

**Team: Hitachi #2**

Weekly Report #6

03/31-04/07

Team Members: Members: Jay Naidu, Daniel Day, Andrew Urquhart, and Kebba Leigh

# Summary

Over the course of the past week our team met and focused our efforts making substantial progress on all fronts in the pipeline from data extraction to visualization. Furthermore, we established our research questions for the project, and though it’s not too dissimilar from what we’d already discussed, we’ve now solidified it.

# Project Status

* The project is organized into four primary phases: Data Extraction, Cleaning, Vectorization, and User Integration & Visualization.
* Data Extraction - 100% | Data extraction is technically complete, but Daniel is continuing to add data and iterate through his code to include as much data as possible.
* Cleaning - 90% | We’re currently operating on data that has been cleaned through Pentaho’s regular expression formula capabilities, but we’re planning on transitioning this to Python for the final process (more on this later)
* Vectorization - 67% | We’ve completed the vectorization of our lyrics, and incorporated our first metadata analysis. We expect to iterate on this version to improve later
* UI - 0%? |
* Research Questions:

1. **Is today’s music less original than the music of previous decades?**
2. **How have music genres evolved over time, lyrically?**
3. [Additional] - **What topics do we commonly see in pop music today, and how has that changed?**

## Key Achievements This Week

* **Daniel**
  + Finished extracting data from original 10k song dataset
  + Created automatic Web-scraping script that takes the top 100 songs from every year since 1950
  + Added additional data cleaning methods to protect the dataset from duplicates and foreign songs.
  + Added robust try/except troubleshooting to all code.
* **Andrew**
  + Loaded and cleaned the data set (removed duplicates and short lyrics).
  + Embedded lyrics using all miniLM-L6-v2.
  + Formatted and uploaded vectors to Pinecone New-Lyrics-Index.
  + Queried the index using a sample input and returned metadata.
  + Added one metadata field.
* **Kebba**
  + Created initial Figma prototype for the user-facing website UI, including search functionality, cluster visualizations, and song-level detail cards.
  + Identified and documented necessary metadata fields to enhance user filtering options (e.g. genre, popularity, release year).
  + Set up basic Power BI dashboards to explore the cleaned data and test visual layout of similarity graphs and user interactions.
* **Jay**
  + Began the transition away from Pentaho
  + Put in charge of exploring our third research question.
  + Explored basic sentiment analysis through nltk and Textblob, will begin ci/cd of feature extraction and similarity calculation.

Meetings/Submissions:

* Team meeting with Dr. Vllasi an Sadovnik on 04/01/2025
* Had 3 hour internal team meeting on 04/02/2025
* Submitted project Gantt chart 04/03/2025

## Next Steps

* **Kebba**: Build a working website using React or a lightweight front-end framework, incorporating the finalized Figma design and clustering visualizations.
  + Fully develop the Power BI dashboards for presentation, including genre breakdowns, similarity score distributions, and lyric-based comparisons.
  + Prepare narrative and transitions between UI elements for live demo during final presentation.
  + Collaborate with Andrew to embed additional vector metadata into the visual layers of the UI.
  + Work with Jay and Daniel to define how sentiment or genre-based NLP results should be visually conveyed on the platform.
* **Daniel:**
  + Finish compiling every year’s top 100 songs
  + Perform regular expressions in Python
  + Prepare for the final presentation.
* **Jay:** 
  + Finalize the analysis methods used, leveraging word embeddings and topic modeling to determine lyrical context and themes present, and then use NLP models for similarity analysis of lyrical similarities across genres and time.
* **Andrew:** 
  + Add additional metadata
  + Explore filters to better address research question
  + Normalize additional metrics to combine them with similarity scores
  + Test additional embedding methods.

**Risks/Roadblocks**

* We may not have enough time or manpower to address the additional research question
* Everyone is contributing to their own sections, which means our current progress hinges more on certain team members than others.