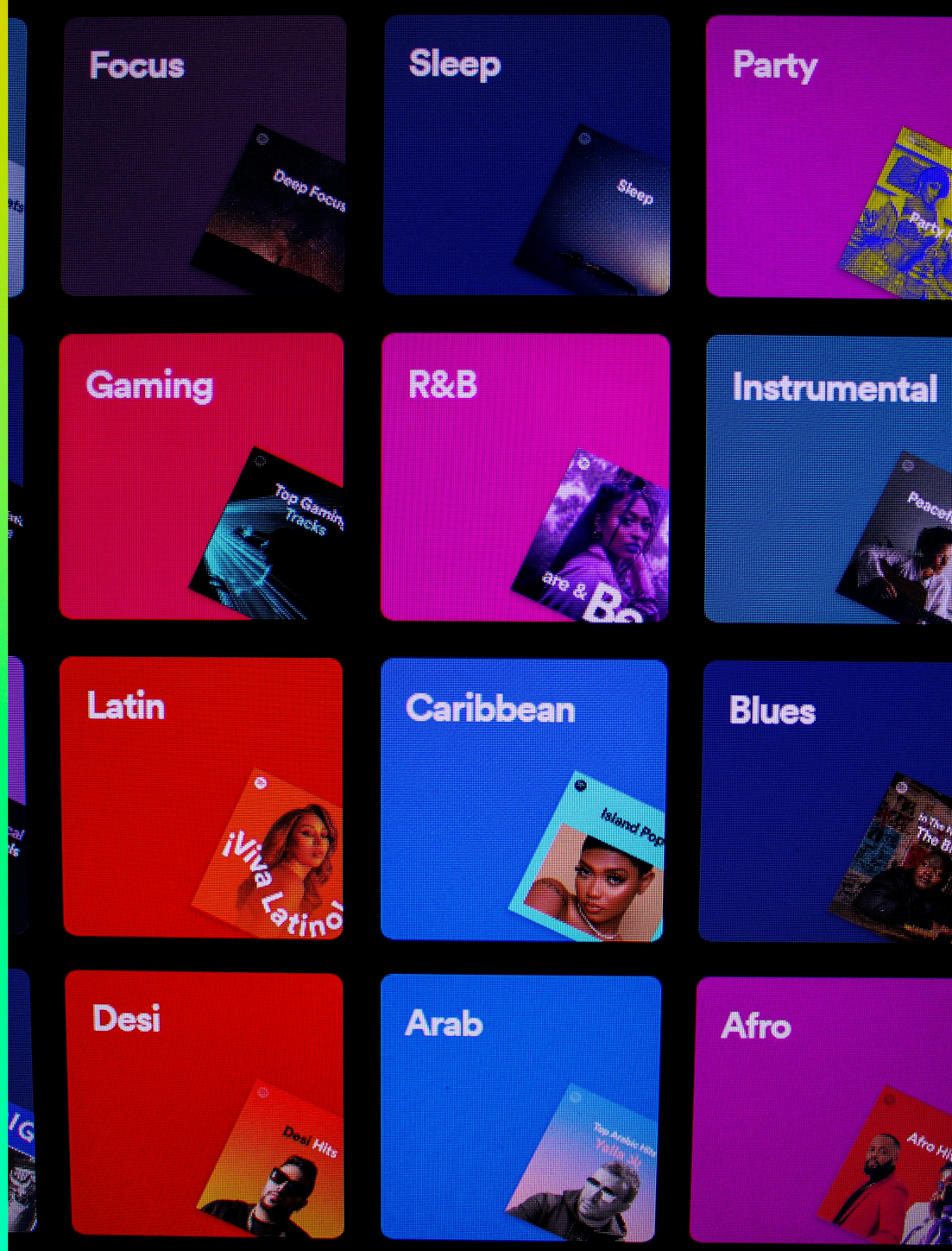
MODELING SONG SIMILARITY USING LYRICS

DUYGU GÖKSU

MARCH 8, 2024

CAPSTONE PROJECT, SPRINT 1



SPOTIFY PLAYLISTS

MADE FOR EVERYONE

Featured Charts:

- Top Songs Global/USA
- Top 50 Global/USA
- Viral 50 Global/USA

Try Something Else:

- Viva Latino
- New Music Friday
- La Lista Pop
- Rap Caviar
- Teen Party
- Baila Reggaeton
- Hot Country
- Rock This
- Songs to Sing in the Shower

