This form is due Thu Nov 12, 2015 at 11.59pm EST. Based on your proposals we will assign a TF to your team who will guide you through the rest of the project. You will schedule a project review meeting with your TF (mainly during regular lecture times of the week marked in the schedule). Make sure all of your team members are present at the meeting. Online students can schedule a Skype/Hangouts/Slack meeting with their TF.

**The proposal is submitted, we can edit the response here.**

**ADDITIONAL (AFTER SUBMISSION)**

**For the screencast, consider VideoScribe and Tawe from Sparkol**

**Can use Slack for Team coordination**

**Can use GitHub Pages for code / website**

**Refer to Lab11 for approach to final project**

**NPR recent article from NOV 08 The Art Of The 'Clean Version'**

- **"When I spoke with Guerini [Radio Disney], he said almost half of the more than 50 songs on rotation at the time had been edited."**
- Billboard Mag article from 2014 on the cleaned up Meghan Trainor song "All About that Bass" which was referenced in the NPR article.

Article: Spotify teams up with Musixmatch for lyrics to 9 million songs

Rap Genius offers lyrics and explanations for a variety of genres

| Category | Method | Comments |
|---|---|---|
| Regression | Elastic Net Regularization | logistic ridge regression with lasso (lab11) |
| Dimensionality Reduction | PCA | |
| Dimensionality Reduction | LDA | |

| | | |
|---|---|---|
| Unsupervised Learning | K-Means | random init, fixed k (you decide); address with “Knee” or “Elbow” method after computing sum of squares<br>cross-validation to apply to new unseen data |
| Unsupervised Learning | Mean Shift, also (multi-feature object trajectory clustering) | don’t need to know k, specify a window around a point, handles arbitrary sized cluster, embarrassingly parallel |
| Unsupervised Learning | Hierarchical Clustering | no k or window size needed but must decide what defines min and whether to do single (forms long chains), average, or complete linkage (sensitive to outliers). Essentially start with k=n, compute distance from all point to all point and cluster by shortest distance order (in hierarchy), then choose threshold |

Evaluate Clusters (looking for something that generalizes well to new data)

- Rand Index: Percentage of correct classifications, compare pairs of elements (TN,TP,FN,FP), requires knowing results for assessment
- **Stability: split data, second set should explain the first, find where they agree, pick k based on that**
  - turn this into a supervised learning problem (assign ad-hoc labels)
  - then train a classifier of choice on set 1
  - then test classifier on set 2

FILL IN JUST ONE FORM PER TEAM.

I REPEAT: ONE FORM PER TEAM.

We strongly encourage you to fill this out as early as possible. Then we can have your assigned TF meet with you earlier, and get you jump-started. You DO NOT get your project approved until you meet your TF.

* Required

REMEMBER: PROPOSALS DUE THU 12th NOV, 11.59PM EST

## Title of your project proposal *

Modern Music Sentiment Analysis

## Team Name *

Lyrics Lab

## Team Member 1's Github username *

cs109-kbuhrer

## Team Member 2's Github username *

financedoc

## Team Member 3's Github username *

michaeljohns

## Team Member 4's Github username

N/A

## Background and Motivation: Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.

Music may tell us what the artist wants to say, but popular music tells us what the people want to hear. And that can tell us much about who they are. This project aims to discover the meaningful messages about cultures and their times, as revealed by the lyrics of popular music.

It has long been recognized that studying a culture's popular music can yield insight into its values and mood, but until recently analysis was confined to subjective examination of small samples. Over most of the history of cultural musicology, research was hampered by the lack of available data. And when songs themselves became data stored on Compact Discs, the computing power necessary to conduct analysis was being devoted to other uses. Only recently has the abundance of storage and computational capacity made it feasible to study music's messages empirically.

In recent years, considerable effort has been devoted to studying and classifying two of the three basic ingredients of music- tones and rhythms. We propose to study the third one, words, which have received comparatively less attention. In the 2008 Proceedings of the Association for Computational Linguistics Xia et al wrote "research efforts on lyric-based song classification are very few." (Sentiment Vector Space Model for Lyric-based Song Sentiment Classification, Proceedings of ACL 2008). Other representative papers include Zhong, et al (2012), Music Sentiment Classification Integrating Audio with Lyrics, JOICS 9:35. The most prominent work in empirical musicology to date is the Million Song Dataset which primarily contains features attributed to the tones and rhythms. The set does contain lyrics in reduced form, but not full text.

We expect to apply the tools of machine learning to extract thematic content from the lyrics of popular songs in the United States and, if possible, internationally, reaching as far back in time as we can get usable data. We hope to find themes that change over time and correlate them with significant historical events. And if it goes better than we expect, we may be able to produce a classifier that can accept text and predict its genre and era.

We have not yet found an existing study that assembles a large corpus of lyrics and classifies them based on their semantic content. Studying song lyrics is different than general sentiment analysis of, say, a culture’s literature. Lyrics can have attributes that would be out of place in literature or formal prose, such as repetition, rhyming, and rhythmic delivery. Moreover, one might speculate that lyrics may have a higher tendency to be formulaic or even clichéd, making them more amenable to pattern analysis than prose generally.

At this early stage, we do not offer much in the way of predictions about the results. We look forward to discovering them along the way.

## Project Objectives: What are the scientific and inferential goals for this project? What would you like to learn and accomplish? List the benefits.

This project will focus on the lyrics of songs that have been elevated into the popular consciousness, considering hits from this and previous decades. While music is often layered with context and nuanced with subtext, this project hopes to uncover those themes and sentiments that both reflect and have served shape generations, to identify what is fleeting and what endures.

Lyrics Lab will employ powerful tools of data science to myriad experiences and ideals as articulated in lyrics, to approximate how they express what it means to be human. It will leverage modern statistical prediction, machine learning, and data mining techniques to accomplish our objectives. A guiding intuition to our approach will be that vice and virtue are classifiable dividing lines in song lyrics. The nouns will give us the vice/virtue split while the adjectives will offer the positive/negative sentiment towards the referenced topic. Through careful design and documentation, our process will be able to iteratively incorporate new hits, or be re-implemented to train any conforming set of lyrics.

Here are primary questions Lyrics Lab should assist in answering:

1. What are the topics raised within popular music? Are there any enduring topics?
2. When are select topics present? At what ranks? How long are they present?
3. Some genres of music are criticized for lack of depth (Rap, Pop, and others) while some have greater variation. Are those variations observable?
4. What artists, songs, and topics reflect vice? What reflect virtue? What is their sentiment towards vice/virtue?
5. Given any combination of artists, songs, and topics, what else might be of interest?

Here are secondary questions that Lyrics Lab could assist in answering:

1. Given any song's lyrics, how well might the song perform in select year? Given a sample of text, can we infer what genre it fits, and in what time period?
2. What topics has an artist expressed? Does an artist consistently express any topics?
3. What are topic preferences for a given geographic area?
4. Do musical trends correlate to select cultural trends? Do musical trends correlate to select historic / marked events?

## What Data? From where and how are you collecting your data? Is the data publicly available? How big is it?

Song lyrics offer an accessible corpus to analyze for topics and sentiments. Billboard provide numerous ranked charts that capture popularity of music. The charts consider both genre / sub-genre and artist and are backed by airplay, sales, digital downloads (since 2005), and streaming (since 2007). Though Billboard no longer supports a public API, Wikipedia preserves the lists. We have chosen Billboard's US Year-End Hot 100 Singles as the foundation data source for this project, which has been maintained since 1951 (only top 30 available from 1951 to 1955). We will explore drawing in other charts for either training or prediction once the scope of the project is settled. Of interest chart options include the following:

- International
  - Canadian Number One Hits, e.g. 2015 (available 2007-2015)
  - European Number One Hits -- Digital (available 2008-2015)
  - Brazil Hot 100 -- Airplay (available 2009-2015)
  - Japan Number One Hits (available 2008-2015)
- Genre/Sub-Genre: Pop, Dance/Electronic, R&B/Hip-Hop, Country, Latin, Religious
- US Number One Artists (explore body of work)

We will need to obtain the lyrics for each identified song within the charts. LyricWiki has a public API that is most promising. For one-off gaps, we can use other sites such as songlyrics.com.

