

Дипломный проект

по профессии Data Scientist

Тема: Рекомендательные системы по подбору музыкальных треков с использованием методов ML

Сопроводительное письмо

Студентка: Шурпаева Р.

Ментор: Баданина Н.

2022

г. Москва

Мной выбрана тема дипломной работы «Рекомендательные системы по подбору музыки», так как за время обучения блок рекомендательных систем был одним из более интересных. Почему в музыке? С ней сложнее работать.

*Сервисы анализируют каждый загруженный в систему аудиофайл. Трек нарезается на фрагменты — по ним делаются спектрограммы, понятные машинам. Так нейросети обучают на «внутренностях» популярных треков, по которым много отзывов. После этого алгоритм подбирает что-то похожее по звучанию из тех холодных треков, которые мало кто слышал.*

*Но разложить песню на ДНК у машины до конца пока не получается — так же, как из мельчайших частей написать что-то новое.*

*По словам руководителя лаборатории машинного обучения «Яндекса» Александра Крайнова, генерировать текст для машины гораздо легче, чем музыку: «В тексте есть символы, есть некоторая закономерность. Символы — это уже закодированная, сжатая информация. С изображением тоже просто, просто потому что там объем данных меньше. А звук — он большой, сложный, у него очень длинные и большие временные зависимости. В звуке очень много чего может объединяться: кто-то петь начинает на заднем плане, то есть это не просто звук голоса, а слова» \**

Все люди, вне зависимости от того, какую музыку они предпочитают, сталкиваются с одной и той же проблемой: личная фонотека заслушана и хочется чего-нибудь новенького. С одной стороны, выбор огромен — мир музыки очень богат, он насчитывает сотни миллионов треков и пополняется каждый день. С другой стороны, сориентироваться в этом разнообразии бывает непросто: музыки много, а человек один. Собственно, целью моей работы является решение этой проблемы.

Исходными являются данные мирового стримингового сервиса Spotify. Ссылка на датасет: <https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db>

Этапы работы:

1. Загрузка данных, первичная обработка, визуализация;
2. Нормализация данных с использованием StandartScaler, векторизация жанров с использованием TF-IDF, рассчитана метрика RSME;
3. Нормализация данных с помощью MinMaxScaler, кластеризация KMeans.
4. Построена первая рекомендация музыки, на основе численных характеристик в датасете. Указана дистанция между треками.
5. Обнаружение и исправление проблемы дубликатов. (Один и тот же трек относится к разным жанрам, т.е. отображается в несколько строк).
6. Построена вторая программа для рекомендации, с исключением жанров музыки из датасета и дубликатов песен. Указана дистанция между треками.

Вторая часть работы:

1. В качестве примера – добавлены к датасету пользователи и рейтинг.
2. Преобразованы данные о модальности и жанрах с помощью LabelEncoder.
3. Построена рекомендация для пользователя.

# Заключение

В работе была разработана модель рекомендательной системы для музыки на стриминговом сервисе. В ходе работы проанализирован исходный датасет, данные проверены на наличие пустых значений, выявлены дубликаты.

Для построения рекомендательной системы выбран Item-based подход, так как датасет имеет числовые характеристики для каждого трека (акустичность, темп, модальность и т.д.). Эти данные использованы, чтобы найти наиболее похожие песни с расчетом дистанции между ними.

Чтобы найти разницу между песнями, рассчитано расстояние по Манхэттену между всеми из них. И, как результат, в рекомендации попадают песни с наименьшими расстояниями.

Также в работе применена векторизация жанров, рассчитан RMSE ~ 0,62

Дополнительно добавлены пользователи и рейтинг, построена рекомендация для отдельного пользователя в рамках одного жанра и в 5 ближайших.

Код проекта доступен по ссылке

<https://github.com/shurpaeva/Recsis_Diplom_DS/blob/main/Рекомендательные_системы_по_подбору_музыкальных_треков_с_использованием_методов_ML.ipynb>

Список литературы:

* + 1. \*<https://trends.rbc.ru/trends/industry/5f2c271c9a7947ead6c655f3>
    2. <https://www.programmersought.com/article/3755788463/>
    3. <https://habr.com/ru/post/486802/>
    4. <https://habr.com/ru/post/666234/>
    5. <https://www.codespeedy.com/music-recommendation-system-project-using-python/>
    6. <https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features>
    7. <https://machinelearningmastery.ru/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0/>
    8. <https://habr.com/ru/post/101338/>
    9. <https://apptractor.ru/info/articles/kak-mashinnoe-obuchenie-v-spotify-nahodit-vashu-novuyu-lyubimuyu-muzyiku.html>