DUYGU GÖKSU | MARCH 8, 2024

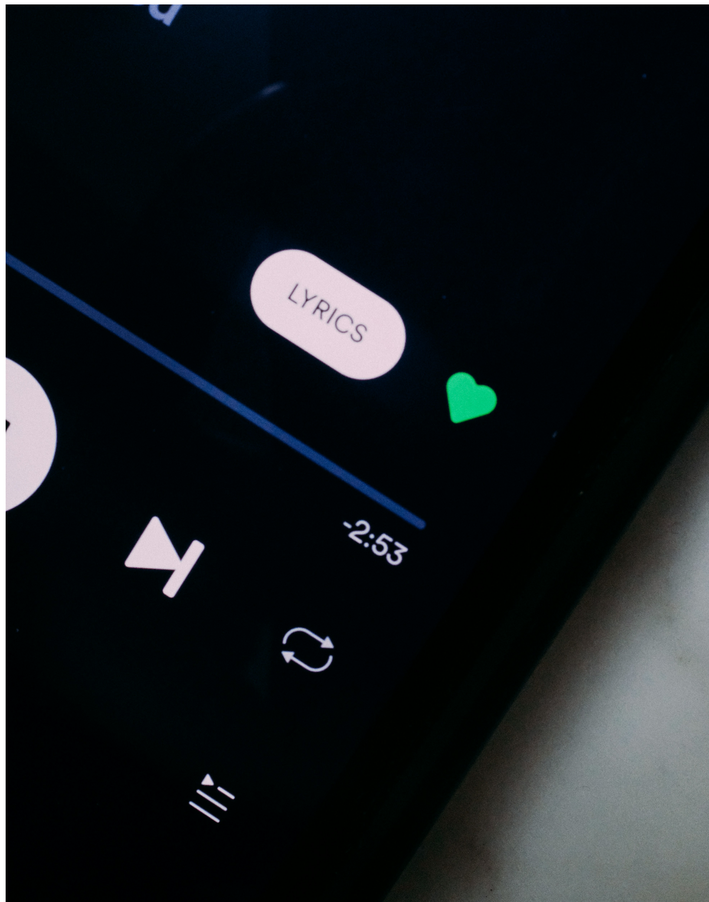
MODELING SONG SIMILARITY FROM LYRICS TO CREATE NEW PLAYLISTS

HOW?

- Find a large dataset of songs with lyrics

- Perform some preprocessing magic

- Use gensim library for topic modeling and word embedding models



DUYGU GÖKSU | MARCH 8, 2024

image source: unsplash.com

POTENTIAL IMPACT

236 MILLION

PREMIUM SUBSCRIBERS*

4.27€

AVERAGE REVENUE PER USER*

*Source: https://www.businessofapps.com/data/spotify-statistics/

DATA

CLEANING AND MERGING

3 datasets from Kaggle:

i. ‘30000 Spotify Songs’ 5 nulls, duplicates ~4500 (track_id) + 5 (song, artist, album)

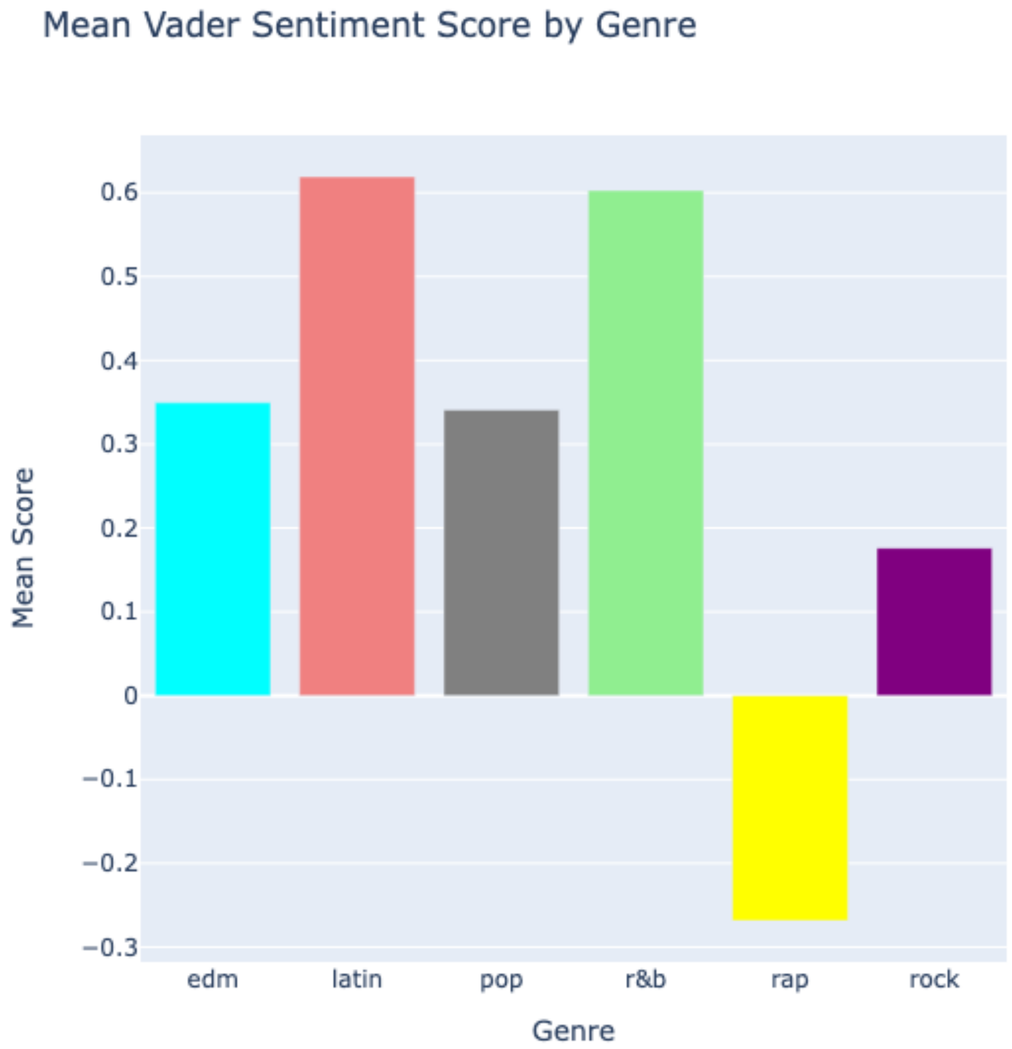
ii. ‘150K Lyrics Labeled with Spotify Valence’ ~1850 lyrics match (artist, song)

