов Apple Music и Genius, что может представлять риски в плане долгосрочной доступности и изменчивости данных, таких как, изменение расположения кнопок, плейлистов, элементов и разметки сайтов в целом.

Процесс сбора данных

Получение данных: Данные были получены напрямую с веб-сайтов, таких как Apple Music и Genius. Информация включает необработанные тексты песен, информацию о треках и метаданные в виде скрытых с основных страниц ссылок.

Механизмы сбора данных: Для сбора данных мы использовали программное обеспечение Selenium, который позволяет взаимодействовать с веб-страницами, как если бы это делает человек. Selenium использовался для автоматического перехода по плейлистам, сбора информации о треках на Apple Music и последующего извлечения текстов песен и прочей информации с Genius.

Проверка процедур сбора данных: Механизмы и процедуры сбора данных были протестированы сначала на отдельных страницах, что позволило убедится в корректности

и полноте собранной информации, а потом были заключены в функции для автоматической обработки.

Стратегия выборки: Набор данных представляет собой выборку топ-100 музыкальных треков за определенные годы (2020-2024), что позволяет анализировать тенденции на российской музыкальной сцене до и после определенных событий. Выборка не является случайной, она специфически направлена на анализ изменений в музыкальных предпочтениях в период социально-политического кризиса.

Период сбора данных: Данные были собраны за конкретные временные периоды — до и после СВО (2020-2021 и 2022-2023 годы), что позволяет провести сравнительный анализ.

Этические проверки: Проект в общем реализован с учетом этических норм и стандартов, так как слабо взаимодействует с этически конфронтационными материалами. Ввиду того, что данные являются общедоступными и не включают конфиденциальную информацию, связанную непосредственно с пользователями сети Интернет, специфические этические проверки, такие как рассмотрение институциональным наблюдательным советом, не требовались.

Анализ воздействия: Ввиду публичного характера собираемых данных анализ возможного потенциального воздействия на субъекты данных был сосредоточен на обеспечении корректного использования и интерпретации собранных данных, а также на соблюдении правил интеллектуальной собственности.

Предварительная обработка/очистка/маркировка

На предварительном этапе (нулевом) мы обработали тексты песен, собранные в список lyrics. В цикле для обработки всех строк мы использовали функцию re.sub() из модуля re для замены всех вхождений заданных символов (в данном случае это пробелы \u2005 и \u205f, а также символ новой строки \n) в строке lyrics[i] на пробел. Это позволило удалить эти символы или заменить их на пробелы. Далее мы использовали регулярное выражение r'\[.*?\]', чтобы найти и удалить все подстроки, которые начинаются с символа [и содержат любые символы (но минимальное количество символов), и заканчиваются на]. Это используется для удаления текста в квадратных скобках, который может содержать метаданные или комментарии в тексте песни. Затем метод .strip() удаляет любые начальные и конечные пробелы из строки. Далее мы использовали функцию datemaker для преобразования даты в объект datetime. Эта функция принимает на вход список дат в формате "Мес. день год" и превращает его в список с объектами формата datetime.

На первом этапе мы выявляем пропущенные данные по значениям 'No Data' и применяем функцию resolve_mistakes, которая принимает на вход список названий песен, список имен исполнителей, большой датасет и датасет с данными, в которых есть пропуски. При помощи Selenium функция отправляет повторный поисковой запрос и пробует забрать первую ссылку в категории Songs, затем открывает эту ссылку и забирает текст песни, заполняя пропуски в датасете. Функция возвращает pandas.DataFrame с заполненными пропусками. Таким образом, удается обработать некоторые пропуски.

На втором этапе мы форматируем данные. Для столбца "Альбом" создадим дамми-переменную (1-0), которая будет указывать, входит ли трек в альбом или является синглом. Значения в столбце "Длительность" представим в формате int, переведя запись формата {мм:сс} в секунды с помощью функции timemaker, которая принимает на вход список из строк, обозначающих длительность трека в формате "мм:сс", и превращает его в список с числовыми значениями, обозначающими длительность трека в секундах. Далее мы меняем столбцы местами в нашем merged_df таким образом, чтобы это выглядело более логично.

