Scope of Work for Spotify Project

-- Heidi Chen, Wenxuan Dai, Cindy Zhao, Judy Zhou

Background:

Nowadays, digital music has become the most popular source for music publication and sharing. According to news report, in the year 2015, revenue generated by digital music first overtook the traditional physical copies. Besides digital download, online streaming is a key component to the digital music industry. Companies including Spotify, Pandora, Radio, Groove Shark, Amazon and many others have stepped into the industry, and mostly offers complimentary and paid music subscription services.

Spotify, launched in 2008 as a technology startup company in Sweden, provides music, podcast and video streaming services and has now prevailed in the United States and many other markets around the world. With Spotify, users can expect to find over millions of tracks which can meet a great variety of needs and from multiple platforms – whether it is listening with your phone when working out or playing out of your computer for relaxing at home. The key mission from Spotify is believed to be helping people find the right music at every moment.

Song selection on Spotify, similar to many other music/video playing and sharing website, can be made through direct song search, peeking into friends' collections, and playlist recommendations. In fact, the recommended playlists and resulting playlists from keyword search (including genre or broader descriptive words) are potentially the secret recipes for Spotify's success in its fast growth and its ability to maintain large user base. That is, only with the capability of matching the users' tastes and offering appropriate playlists even with vague descriptive words like "rainy day" in addition to the traditional genre-based search will Spotify be able to surprise its users and increase its users' stickiness to the App.

In accordance with general knowledge about songs, many researches have been done focusing on the connection between the tune or pitch and the classification or grouping of songs. Micro-genre or specific tags, including "synthpop" or "00s songs" are created in addition to the traditional "rock" and "country" genre classification to more delicately meet the users' interests. Their main interest lies in making more flexible but accurate song and playlist recommendations for users – potentially with mood or specific occasion as inputs in addition to genre types.

However, the music itself is not the only component of a song, and thus may not be the only reason for users to enjoy certain types of songs. More specifically, emotions encrypted in the wordings may be worth considering, and thus can be useful when Spotify is responding to a general user search.

Problem statement:

Goal:

The goal of this project is to leverage the rich content of song lyrics to connect each song with relatable concepts such as moods, occasions, and themes. A direct application of this automatic

tagging system would be to produce playlists associated with different emotions or serve specific purposes (after break-up songs, holiday music, party mix, et cetera). An initial target for final product would be a collection of moods and topics that a user can select to retrieve an associated list of songs. A more advanced version would allow the user to type in a specific emotion or adjective and listen to a list of related songs. The ultimate goal is to help create an interactive and highly personalized music experience for the users.

If time permits, we might be able to extend the project further in either modeling or research directions. A modeling enhancement would be to not only process lyrics but also take into consideration other characteristics of songs such as genre, vintage, writer, singer, et cetera, when making connections between them. A research potential of interest, on the other hand, would be to analyze and/or visualize lyrical themes across time. Overall, depending on the data accessibility and quality, we see many potentials in this project and aim to explore various options along the way with the end goal of producing personalized music experience for users in mind.

Resources available:

We explored some existing datasets including Million Song Dataset, related complementary datasets and Yahoo music dataset, as well as several music APIs including Spotify API, YouTube API and Genius API. Data that may be useful for this project are summarized as following:

- **Million Song Dataset (MSD):**
  The core of MSD is the feature analysis and metadata for one million songs. The derived features include sample rate, duration, loudness, energy, and etc. Other metadata include information about the song, album and artist, such as releasing date, artist location. There are also algorithm estimated features: artist familiarity and artist hotness.
- **MusiXmatch (MXM) Dataset:**
  The MXM dataset provides lyrics for 77% of the MSD tracks. The lyrics come in bag-of-words format: each track is described as the word-counts for a dictionary of the top 5,000 words across the set. All lyrics can be directly matched to MSD using MSD IDs and MXM IDs.
- **Last.fm Dataset**
  This dataset provides a song's tags and most similar songs for most of the tracks in MSD. The tags are generated by users from Last.fm API. There are 33,355 different tags in total for 9,330 songs from the training subset. One song can be associated with multiple tags which cover information about genre, emotion, occasion, and etc.

Deliverables:

Deliverable 1	Predictive **model** trained on lyrics and existing articles which: • Predicts theme/mood tags of an input song • Lists a set of songs with given tags in the tag database This will be built with word2vec, NLTK(Natural Language Toolkit), LDA.

Deliverable 2	Advanced predictive **model** which: • Creates a separate model to generate the similar tags that relate to the input tag (not in the tag database) • Based on given tag, generates a list of songs that relate to the tag according to the songs tag distribution and possibly songs' popularity
Deliverable 3	Python **module** which: • Can be run as a standalone script and ready for demo purpose **Website** which: • Can show the result and workflow of the entire project

Project timeline:

Sprint ending	**Tentative milestone or goal(*color grey for past milestones)**
2017-02-07	• Project set up o Private git repository created, TF and professor shared o Team communication channel (e.g. Slack) selected, TF added o Project management tool selected (e.g. Github projects, Trello, waffle.io, ScrumDo, etc.), TF added
2017-02-14	• Setting up Goals & Data Exploration o Decide on final goal of the project o Explore MSD, Spotify API, MXM API, Genius API o Confirm the data on hand is adequate for the rest of the project o Complete first draft of scope document and send to Client for review and app
2017-02-21	• Data Scraping/Cleaning & Tool/Method Learning o Extract data from Spotify API and MSD o Preliminary data visualization, compile list of technical/data and business questions for Client

	○ Research into potentially relevant machine learning and NLP algorithms
2017-03-07	• Learning selected algorithms and toolkits ○ Word2vec (Continuous Bag of Words), LDA and NLTK toolkit ○ Confirm what methods are appropriate given our existing data and goal • Preliminary classification experiment with own data
2017-03-21	• Analyze and report model fitting progress for tags prediction ○ Further process the data into the desired form for different algorithms ○ Train and test on existing tags for model accuracy
2017-04-04	• Extend limited tags to groups of similar tags to perform lyrics-to-tag analysis • More Visualization with fitted predictive models • Prepare for Midterm 2 presentation to Class and Partner
2017-04-25	• Incorporate additional song features to further improve emotion/theme prediction: ○ Could potentially use genre, artist information for better • Finalized models and deliverables for final presentation • Organize codes and create demos for predictive models and recommendation results
2017-05-02	• Poster and presentation preparation and review