MUSIC RECSYS HSE PT.1

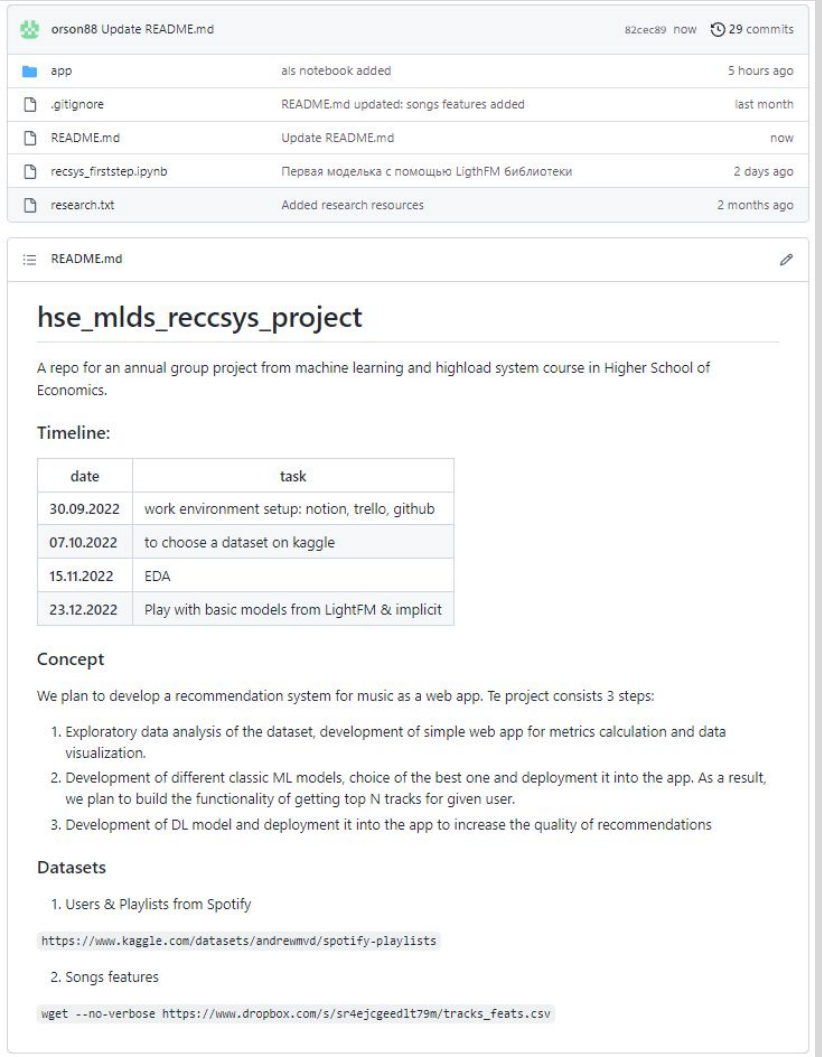
Герасимов Владислав
Юркин Андрей
Поздняков Арсений

Чекпоинты:

1. Организационный
2. Сбор данных
3. EDA
4. Базовые модельки

Организационный

- Собрались, сделали репозиторий
- Выбрали тему – музыкальные рекомендации
- Датасет – Плейлисты юзеров из спотифая и описания музыки, заскрапленные со спотифая
- Конечный формат – веб приложение



orson88 Update README.md — 82cec89 now — 29 commits

app	als notebook added	5 hours ago
.gitignore	README.md updated: songs features added	last month
README.md	Update README.md	now
recsys_firststep.ipynb	Первая моделька с помощью LigthFM библиотеки	2 days ago
research.txt	Added research resources	2 months ago

README.md

hse_mlds_reccsys_project

A repo for an annual group project from machine learning and highload system course in Higher School of Economics.