На этапе обработки текстовых данных, первоначально, собранные тексты песен, подверглись токенизации и очистке от специальных символов, знаков препинания,неинформативных слов (стоп-слов), а также от местоимений и слов, длина которых короче 3 букв. Эти действия были реализованы с помощью библиотеки Natural Language Toolkit, функции word_tokeniz, модуля MorphAnalyzer и списков стоп-слов stopwords, а также с помощью функции clean_text, которая возвращает очищенный, токенизированный и лемматизированный текст и функции remove_short_words.

Использование

Наш набор данных, собранный с помощью web-scraping с сайтов Apple Music и Genius, предназначен для анализа тенденций на российской музыкальной сцене в контексте социально-политических изменений, может служить разнообразным целям в различных областях исследований и смежных областях:

Текущее использование: Набор данных использовался для анализа изменений в музыкальных предпочтениях российских слушателей в условиях до и после социально-политического кризиса.

Потенциальное использование:
-Исследование влияния социально-политических событий на культурные тренды.
-Анализ изменений в языке и темах песен в контексте социальных изменений.

- Использование в образовательных целях для изучения музыкальной индустрии, анализа данных и культурологии.
- Разработка алгоритмов машинного обучения для предсказания трендов в музыке.
- Исследование музыкальных предпочтений российских слушателей (например, дополнение собранного датасета опросами россиян, интервью).
- Проект “Формула хита”, направленный на выявление особенностей треков, попадающих в топы.

Ограничения и предупреждения:

- Набор данных не должен использоваться для создания стереотипов или предвзятых выводов о культуре или предпочтениях определенных социальных групп.
- При использовании данных для машинного обучения необходимо учитывать, что предварительная обработка и маркировка могут существенно влиять на результаты анализа.
- Необходимо соблюдать авторские права и этические нормы при использовании текстов песен.

Нежелательное использование:

- Набор данных не следует использовать для целей, которые могут вести к дискриминации, нарушению чьих-либо прав или ущербу репутации исполнителей или, как вариант, каких-то жанров.
- Использование данных для коммерческих целей без учета авторских прав может привести к юридическим последствиям.

Распространение

Распространение данных: Набор данных может быть предоставлен третьим сторонам для исследовательских целей или для продвижения знаний в области музыкальной индустрии и анализа данных. Это распространение должно будет производиться с учетом всех необходимых юридических и этических соображений.

Способы распространения: например, публикация на платформах для совместной работы, таких как GitHub.

Поддержка

Поддержка набора данных не планируется.

Этическая заметка

В нашем проекте мы собираем данные, которые являются общедоступными и касаются музыкальной информации, включая информацию о треках и их тексты, с публичных платформ. Несмотря на публичный характер этих данных, мы придерживаемся этических и правовых норм для обеспечения соответствия нашей работы общепринятым стандартам ответственности и уважения.

Во-первых, уделяется внимание соблюдению авторских прав. Тексты песен, которые являются объектами интеллектуальной собственности, используются в нашем исследовании исключительно для аналитических целей. Мы гарантируем, что наш анализ не приведет к несанкционированному распространению или незаконной публикации этих материалов.

Вторым важным аспектом является уважение к труду создателей музыкального контента. Наш подход исключает субъективные толкования или негативные оценки, обеспечивая объективность и нейтральность в анализе и оценке музыкальных произведений.

Мы также стараемся обеспечить прозрачность нашей методологии. Мы стараемся подробно описать используемые методы сбора и обработки данных, чтобы наши выводы были воспроизводимы и понятны для сообщества.

Ответственное использование собранных данных — наш приоритет. Мы строго ограничиваем использование данных рамками и целью нашего исследования и исключаем любую возможность их применения в дискриминационных или других неприемлемых целях.

Наконец, наша работа соответствует требованиям Общего регламента ЕС по защите данных (GDPR) и других соответствующих правовых стандартов, даже несмотря на то, что мы не обрабатываем личные данные. Мы считаем, что все аспекты нашей работы учитывают необходимые нормативные и юридические рамки.