Предварительный поиск информации показал, что на текущей момент не существует одного источника данных, который позволяет решить все задачи проекта. Поэтому решено воспользоваться информацией из двух ресурсов:
- сайт **last.fm** (этап 1)
- база данных **FMA** (этап 2)

Этап 1

Цели:
- получить оценку популярности/ меру предпочтения музыкальных треков с детализацией по каждому слушателю
- собрать минимально достаточную выборку предпочтения слушателей

Сайт www.last.fm — это сайт, посвящённый музыке. С помощью плагинов к медиаплеерам он собирает информацию о музыке, которую слушают пользователи, и на основе полученных данных составляет индивидуальные и общие хит-парады. Содержит базу данных об исполнителях музыки, альбомах и композициях.

По итогам парсинга **last.fm**, при помощи API получены данные о предпочтениях 10 тысяч пользователей ресурса и подготовлено два датасета.

В датасете **«user_info.csv»** содержится общая информация по каждому клиенту:

- **Login** - уникальный идентификатор на сайте *[тип данных: категориальный]*
- **Name** - имя пользователя (как правило не совпадает с **Login**) *[тип данных: категориальный]*
- **Country** - страна происхождения клиента *[тип данных: категориальный]*
- **Age** - возраст (много пропусков) *[тип данных: int64]*
- **Gender** - пол (много пропусков, но можно попробовать проставить пол вручную, исходя из имени) *[тип данных: категориальный]*
- **Registered** - дата регистрация в сервисе last.fm *[тип данных: категориальный]*

В датасете «**taste.csv**» содержится информация о личном хит-параде (первые **100** наиболее прослушиваемых треков) каждого из 10 тысяч клиентов:

- **Name** - логин пользователя из датасета **«user_info»** *[тип данных: категориальный]*
- **Rank** - ранг популярности трека у конкретного пользователя (чем выше ранг, тем чаще трек прослушивался этим клиентом) *[тип данных: категориальный]*
- **Mbid** - уникальный номер трека от musicbrainz *[тип данных: категориальный]*
- **Artist** - имя группы/ певца/ певицы *[тип данных: категориальный]*
- **Song_name** - название трека *[тип данных: категориальный]*
- **Duration** - длительность трека в миллисекундах *[тип данных: int64]*
- **Count** - число прослушиваний трека *[тип данных: int64]*

Итого, за счет комбинирования указанных датасетов получена информация о популярности/ числе прослушивания 300K+ треков от 100K+ артистов. Это дает возможность применить арсенал не только ML, но и подступиться к DL (не исключаем, что к этапу DL будет выгружено больше данных).

Этап 2

Цель: добавить по каждому треку из этапа 1 дополнительные фичи из более крупной музыкальной базы данных (**FMA**), которые, потенциально, могут улучшить качество предсказания

FMA (Free Music Archive) - это открытая БД, которая содержит аудио и текстовую информацию о 106,574 треках от 16,341 артистов и 14,854 альбомах. На текущем этапе проекта мы планируем сосредоточиться на добавлении только текстовых фичей.

По каждому треку из FMA могут быть добавлены следующие фичи:

- **Date_recorded** - время создания трека
- **Favourites** - у какого числа пользователей этот трек находится в категории любимых
- **Genre_top** - основной жанр трека
- **Language_code** - язык трека
- **Listens** - число прослушиваний трека
- **Artist_location** - страна исполнителя

Будут добавлены дополнительные признаки из системы распознавания музыки Echonest (измеряются в интервале от 0 до 1):

- **Acousticness** - мера уверенности в том, что композиция является акустической
- **Danceability** - описывает пригодность трека для танцев на основании темпа, стабильности ритма и других показателей
- **Energy** - характеризует «яркость» и «активность» песни. Обычно энергичные композиции быстрые, громкие и шумные. Например, высокой энергией обладает death metal, а прелюдия Баха имеет по этой шкале низкие показатели.
- **Instrumentalness** - прогноз того, что в треке нет вокала. В этом контексте звуки «оу» и «а-а-а» считаются инструментальными. Рэп или треки со словами являются «вокальными». Чем ближе значение инструментальности к 1.0, тем выше вероятность того, что в треке не содержится голоса
- **Liveness** - распознаёт присутствие в записи слушателей. Чем больше значения liveness, тем выше вероятность того, что песня исполнялась вживую
- **Speechiness** - обнаруживает присутствие текста в песне. Если speechiness композиции выше 0.66, то она скорее всего состоит из текста, значение ниже 0.33 означает, что в песне нет никаких слов
- **Valence** - описывает музыкальную позитивность, передаваемую песней. Песни с высокой valence звучат более позитивно (т.е. они передают счастье, радость или эйфорию), а песни с низкой valence звучат негативнее (т.е. они печальные, депрессивные или гневные)

Возможно, будут подключены дополнительные признаки, созданные для FMA с помощью библиотеки Librosa (перечень конкретных фичей - это предмет дальнейших изысканий).

Код и датасеты

Код для парсинга last.fm при помощи API лежит здесь:
Датасет **«user_info.csv»** лежит здесь:
Датасет **«taste.csv»** лежит здесь: