Detecting Release Decade From Song Lyrics

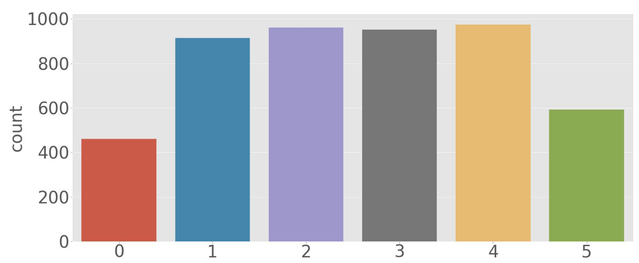
Joe Sommer
jsommer1@stanford.edu

Motivation

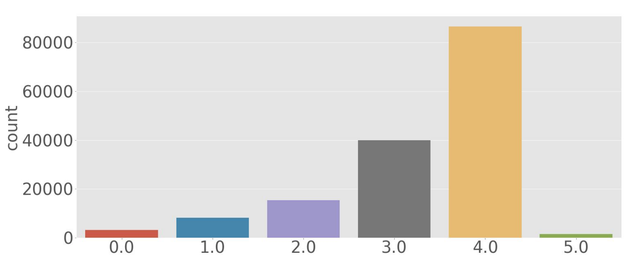
The way people speak changes over the years, and can be reflected in pop culture texts such as books or songs. There may be merit to being able to determine what time period something was written just from text, so I attempted to train a couple different models to determine the decade (1960's through 2010's) that a song was released in. My overall best model achieved 40% accuracy between the 6 different decades.

Data

- **Billboard's Year-End Hot 100 (1965 - 2015)**
  - 4,853 usable songs (dataset from Kaggle)
  - All lyrics in order as a string
  - Distribution mostly equal (60's and 10's small)

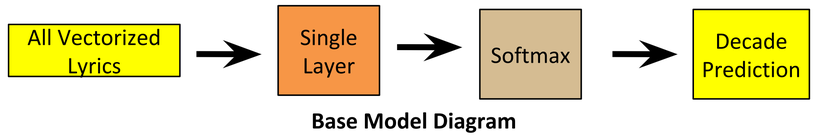


- **Million Song Dataset**
  - 154,527 songs with lyric information
  - Lyrics in Bag-of-Words format, not sequential
  - Distribution heavily skewed to 2000's



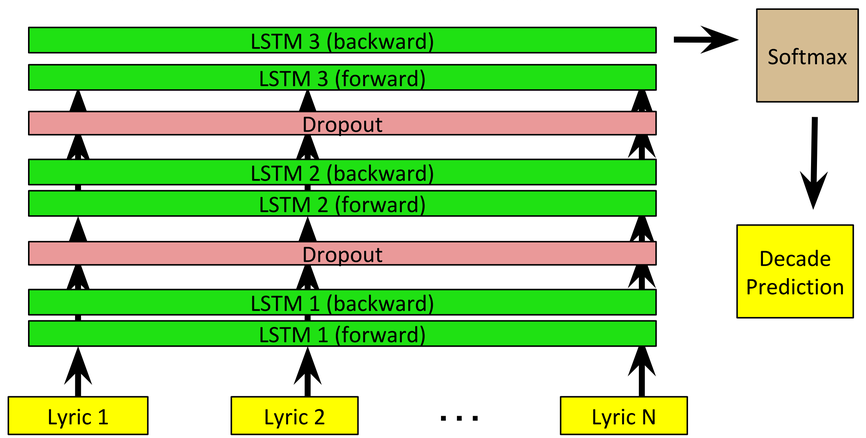
Models

- **Base Models**
  - Single hidden layer -> Softmax
  - One trained on Billboard data, one on Million Song data



Base Model Diagram

- **Sequential Models (Bidirectional LSTMS)**
  - **2-Stack**
    - BLSTM → Dropout → BLSTM → Dropout → Softmax
  - **3-Stack**
    - BLSTM → Dropout → BLSTM → Dropout → BLSTM → Softmax
  - Both models trained on Billboard data



3-Stack Model Diagram

Features

- Vectorize inputs into token counts
- Base Model
  - Vectorizing pads all input lengths to maximum input length (4936)
- Sequential Models
  - Manually determine maximum input length, pad slightly longer (1200)

Results

Model	Train Accuracy (%)	Test Accuracy (%)
Billboard	99.26	37.77
Million	67.23	52.53
2-Stack	99.18	40.11
3-Stack	79.25	37.71

- Not fantastic, but better than randomly guessing
  - 1 / 6 = 16.67%
- Best test accuracy from Million model, but data is too skewed towards 2000's to actually predict other decades
- 2-Stack gives second best test accuracy, but model is extremely overfit
- Base Billboard model is surprisingly not too far behind 2-Stack

Discussion

- Models trained on Billboard are overfit, so we need more data
- Million model is skewed too heavily, so we need more even distributions of data
- Classes could also be hard to distinguish
  - Ex: are songs from 1969 and 1970 that different?

Future

- Potentially promising, but need to address overfit
- Needs more data evenly distributed across decades
- Pre-trained NLP models may help
- Could try tweaking classes
  - Shift start and end, or increase time period