The project will be initially constrained to English songs, though the process could be followed or extended to accommodate any language. Also, for a portion of our analysis we will seek to establish a vice/virtue vocabulary of topics within songs, seeded from an existing word list, then tuned for actual vernacular used in modern songs.

## Must-Have Features: These are features or calculations without which you would consider your project to be a failure.

Lyrics Lab must be able to identify similarities and differences within music messages over various time spans. It will need to offer analysis ranking, inter-genre, intra-genre, and artist. Here is a table of primary questions and what the project should employ to address:

| **Primary Question** | **Feature(s)** |
|---|---|
| 1. What are the topics raised within popular music? Are there any enduring topics? | ● Interactive N-Gram Component<br>● Interactive Vice/Virtue Component<br>● Static products and summary from analysis |
| 2. When are select topics present? At what ranks? How long are they present? | ● Static products and summary from analysis |
| 3. Some genres of music are criticized for lack of depth (Rap, Pop, and others) while some have greater variation. Are those variations observable? | ● Genre Isolated Interactive N-Gram Component<br>● Genre Isolated Interactive Vice/Virtue Component<br>● Static products and summary from analysis |
| 4. What artists, songs, and topics reflect vice? What reflect virtue? What is their sentiment towards vice/virtue? | ● Interactive Vice/Virtue Component<br>● Static products and summary from analysis |

| | |
|---|---|
| 5. Given any combination of artists, songs, and topics, what else might be of interest? | ● Interactive Recommender Component<br>● Static products and summary from analysis |

## Optional Features: Those features or calculations which you consider would be nice to have, but not critical.

Here is a table of secondary questions and what the project could employ to address:

| **Secondary Question** | **Feature(s)** |
|---|---|
| 1. Given any song's lyrics, how well might the song perform in select year? Given a sample of text, can we infer what genre it fits, and in what time period? | ● Song Prediction Component |
| 2. What topics has an artist expressed? Does an artist consistently express any topics? | ● Artist Isolated Interactive N-Gram Component<br>● Artist Isolated Interactive Vice/Virtue Component |
| 3. What are topic preferences for a given geographic area? | ● Spatio-temporal Isolation Component |
| 4. Do musical trends correlate to select cultural trends? Do musical trends correlate to select historic / marked events? | ● Ad-hoc manual analysis |

## Design Overview: List the statistical and computational methods you plan to use. Are you planning to use AWS?

The table below depicts the methods we will heavily evaluate in our design phase. The features are listed in initial prioritized order. Those features which address secondary questions are stretch goals for the project. During the design phase, any of the features may be ultimately combined or re-imagined to better address standing project questions.

We intend to take advantage of MPP using Spark for data conditioning and at least a portion of the machine learning. However, we will decide whether or not to use AWS resources when the project is further along and the processing requirements of the lyric corpus is better understood.

| **Overall Priority** | **Feature** | **Question Level** | **Statistical and Computational Method** |
|---|---|---|---|
| 1 | Interactive N-Gram Component | Primary | ● Use NLP (Natural Language Processing) to extract grammar and semantic meaning from the resulting tokens, e.g. Penn Treebank II tag set.<br>● N-gram (unigram, bigram, trigram)<br>○ Zipf's Law<br>○ Stemming and lemmatization<br>○ Stop Words |
| 2 | Interactive Vice/Virtue Component | Primary | ● LDA (Latent Dirichlet allocation), use gensim -- use to find the topics from the nouns<br>● Naive Bayes (or other) -- use to do sentiment analysis from the adjectives, typically, sentiment analysis is done using an external data set such as SentiWordNet<br>● TF (Term Frequency) IDF (Inverse Document Frequency)<br>● Bag of Words (BOW) Bayesian<br>● Topic Modeling<br>○ Intertopic Distance Map<br>● n-fold cross-validation<br>● SVM |

| | | | |
|---|---|---|---|
| 3 | Interactive Recommender Component | Primary | ● Ensemble Method<br>○ KNN (common support of topics) for recommendations<br>○ Ridge Regression<br>○ SVD<br>○ Latent Factors |
| 4 | Song Prediction Component | Secondary | ● K-Means : Clustering Algorithm, probability density of features<br>● process text according to implementation |
| 5 | Spatio-temporal Isolation Component | Secondary | ● Post-filtering of results for spatio-temporal rendering |

## Verification:

How will you verify your project's results? In other words, how do you know that your project does well?

