Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное

учреждение высшего образования

«Южно-Уральский государственный университет»

(Национальный исследовательский университет)

Высшая школа электроники и компьютерных наук

Кафедра системного программирования

ОТЧЕТ

по практической работе №2

|  |  |  |  |
| --- | --- | --- | --- |
| Сбор и разметка данных | | | |
|  | | | |
|  |  | Руководитель: | |
|  |  | | \_\_\_\_\_\_\_\_\_\_\_К.Ю. Никольская  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2020 г.  Автор работы  студент группы КЭ-220  \_\_\_\_\_\_\_\_\_\_\_\_А.В.Витомсков  «\_\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2020 г. |

Челябинск 2020

1. Тема моего проекта: «Классификация популярной музыки по жанрам».

В проекте будет реализована классификация аудиофайлов в формате .wav, содержащих треки популярной музыки, по следующим 12 жанрам:

1) фолк-музыка (world),

2) кантри (country)

3) латиноамериканская музыка (latin-american),

4) блюз (blues),

5) ритм-н-блюз (r-n-b),

6) джаз (jazz),

7) шансон (shanson),

8) электронная музыка (electronic),

9) рок (allrock),

10) хип-хоп (hip-hop),

11) регги (reggae),

12) поп-музыка (pop).

Жанры популярной музыки я взял из Википедии: [https://ru.wikipedia.org/wiki/Список\_музыкальных\_жанров,\_направлений\_и\_стилей](https://ru.wikipedia.org/wiki/Список_музыкальных_жанров,_направлений_и_стилей%20)

2. Для сбора данных было решено использовать «Яндекс.Музыку». «Яндекс.Музыка» — стриминговый сервис компании «[Яндекс](https://ru.wikipedia.org/wiki/%D0%AF%D0%BD%D0%B4%D0%B5%D0%BA%D1%81)», позволяющий прослушивать музыкальные композиции, альбомы, подборки музыкальных треков и получать персональные рекомендации. На сервисе доступно более 60 млн треков.

Выбор в пользу «Яндекс Музыки» был сделан по трем причинам.

Во-первых, это наличие адекватного API: существует пайтон-библиотека с подробной документацией, правда неофициальная.

Во-вторых, это понятная классификация треков по жанрам.

В-третьих, серверы Яндекс Музыки находятся на территории России, скорость соединения и скачивания музыки велика.

3. Для начала было решено собрать базу данных об исполнителях и их принадлежности определенному музыкальному жанру. Сначала для этой цели я попробовал использовать ресурс <https://arzamas.academy/materials/1318> . Но на нем содержатся данные всего о 187 исполнителях (рисунок 1), что очень мало для поставленной задачи.

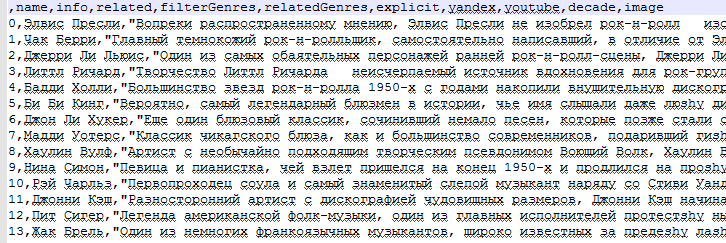


Рисунок 1 – Данные <https://arzamas.academy>

4. Поэтому мой дальнейший выбор пал на ресурс [https://music.yandex.ru/genre/<название\_жанра>/artists](https://music.yandex.ru/genre/%3cназвание_жанра%3e/artists) .

Для начала, используя python модули beatifulSoup, regex и json, я получил информацию о музыкальных жанрах. То, что я получил от Яндекса, представлено на рисунке 2.

Видим, что всего имеется 30 жанров, в каждом из которых имеются поджанры. Я взял отсюда 12 нужных мне жанров популярной музыки. Единственно, latin-american музыка оказалась поджанром world, я учел это в дальнейшем.

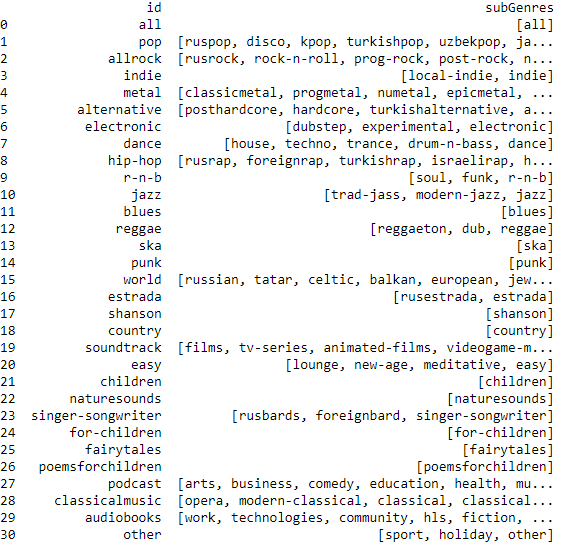


Рисунок 2 – Жанры Яндекс.Музыки

5. Было решено представлять каждый жанр конкретными исполнителями. Понятно, что один и тот же исполнитель может по классификации попадать в несколько жанров. В этом случае я его не исключаю, просто считаю, что все его произведения находятся «на стыке жанров» и могут рассматриваться как принадлежащие сразу нескольким жанрам.

