Statement of Work - Spotify Capstone Project

Spandan Madan, Mehul Raje, Timothy Lee, Benjamin Sanchez

1 Logistics

Prepared for:

- Aparna Kumar (Spotify)
- Joe last name(Spotify)
- Pavlos Protopapas (Harvard IACS)

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

2 Background

2.1 RecSys Challenge 2018

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[1] which outlines the existing challenges fad by state of the art music recommendation systems. This has been discussed in greater detail below.

2.2 Ideas discussed in Partner Meeting with Spotify

Some of the points that came out of the partner's meeting are:

- Spandan gave a demo of his word2vec model, and asked Aparna and Joe if we could get a sense of how spotify's existing systems work so that we can build on top of it, and not duplicate any previous failed experiments.
- As it is not possible divulge any information about spotify's existing systems, it's best if we approach this project as independently, as opposed to build on top of Spotify's existing systems.
- They suggested working on incorporating different types input information and data as opposed to focusing our energy on only different kinds of models. That is, data over model.
- They suggested that designing evaluation metrics will be important as it is open ended, and that we should spend some time thinking about how we will evaluate whatever task we decide on.
- They thought scraping lyrics and youtube comments were both good ideas.

- They mentioned that order of songs is an important factor and that we could think about how to incorporate that information.
- Joe agreed to Spandan's idea that the process could be split into two parts - 1) Finding an approximate neighborhood i.e. a pool of interesting songs, and 2) Ranking songs in that pool only.
- Ben suggested Reinforcement Learning approaches. Joe said that is a interesting venue to look at but the metric is going to be hard.

2.3 Literature Review

The three main challenges listed out by the paper are -

1. **Cold-Start Problem:**
2. **Difficulty in incorporating contextual information:**
3. **Evaluation Metrics:**

3 Problem Statement

3.1 Goals

Based on the problems mentioned in the literature review and the discussions with the partners, our team has decided to work on the following concrete goals and extensions.

3.1.1 Concrete Goals:

1. Automatic Playlist Continuation
2. Metric Designing

3.1.2 Proposed Extensions

3.2 Data and other Resources Available

4 Preliminary Experiments and Results

4.1 Pool Identification:

We trained a word2vec like model.

5 Deliverables

5.1 Deliverable 1:

A Predictive Model to identify a pool of possible songs, given a playlist

5.2 Deliverable 2:

A ranking model which takes a pool of songs, a playlist, user information and ranks songs in order to give suggestions based on tunable parameters

5.3 Deliverable 3:

Evaluation metrics designed for the task of playlist continuation which take into account contextual information about the playlist

6 Contributions

Our work offers the following contributions, which will be helpful not only to Spotify but the greater community working on music data. Firstly, we augment the Million Playlist Dataset with additional information which can be used for content based recommendation systems - lyrics and youtube comments scraped from the internet. Secondly, we provide playlist level information by scraping youtube comments and augmenting them. Thridly, we provide a hierarchical model which utilizes not only the content and user information, but also information about people's conversations about the songs. (articulate better). Fourthly, we provide a novel model which first identifies a pool of songs and then ranks songs in that pool. (articulate better). Finally, we provide metrics for the specific task of playlist continuation, which is different from the task of playlist generation. articulate about how we are collecting Playlist level information.

7 Project Timeline

7.1 11th Feb

- Complete first draft of scope document and send to Client for review and approval Project set up
- Private git repository created, TF and professor added. Team communication channel selected to be slack, TF added. Project management tool selected as Trello, TF added.

7.2 14th Feb

- Receive data from client, Preliminary data exploration completed.
- SOW finalized and submitted, Clear goals and possible extensions defined.
- Literature Review submitted as part of SOW.
- Preliminary model trained, have functional results.

References

[1] 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.