Million Songs Recommendation System

We want to make it easier for Music Lovers (users) to select songs from millions of songs based on user's likes/dislikes. By recommending relevant, targeted and personalized songs, the company intends to increase stickiness of its platform. This stickiness creates a happy and expanding customer base with increased revenues for the company.

We intend to build a recommendation system to propose top 10 songs for a user based on the likelihood of listening to those songs.

Data Description

The core data is the Taste Profile Subset released by The Echo Nest as part of the Million Song Dataset. There are two files in this dataset are:

song_data

1. song_id - A unique id given to every song
2. title - Title of the song
3. Release - Name of the released album (Album can have more than 1 song)
4. Artist_name - Name of the artist
5. year - Year of release

count_data

1. user _id - A unique id given to the user
2. song_id - A unique id given to the song
3. play_count - Number of times the song was played

Data Observations and Insights

Total Users	76,353
Total Artists	72,665
Total Songs	999,056 (Song_ids)
Total Song Titles	70,242
Total Albums Released	149,288
Song release years	1922 to 2011
User-Song Density	0.0026% Data is very Sparse

Some Song titles appear in more than one song hence may not be reliable - how to impute them?
For example Title ‘Intro’ appears 1510 times:

	song_id	title	release	artist_name	year
438	SOKTGIX12AB018B354	Intro	The Anatomy Of Melancholy	Paradise Lost	0
657	SOPILLI12A6D4FB80D	Intro	Heat	H.e.a.t	2008
1224	SOVHTPX12A8C140778	Intro	Kinfolk	Ali & Gipp / Big Rube	2002
2081	SOBHCGX12A8C144FC6	Intro	Killing Ground	Saxon	2001
2691	SOFBNLI12A6701E4C0	Intro	Searching For A Land	New Trolls	1972
...	...	...	...	...	...
997004	SOHLOZQ12A67ADB361	Intro	God's Project	Aventura	2003
997496	SODOIRT12AC468926D	Intro	Dub_ Weed & Fyah	Cañaman	2006
998039	SONYNMZ12A8C13D510	Intro	Schritt Für Schritt	Nadja Benaissa	2006
998338	SOQLWLX12AB018D24A	Intro	Sámán	Sámán	0
998819	SOIKTHE12A8C132B0E	Intro	Karma.Bloody.Karma	Cattle Decapitation	2006

1510 rows × 5 columns

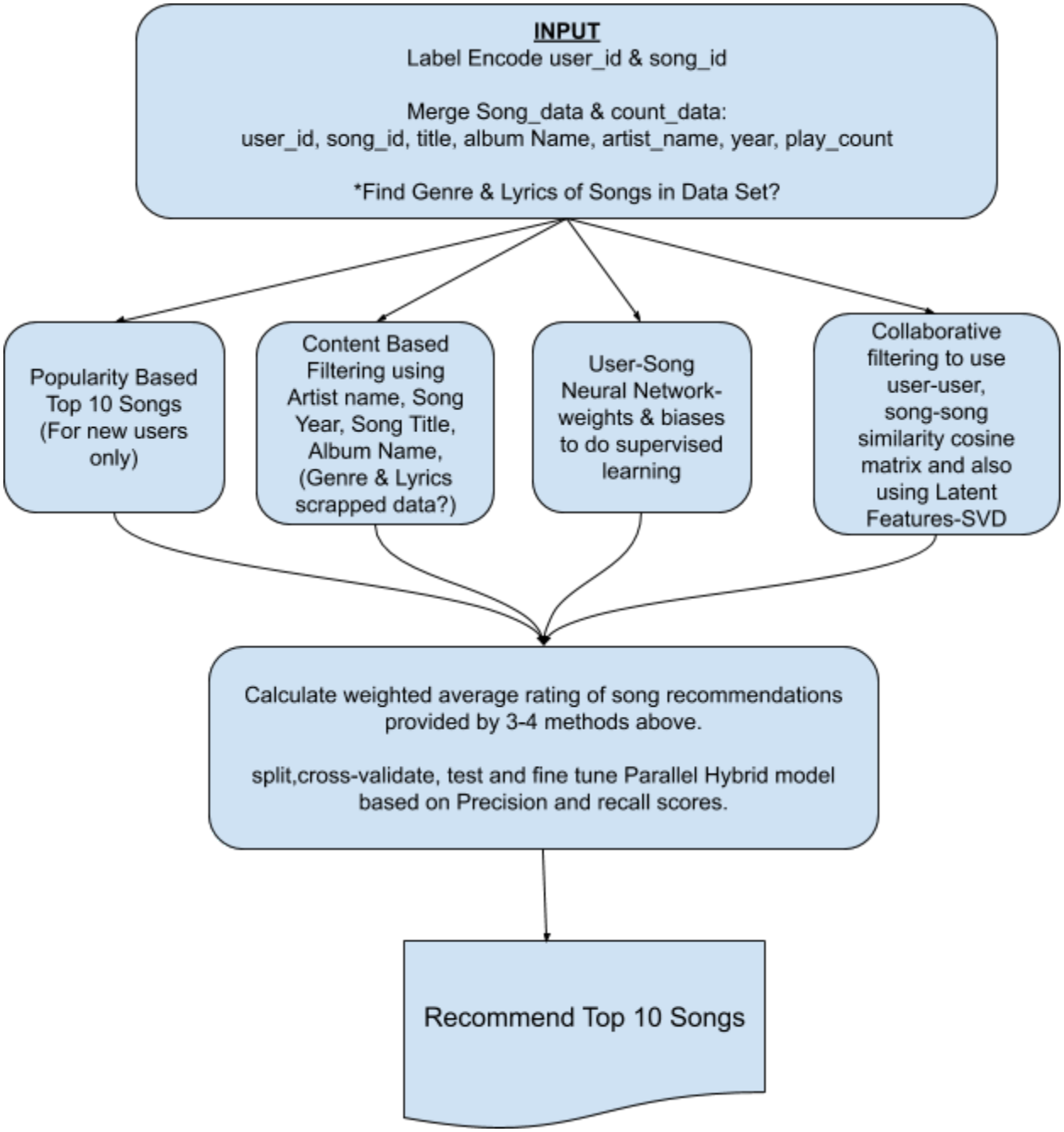
Proposed Approach

1. For Brand New Users :
   This is a Cold Start Problem so use popularity to show Top 10 Songs. (If we were developing front end GUI , would have asked user few questions(Artists they liked/disliked, songs they liked, What decade sogs the liked most- 80’s,90’s ,2000’s, 2010’s etc) to give slightly better recommendations

2. For existing Users:
   Since these are songs and words may not have direct meaning to convey direct meaning to find similar songs. Songs are usually liked based on genre,Artist,decade, tune etc. Song Title may not be much helpful in finding similar songs. Album Name might have some similarity. Lyrics and Genre are not part of the dataset provided.
   Should we scrap Genre and lyrics from web/other sources?

   Because users have rated songs, we can use collaborative filtering to find similar users and recommend songs most similar users to a given user.

A Parallel hybrid Approach will most probably provide us the best recommendation:



Appendix:

Please see accompanying Prelim exploratory Data Analysis Python Notebook