RECONSTRUCTING THE HISTORY OF MUSIC RECOMMENDATION

PROJECT GOAL

1

Rebuilding 3 major recommendation paradigms

DATASET & SETUP

2

Spotify 30K + preprocessing & tools

CONTENT-BASED FILTERING

3

Feature similarity (cosine + PCA)

COLLABORATIVE FILTERING

4

User–item matrix + SVD (latent factors)

NEURAL RECOMMENDATION

5

Item2Vec + playlist context + t-SNE

INTRODUCTION & MOTIVATION

Introduction

The explosion of digital music platforms created an overwhelming amount of content.

Users rely on recommendation systems to navigate millions of tracks.

Understanding *how* these systems evolved helps us design better, more transparent models.

Motivation

Shift: rules → statistics → neural models

Need to understand their logic & limits

Build transparent, comparable systems

Objectives

Reconstruct the three recommendation paradigms.

Analyze their behavior under the same dataset.

Identify what each method captures uniquely.

Contribution – Andra Mihaela Andruță

BUILD

Dataset preprocessing & cleaning

EXPLORE

Exploratory Data Analysis (EDA)

MODEL

Content-Based Filtering model implementation

ANALYZE

PCA visualizations & Word Cloud analysis

EMBED

Training the Word2Vec embedding model

INTERPRET

Semantic playlist transition analysis (t-SNE interpretation + transition paths)

Contribution – Ioana Alexandra Tunaru

1. **GENERATE** – Simulation of user–song interactions + generation of playlist corpora
2. **MODEL** – Implementation of Collaborative Filtering (SVD) and Item2Vec neural embeddings
3. **CONNECT** – Co-listening & playlist-context logic (user–item co-occurrence + Item2Vec neighbors)
4. **PREDICT** – Personalized top-N recommendations using latent factors
5. **EVALUATE** – Popularity vs Personalization analysis (Venn diagram study)
6. **INTERPRET** – Comparative study: Content-Based Filtering vs Collaborative Filtering

DATASET OVERVIEW

Dataset Summary

- Dataset extracted from Kaggle's Spotify Tracks Dataset.
- Contains 29,865 unique songs after cleaning.
- Each entry includes track name, artist, and playlist-based genre.
- Includes 7 audio features describing musical characteristics.

Data Cleaning

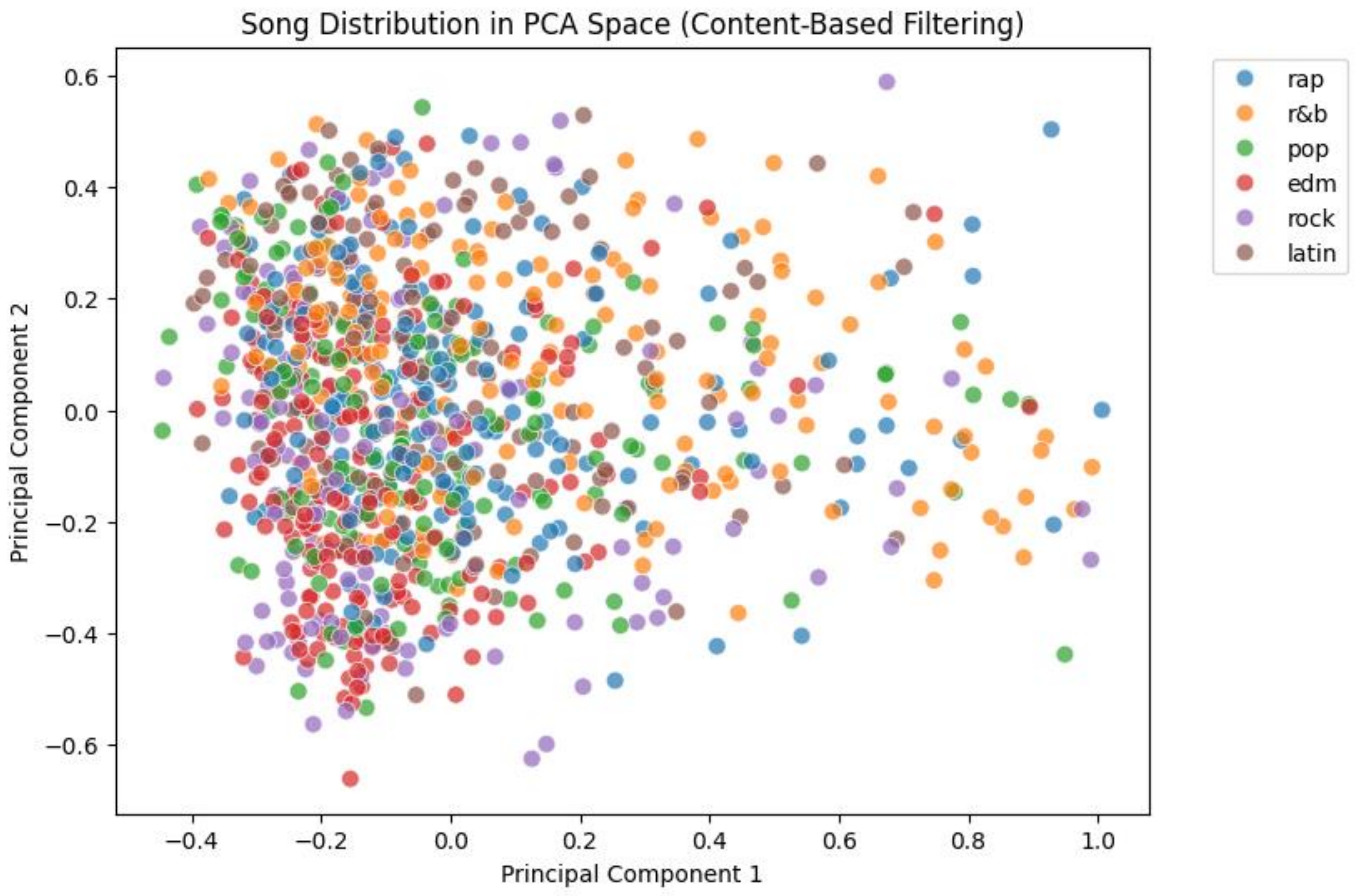
- Removed missing values and duplicate tracks.
- Kept 10 essential columns used in our experiments.
- Standardized genre labels for consistency.
- Verified feature distributions before modeling.
- Final dataset prepared for all three recommendation paradigms.