Verification will be conducted on a variety of levels. First and foremost common sense and a logical approach will be taken so that results are interpreted and understood during analysis. This will avoid issues related to “plug-n-chug” or blindly accepting output. The methodologies will account for folds, test sets, and validation sets as appropriate to help ensure valid results.

Automated approaches can be used for a variety of the factors. For example genre may be associated with a given song, artists also have associated genre(s). These can be cross referenced between various sources (Billboard, Wikipedia, lyrics). Similar cross references can be made for sales, artist income, and general song content/themes.

Manual methods can also be employed for other results: spot checking lyric content and sentiment compared to findings; comparing literal word counts from lyrics to n-gram results to confirm they “make sense”; among other approaches depending on the specific methodologies that are ultimately used.

## Visualization & Presentation: How will you visualize and communicate your results in your video and website?

The most important feature of the final deliverables is telling a compelling narrative of results observed through our study. All visualizations will be selected to support the narrative which will only become clear during the analysis process.

For the website, from a high-level perspective, a single page will be used with clean modern design elements. Beginning with summary information and an engaging headline the users will be introduced to the topic. While scrolling from top to bottom the narrative will unfold giving insights that are comprised as major sections. This format will derive inspiration from infographics, but will expand into further detail and allow room for large charts and graphs. It will close out with key take-aways.

The visualizations may come in any combination of static images generated from Seaborn and PyPlot, embedded Tableau graphs and dashboards, and D3.js based visualizations. The determination of specific chart-types (line, bar, word cloud, etc...) will depend directly on our findings. Every effort will be made to provide a balance of appealing design and accommodation for accessibility through means of color selection, iconography, and the like.

The video will largely contain elements that are created for the website and feature voice-over narration to provide an overview of the content. The website elements may be taken directly (preferred to save on redundant effort) or reworked specifically due to limitations of the video presentation. Rework may include converting an interactive graph to an animation or adjusting ratios and form factors.

## Schedule / timeline: Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.

| Date | Milestone | Lead | Detail |
|---|---|---|---|
| 11/12 | Proposal Due | Michael Johns | This document |
| 11/16 - 11/22 | Project Review | N/A | Meet with Assigned TF |
| 11/22 | Data Conditioning, Design Finalized, and | Michael Johns | All charts, lyrics, and vocab ingested and |

| | | | |
|---|---|---|---|
| | Team Huddle | | available for analysis and design finalized |
| 11/29 | Discovery and Analysis and Team Huddle | All | All statistical and machine learning methods established |
| 12/06 | Project Website Finalized and Team Huddle | Kevin Buhrer | Summarize results |
| 12/10 | Project Due | Scott Stephens | All required artifacts delivered (process book, screencast, and website) |

## Team Member Contributions: List the contributions each team member will make.

All: Narrative writing, analysis and data exploration as divided by group, participation in team huddles, contributions to IPython Process Book.

- Kevin Buhrer: Website (design and visualizations).
- Michael Johns: Data conditioning (charts, lyrics, vocab).
- Scott Stephens: Final project artifacts (screencast and summaries for website)

Project proposals are due Nov 12. Final project process book (ipython notebooks) is due Dec 10. Project webpage and 2 minute screencast is due Dec 10. For more information on the milestones, deliverables and other due dates on the course website:
http://cs109.github.io/2015/pages/projects.html

….

**PROPOSAL FORM**

**This form is due Thu Nov 12, 2015 at 11.59pm EST.**
Based on your proposals we will assign a TF to your team who will guide you through the rest of the project. You will schedule a project review meeting with your TF (mainly during regular lecture times of the week marked in the schedule). Make sure all of your team members are present at the meeting. Online students can schedule a Skype/Hangouts/Slack meeting with their TF.
**FILL IN JUST ONE FORM PER TEAM.**
**I REPEAT: ONE FORM PER TEAM.**
We strongly encourage you to **fill this out as early as possible**. Then we can have your assigned TF meet with you earlier, and get you jump-started. **You DO NOT get your project approved until you meet your TF**.
The form is at:
http://goo.gl/forms/LdTEZL2c2S