Timeline:

date	task
30.09.2022	work environment setup: notion, trello, github
07.10.2022	to choose a dataset on kaggle
15.11.2022	EDA
23.12.2022	Play with basic models from LightFM & implicit

Concept

We plan to develop a recommendation system for music as a web app. Te project consists 3 steps:

1. Exploratory data analysis of the dataset, development of simple web app for metrics calculation and data visualization.
2. Development of different classic ML models, choice of the best one and deployment it into the app. As a result, we plan to build the functionality of getting top N tracks for given user.
3. Development of DL model and deployment it into the app to increase the quality of recommendations

Datasets

1. Users & Playlists from Spotify

`https://www.kaggle.com/datasets/andrewmvd/spotify-playlists`

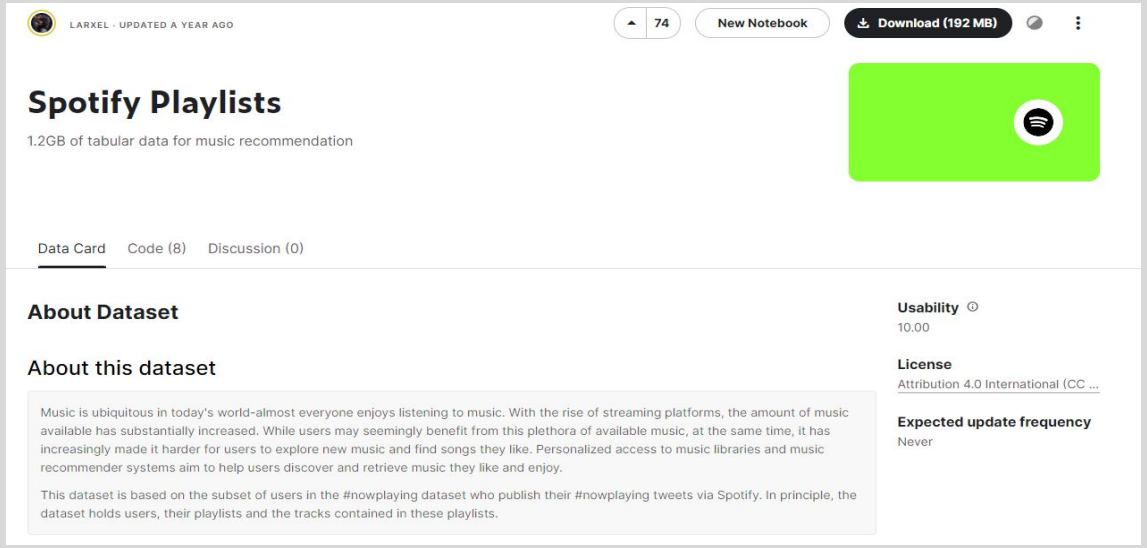
2. Songs features

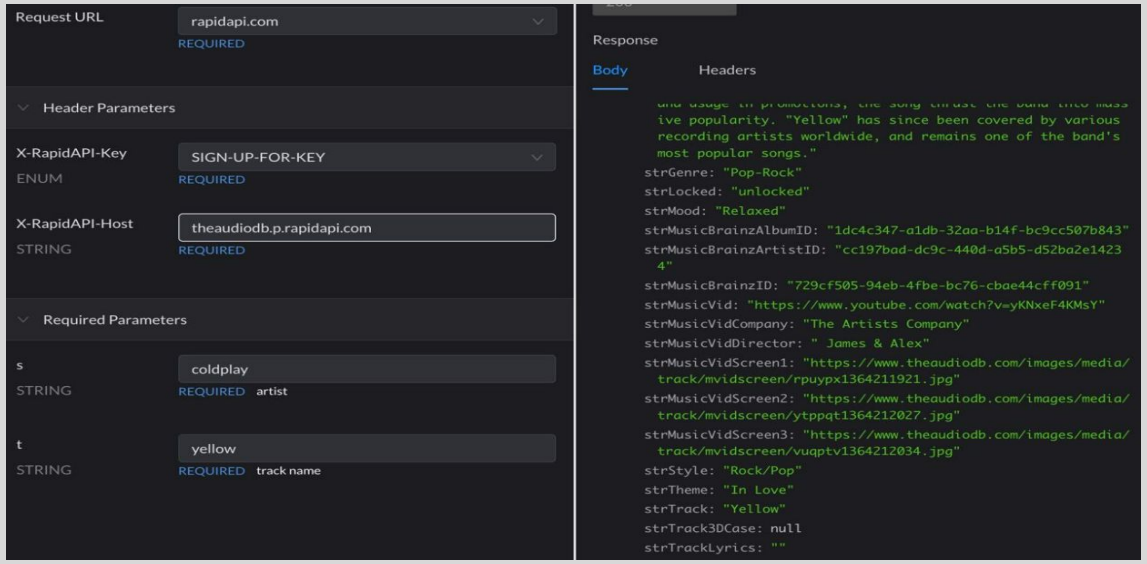
`wget --no-verbose https://www.dropbox.com/s/sr4ejcgeed1t79m/tracks_feats.csv`