Feature Normalization

- Most Spotify audio features are already scaled 0–1.
- All 7 columns normalized using Min–Max scaling.
- Ensures fair comparison between all features.
- Normalized matrix used in similarity scores, SVD, and Word2Vec.
- Ready for content-based, collaborative, and embedding models.

CONTENT-BASED FILTERING (CBF)

- Uses **audio features** (danceability, energy, valence, tempo, acousticness, liveness, speechiness).
- Computes similarity using **cosine distance**.
- Recommends songs with **similar sound** characteristics.
- Does not require user interaction history.



This PCA plot shows how **songs cluster based on audio features**. Tracks with similar sound profiles appear closer together, supporting the idea behind content-based filtering.

COLLABORATIVE FILTERING (SVD)

- Learns from **collective user behavior** instead of audio features.
- Uses User–Item interactions to detect similar users or co-listened songs.
- Applies Matrix Factorization (SVD) to learn **latent taste factors**.
- Reconstructs missing ratings and predicts user preferences.
- **Co-listening logic**: songs frequently played together are considered similar.

SVD recommends new songs by predicting which tracks best match the user's **hidden taste profile**.

```
=== SMART-SVD Recommendations for user_id = 15 ===

Songs already listened to (with rating):
```

	track_name	track_artist	playlist_genre	rating
2	Only God Can Judge Me	2Pac	rap	2
3	Insta	Bizzey	latin	2
4	Being Alive	Hardwell	edm	4
7	Dance?	Night Tempo	r&b	1
11	Idealne Połączenie	Young Igi	rap	2
...	...	...	...	...
492	In My Mind	The Amazons	rock	5
493	Extraños En El Escaparate - Live	Miguel Rios	rock	4
495	Sparkles	Daxten	latin	2
496	Reflections	Jean Juan	edm	4
497	Cryin' Like A Bitch!!	Godsmack	rock	5

