Group 8
Saurabh Donthin
Jamie Wheeler
STAT 515-001
Final Project Proposal
05 November 2016

Prince - Computationally True

1. Problem description

Our project will be investigating the music artist, Rogers Prince Nelson, also known as Prince. We will perform both a statistical analysis and a visual analysis of his music, lyrics, and objective metrics. We will attempt to identify patterns and critical factors in his music and the success of his music. Additionally, we will identify the sentiment (happy or sad, on relative scale) of each song. We chose Prince due to the varying nature of his songs sentiments.

The ability to classify songs can promote subscriber retention or value to users for streaming and music recommendation systems. The ability to identify factors of a hit is valuable for artist and record label song promotion and identifying potentially new and successful artists. Further, the ability to quantify things such as sentiment and album structure may offer new art or cultural study of an artist, song, album, and/or genre. Typically, these studies have been done almost solely by opinion, case study, and relative comparison without much quantification.

2. Data description

- For the lyrics, the tool import.io will be used to scrape song lyrics website from the website www.azlyrics.com. After that is done, the data will be stored in a .csv format and delimiters (mostly spaces and commas) will be used to separate each word of the lyric into a single cell.

- The website www.azlyrics.com will also be used to gather a list of songs and albums that have been released by the artist. Import.io will be used for this.
- We will be gathering billboard rankings for the artist from billboard.com. Import.io will be used for this.
- If sufficient data is available, we also plan to gather the chords that have been used in the songs by the artist from www.azchords.com.
- The bpm (beats per minute) data for each song will be gathered from http://www.notediscover.com/, where the BPM value for each song is available. The data for this part of the project will also be gathered via import.io.

3. Methods/Statistical Modeling

Our project will attempt to calculate song sentiment by scraping the lyrics from the websites and performing an analysis compared to a list of 6000 words with a pre-defined positive or negative sentiment rating. Once the sentiment for enough songs is created, we will create several comparative visualizations:

Album Sentiment Curve: A graph showing the change in sentiment by song for an album. Several albums will be shown on the same graph to identify any common pattern. This will also be modeled by regression to see if there is a statistically significant relationship between song location on album and sentiment.

Sentiment to Billboard Position: A visualization will be completed to show song sentiment related to billboard position.

Album Sentiment by Age: A visualization will be created to identify any relationship between Prince's age and the overall sentiment of the album.

We will perform several regressions:

Linear regression: Identify the relationship between sentiment, position on album, billboard position, year, song length, beats per minute, and song key, album. Target variables are likely to be sentiment, billboard position, and position on album. Other target variables may be attempted to verify if any relationships exist.

Logistic regression: Classify if a song will be a hit (position on billboard at 100 or lower). A confusion matrix will be created to show the accuracy, precision, and specificity of these predictions. It will be cross validated to actual billboard positions.

4. Expected Results

- Identify relationship between song sentiment, position on album, year (Prince's age), billboard position, and whether nor the song was a hit.
- Identify possible common album structure as regards to song sentiment.