CS 410 Final Project Progress Report

Team “Bae Area”

Free Topic

Improved Media Content Search

Team Members

Brayden Turner - brturne2@illinois.edu - *Captain*
Joshua M. Smith - jms28@illinois.edu

Progress Made

So far, we have completed the following tasks:

- Set up the environment and work through initial build issues: we are using the MeTAPy library for tokenization and n-gram creation, which requires an earlier version of Python, and we are committing code in Jupyter. We tested using an online collaboration notebook but abandoned this due to the size of our dataset. We have landed on self-hosting Jupyter notebooks on Python 3.6.8 and using git for syncing progress.

- Dataset download: we found a data set of 10,000 top music artists and successfully implemented the Lyrics Genius API to download the lyrics for all songs by these artists. This code is using the Multiprocess library to speed up downloads.

- Early tokenization and pre-processing: we have begun to implement MeTAPy, creating functions to process the lyrics in our dataset file. We have discovered that we may require some additional pre-processing prior to tokenization, and are currently working on the strategy here.

Remaining Tasks

The following core tasks remain:

- Implement a workflow for fuzzy relations between words: Ideally, we would like our search to be able to account for paradigmatic relations and rank related words to what the query contains, not just exact matches.
- Implement the search model, e.g. PLSA to retrieve documents with relevant words
- Evaluate search performance
- Write final report with conclusions
- If time allows: create a web frontend for the application

Unknowns

Our original proposal scoped the project as both music lyrics and movie summaries, whether we are able to easily implement movie search will depend on how adaptable our existing code is to a different dataset of movies. If it proves too difficult this may fall out of scope.