https://github.com/orson88/hse_mlds_recsys_project

Сбор данных

- Нашли данные, заскрапили прочие фичи со Спотифая
- Набрали статей по акустическим фичам – данных не нашлось
- Датасет – Плейлисты юзеров из спотифая и описания музыки, заскрапленные со спотифая
- Конечный формат – веб приложение





https://www.kaggle.com/datasets/andrewmvd/spotify-playlists

https://rapidapi.com/theaudiodb/api/theaudiodb/

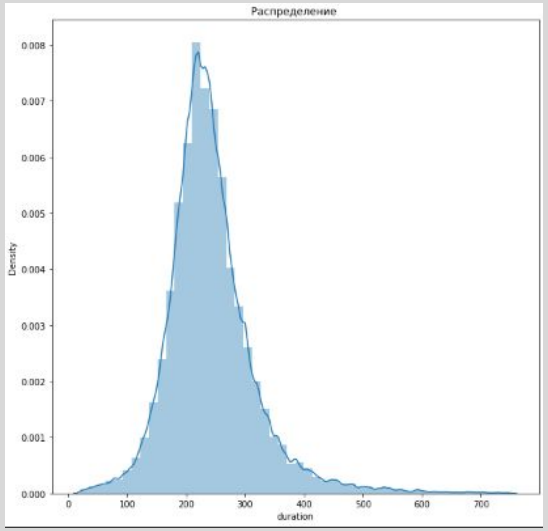
EDA

- Принято решение выбирать топ 10000 песен (артист+песня). Получили общее представление о структуре данных.
- 16 т. юзеров
- 2.8 млн. песен

	uid	artist_x	song_name_x	playlist	num_id	artist_y	song_name_y	album	genre	mood	style	theme	duration
0	9cc0cfd4d7d7885102480dd99e7a90d6	Elvis Costello	(The Angels Wanna Wear My) Red Shoes	HARD ROCK 2010	751205	Elvis Costello	(The Angels Wanna Wear My) Red Shoes	My Aim Is True	Pop-Punk	NaN	NaN	NaN	170.0
1	9cc0cfd4d7d7885102480dd99e7a90d6	Elvis Costello & The Attractions	(What's So Funny 'Bout) Peace, Love And Unders...	HARD ROCK 2010	750788	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	9cc0cfd4d7d7885102480dd99e7a90d6	Tiffany Page	7 Years Too Late	HARD ROCK 2010	2548456	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	9cc0cfd4d7d7885102480dd99e7a90d6	Elvis Costello & The Attractions	Accidents Will Happen	HARD ROCK 2010	750791	Elvis Costello & The Attractions	Accidents Will Happen	Armed Forces	New Age	NaN	NaN	NaN	178.0
4	9cc0cfd4d7d7885102480dd99e7a90d6	Elvis Costello	Alison	HARD ROCK 2010	751243	Elvis Costello	Alison	My Aim Is True	Pop-Punk	NaN	NaN	NaN	205.0

```
Daft Punk                37640
Coldplay                 33399
The Rolling Stones       30929
Queen                    30837
Radiohead                30738
Kanye West               28978
Michael Jackson          28964
Eminem                   28842
JAY Z                    28532
Muse                     28270
David Bowie              27914
Lady Gaga                27712
Bruce Springsteen        25763
Bob Dylan                24677
U2                       24403
Beyoncé                  24248
Arctic Monkeys           23945
Pearl Jam                23844
Rihanna                  23508
Nirvana                  23048
Foo Fighters             22878
Johnny Cash              22571
David Guetta             21798
John Mayer               20690
Red Hot Chili Peppers    20527
Name: artist_x, dtype: int64
```

```
Starred                  1358091
Liked from Radio          185894
Favoritas de la radio      31707
Rock                       30807
Christmas                  22751
2014                       22419
2013                       20425
Work                       18778
Jazz                       18371
Indie                      18145
everything                 16443
Classical                  16394
Country                    15660
New Playlist               15451
Music                      15161
Generos                    15021
Rap                        14506
iPhone                     13483
Chill                      13481
2012                       12968
Metal                      11277
All Live Files             11174
My Shazam Tracks           11097
2015                       10996
Random                     10876
Name: playlist, dtype: int64
```

