Statement of Work - Spotify Capstone Project

Mehul Smriti Raje, Spandan Madan,
Timothy Lee, Benjamin Sanchez-Lengeling

1 Logistics

Prepared for:

- Aparna Kumar (Spotify)
- Joe Cauteruccio (Spotify)

Running summary of changes: N/A as first version of doc.

2 Background

Spotify has released the Million Playlist Dataset which comprises of 1,000,000 playlists created by Spotify users. This data includes playlist titles, track listings and other metadata. While the exact deliverables of the challenge are yet to be announced, the website lists a paper[3] which outlines the existing challenges faced by state of the art music recommendation systems. While the purpose of this project is not completely aligned with the RecSys Challenge, a possible submission to the challenge may be a possible outcome of the project.
For this project, we will be focusing on two closely related problems - Automatic Playlist Continuation, metrics for evaluation of playlists.

1. **Automatic Playlist Continuation:** It is the most generic problem in music recommendation system. Music playlist is a series of music grouped together based on various factors such as genre, user's musical taste and etc. The task is to recommend an item that most properly fits into the playlist.
2. **Evaluating Music Recommender System:** Metrics used for recommender system are usually not easy to quantify, because musical taste is different for everyone, and such tastes are hard to define.

3 Problem Statement

Goals

Based on the problems mentioned in the literature review and the discussions with the partners, our team has decided to focus primarily on the Automatic Playlist Continuation problem. Our main contribution will come from the use of novel data for this task. Our initial approach lies in building a bayesian hiercichal model based on [1]. We choose this model primarily due it's flexibility in dealing with prior information such as user preferences. We will also explore simpler models, and if time permits, extend our models to neural network architecture for improvements. We also hope to explore different types of metrics that can be used to evaluate our model.

Data and other Resources Available

In addition to the provided data, we have prepared a few more from outside sources. All of the additional data are either scraped from websites or retrieved from an API.

Provided Data

- Million Playlist Dataset from Spotify
  - 5.4 GB Data
  - Created in 2018
- Million Song Dataset from Last.fm
  - Created in 2010/2011
- Lyrics from LyricWiki

Additional Data

- YouTube comments of songs (Scraped)
- Song descriptions from Genius.com (Scraped)
- Spotify Audio Features (from Spotify API)

4 Deliverables

Deliverable 1:

Two predictive models models.

- A predictive models to identify a pool of possible songs that can be used to extend a given playlist. This will be trained in a manner similar to word2vec[2].
- A model trained to rank songs in this pool, to obtain best suggestions for extending the playlist. We aim to emphasize on focusing on new modalities for this purpose, rather than focusing on designing a new, state of the art model using features commonly used in literature. In order to avoid the task of identifying the best representation for each new feature we wish to incorporate, we will try a Bayesian Hierarchical Model inspired by[1].

Deliverable 2:

A github repo with

- Jupyter notebooks (/notebooks)
- A set of python scripts (/src)
- Additional custom datasets (/data/external)

All of the above would be useful for replicating all results from our model. The repo will follow the structure outline as suggested in the ACOMP 297r course.

Deliverable 3:

Brief recommendations on how Spotify can utilize extra information in their playlist generation and continuation efforts.

Stretch goal 1:

An extension of the ranking model mentioned in Deliverable 1 which incorporates additional modalities into the decision making process including user information in the form of lyrics and song attributes obtained from Spotify's API. The model will take an existing playlist, a pool of songs, and rank these songs in order to give suggestions for playlist continuation based on tunable parameters. We also hope to explore and evaluate the importance of order of songs within a playlist.

Stretch goal 2:

Incorporate neural networks, such as

- Convolutional Neural Network (CNN)
- Long Short Term Memory (LSTM)

for our model.

Stretch goal 3:

A Web API for music recommendation running our best model. We will potentially build a website for recommending music playlist, and ask classmates and friends for playlist evaluation. Most likely, we will use a 5-star metric system.

5 Contributions

We aim to offer the following contributions, which will be helpful not only to Spotify but the greater academic community working on music data. Firstly, we hope to augment the Million Playlist Dataset with additional information - lyrics and YouTube comments scraped from the internet. These can be used for content based recommendation systems and most importantly, for obtaining a playlist level representation by aggregating information about songs in a playlist. Secondly, we aim to provide a novel model which first identifies a pool of related songs and then ranks songs in that pool. For the ranking, we plan to propose a hierarchical model which utilizes not only the content of the song and user information, but also information about people's conversations about the songs in the form of scraped YouTube comments. We aim to train this using a variational bayes approach. Finally, we hope to explore metrics for the specific task of playlist continuation, which is different from the task of playlist generation.

6 Tentative Project Timeline

2/09

- Complete first draft of scope document and send to Client for review and approval Project set up
- Private git repository created.

2/16

- Preliminary data exploration completed.
- SOW finalized and submitted, Clear goals and possible extensions defined.
- Preliminary model trained.

2/23

- Define common subset of data to test models for all.
- Furter refinement of initial model.
- Finish scraping spotify audio features.
- Create several ideas for playlist representation.

3/2

- Finish scraping lyric/YouTube features.

3/9

- Begin working on Bayesian Model.
- Begin working on Neural Network model.

3/16

- Begin working on the ranking model.

3/30

- Begin working on the web API and building a website.

5/7

- Final presentations and submission of all deliverables.

References

[1] Shay Ben-Elazar, Gal Lavee, Noam Koenigstein, Oren Barkan, Hilik Berezin, Ulrich Paquet, and Tal Zaccai. Groove radio: A bayesian hierarchical model for personalized playlist generation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 445–453, New York, NY, USA, 2017. ACM.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[3] Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. Current challenges and visions in music recommender systems research. *CoRR*, abs/1710.03208, 2017.