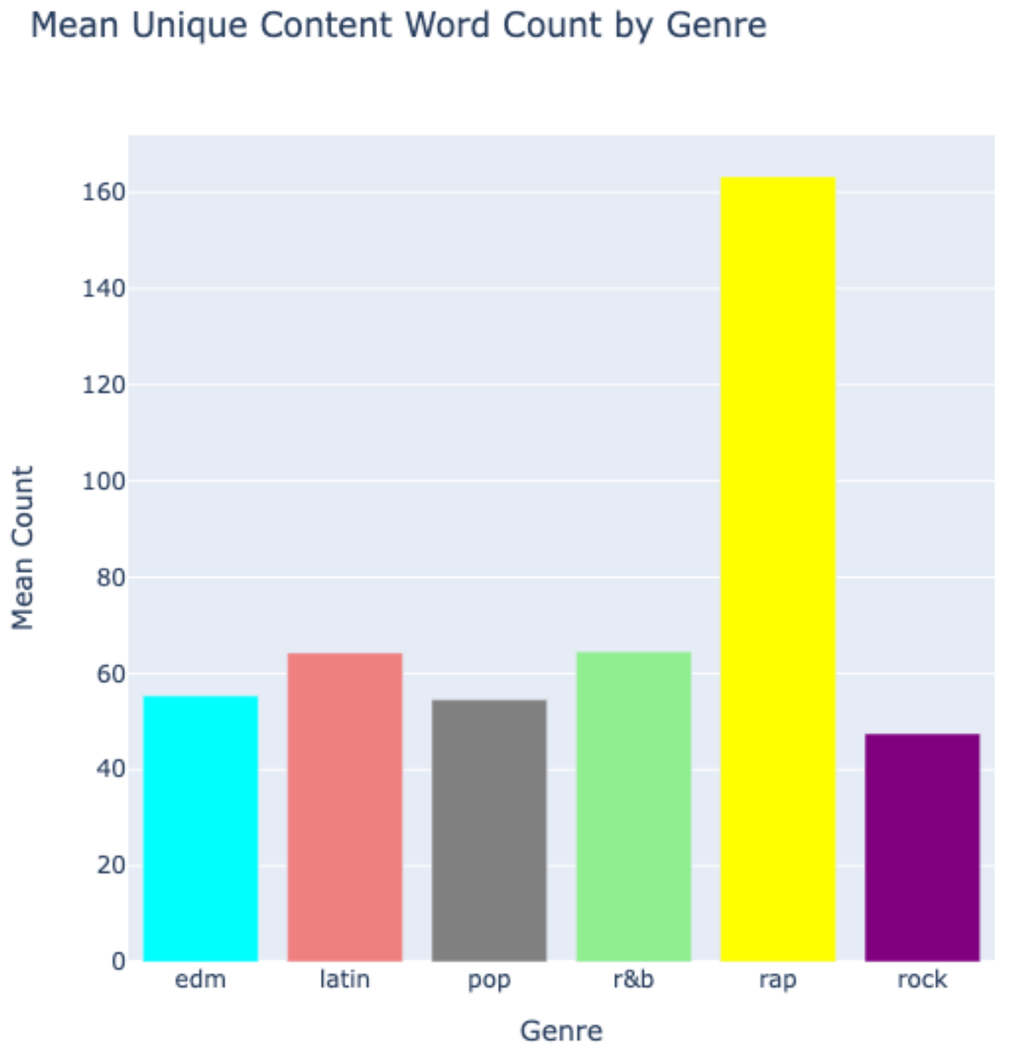
iii. ‘Spotify Lyrics Dataset’ ~100 lyrics match (track_id)

Result:

28347 rows, 24 columns, with 1926 non-null lyrics

PRELIMINARY EDA FINDINGS





NEXT STEPS

- Add more lyrics to the dataset

- More data cleaning

- Learn more about gensim library

- Data preprocessing

- Initial modeling

TEŞEKKÜRLER!