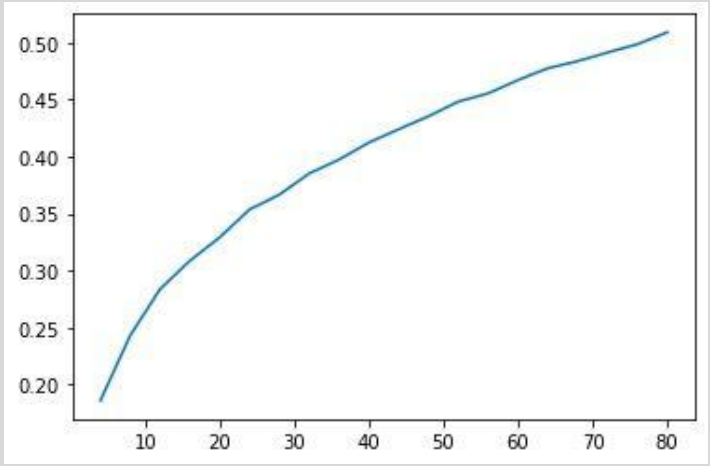


https://colab.research.google.com/drive/1PTA-3IWY0VAryFYCLk2HJx86NWc5ofu1?usp=sharing

Модели

- Практиковались в LightFM и Inmplicit, библиотеках для recsys моделей и нужных матричных преобразований.
- Разбили трейн и тест рандомно, проверив наличие всех тест-юзеров в трейне
- Выбрали Precision@K

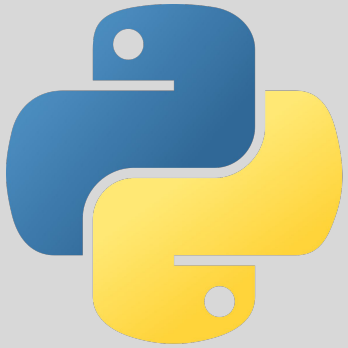
Precision@K = 0.55

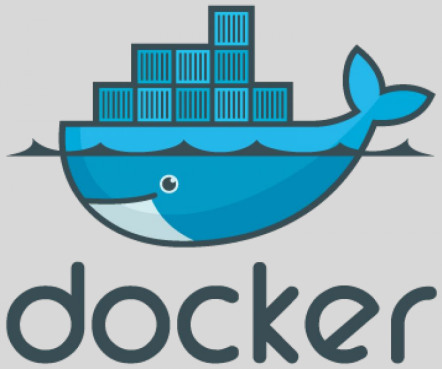


https://colab.research.google.com/drive/1PTA-3IWY0VAryFYCLk2HJx86NWc5ofu1?usp=sharing

Планы

- Сервис на FastAPI
- Использовать доп. Фичи
- Другие модели





Спасибо за внимание!