Используя тот же парсинг bs4 и json, для каждого из 12 жанров и внутри каждого жанра для каждого поджанра я получил список исполнителей.

Фрагмент таблицы, полученной после преобразований, приведен на рисунке 3.



Рисунок 3

В полученном файле 52803 записи.

Как распределено число исполнителей по жанрам на Яндекс Музыке, представлено на рисунке 4.



Рисунок 4 – Число исполнителей в каждом жанре

6. Таблицу исполнителей я преобразовал к более удобному виду, где каждый столбец представляет определенный жанр, а в строках приведены имена исполнителей. На данном этапе я перемешал исполнителей рандомно в рамках каждого столбца.

Первые 50 записей полученной таблицы приведены на рисунке 5. В таблице 2811 строк.



Рисунок 5 – Имена исполнителей по жанрам

7. С помощью модуля yandex\_music для каждого из 12 жанров я выполнил поиск по 300 исполнителям (исполнителей выбирал рандомно), и у каждого исполнителя загрузил 30 наиболее популярных треков.

Именно 300 исполнителей выбрал, так как минимальное число 438 (для шансона) и нужно учесть запас на случай исполнителей, у которых меньше 30 треков – такие исполнители игнорировались и заменялись другими.

30 треков выбрано по соображениям необходимого объема датасета.

8. API Яндекс.Музыки позволяет скачать файл в формате .mp3 или .aac битрейтом 64, 128, 192, 320 kbps. Опытным путем установлено, что .mp3 192 является самым популярным форматом, скачиваются где то 9 из 10 файлов. Поэтому использую для скачивания этот формат. Если произошла ошибка, скачивается другой файл, чтобы общее количество файлов составило 30 на исполнителя.

9. Для каждого файла выполняется предобработка. Используется модуль pydub (конвертер ffmpeg) для конвертации mp3 в wav. При этом применятся следующие параметры:

число каналов = 1 (моно),

число байтов в каждой выборке = 1 (2^8 = 256 уровней амплитуды),

частота дискретизации = 22050 Гц (позволяет сохранить оригинальное звучание в диапазоне частот до 11 кГц, что достаточно для голоса и большинства муз. инструментов).

Затем из полученного wav файла с помощью модуля Librosa вырезаются участки тишины. За тишину принял фрагменты, уровень амплитуды которых на 50 дБ меньше максимального в треке.

Из полученного трека в рандомном месте берется участок 10 секунд. Треки длиной менее 10 секунд игнорируются и меняются на другие.

Полученный трек нормализуется с помощью модуля librosa: максимальная громкость приводится к единице.

10. Полученный размер файла с 10-секундной записью составил 431 кБ. Итоговый размер датасета:

431 \* 30 \* 300 \* 12 / 1000 / 1000 ~ 47 Гб.

11. Ввиду большого размера, полученные файлы для каждого исполнителя упакованы в zip архив.

Фрагмент полученного датасета представлен на рисунке 6.

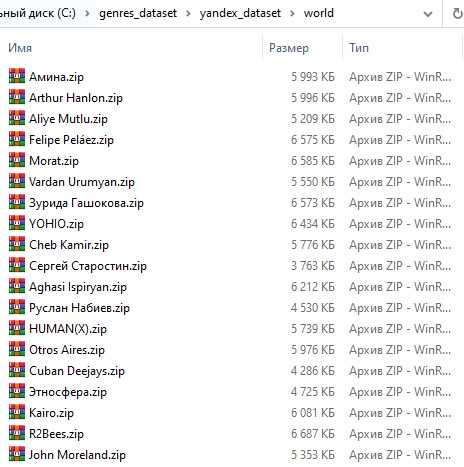


Рисунок 6

Содержимое одного из архивов на рисунке 7.

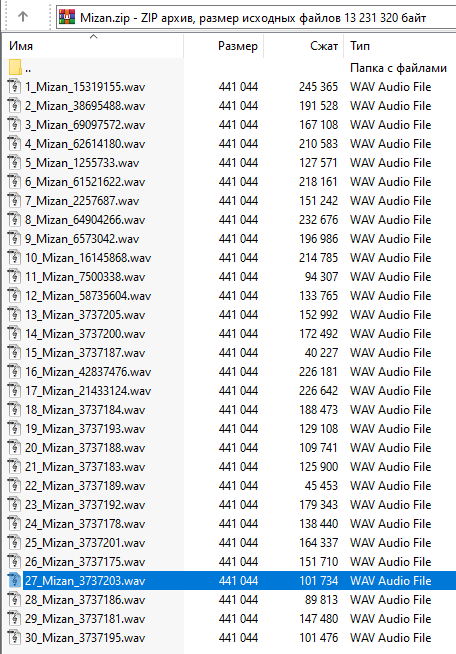


Рисунок 7