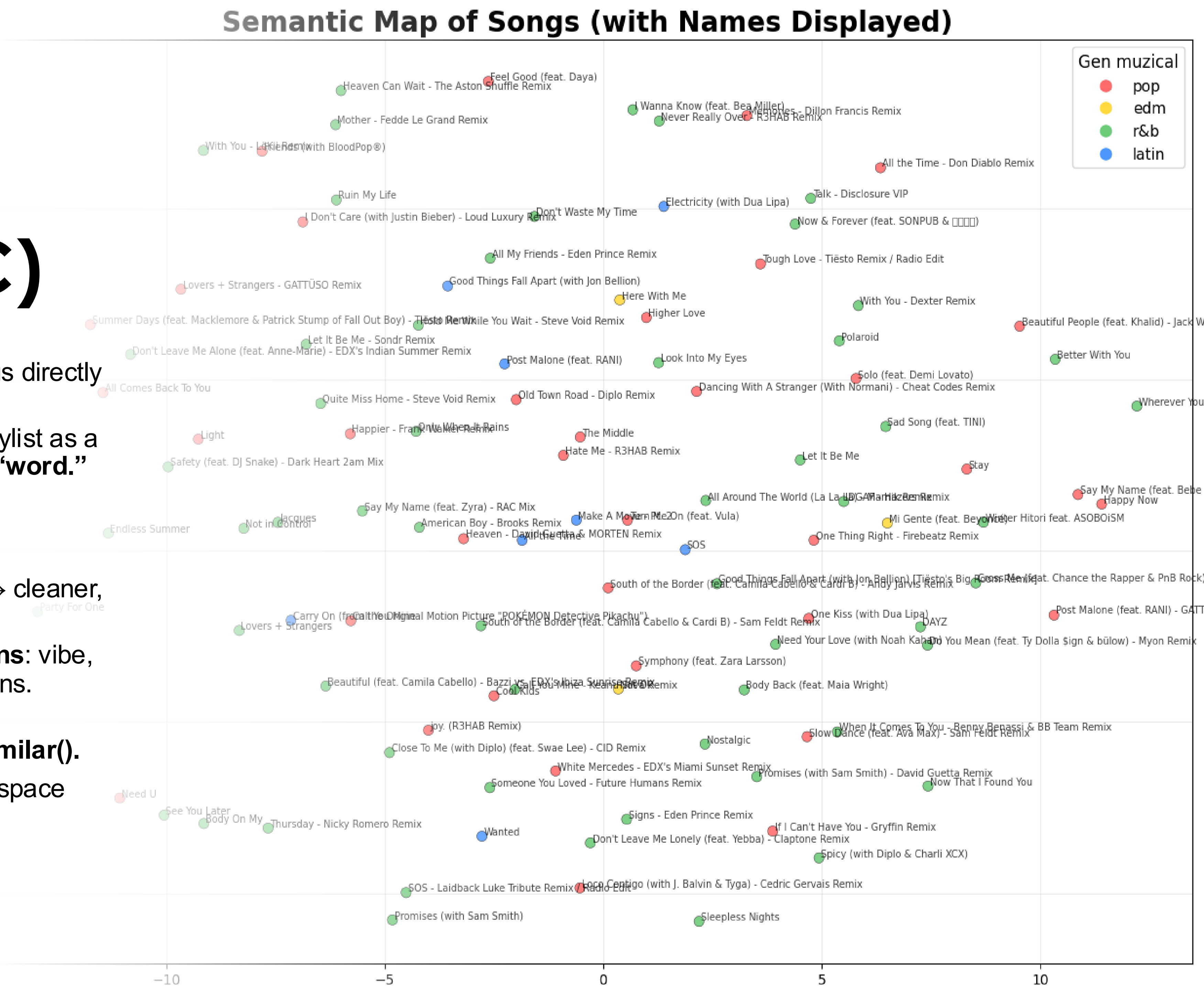
172 rows × 4 columns

```
SVD Recommendations (new songs):
```

	track_name	track_artist	playlist_genre	predicted_score
107	Wrong Way	Sublime	rock	4.295300
241	Sofia	Alvaro Soler	latin	4.109802
96	Lovesong	The Cure	rock	3.986341
370	The Dark	John Clark	rock	3.898507
391	Here I Go Again 87 - 2017 Remastered Version	Whitesnake	rock	3.874134
487	White Flag	Dido	rock	3.792521
90	(Don't Fear) The Reaper	Blue Öyster Cult	rock	3.768494
94	Bedside Radio - Remastered	Krokus	rock	3.753841
328	All In My Head	Tori Kelly	r&b	3.708414
344	The Climb	Miley Cyrus	pop	3.705294

NEURAL ITEM2VEC (WORD2VEC)

- Learn relationships between songs directly from **playlists**.
- Use Word2Vec to model each playlist as a **"sentence"** and each song as a **"word."**
- Train two models:
  1. **Mixed playlists** → general compatibility patterns.
  2. **Genre-structured playlists** → cleaner, more accurate embeddings.
- Capture **subtle semantic relations**: vibe, mood, style, co-occurrence patterns.
- Generate similarity-based recommendations using **most_similar().**
- Visualize the learned embedding space using **t-SNE** (2D semantic map).



LIMITATIONS

1. No real user logs available → CF trained on simulated data.

2. Playlist context was artificially generated → limits Item2Vec.

3. Audio features limited to 10 attributes from Spotify.

4. Models simplified compared to industrial systems.

5. Evaluations restricted by dataset constraints.

CONCLUSION

- CBF, CF and Item2Vec capture **different dimensions of similarity** (audio, behavioral, contextual).
- **Content-Based Filtering** works well for acoustic similarity but lacks personalization.
- Collaborative Filtering **(SVD)** provides stronger **personalized recommendations** through latent factors.
- **Co-listening patterns** are effective for identifying culturally similar tracks.
- **Item2Vec** learns deeper **contextual relationships** between songs from playlist structure.
- **PCA and t-SNE** clearly show how each model organizes the musical space.
- Combining multiple methods creates a **more robust recommender system**.

FUTURE WORK

Designing	Designing a fully custom neural architecture
Building	Building a unified graphical interface
Training	Training on larger, more realistic data
Extending	Extending the analysis to new modalities