Visual Analytics 690V – Midterm

- Ajay Shaan Shanmugam and Siddharth Chandrasekaran

Dataset used:

lyrics.csv (Lyrics and metdata of songs), downloaded from

https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics.

Please sign in to kaggle and download the dataset in the link above. Unfortunately, the moodle submission does not allow a zip beyond 50mb and this amounted to around 90mb.

We've attached a subset of it (30000 records) in a csv file for emergency, but please use it only if the whole dataset above can't be downloaded.

The Jupyter notebooks contain documentation of our analysis and all inferences we made from them.

Notebooks:

1. ExploratoryAnalysis.py.ipynb – Contains our exploratory analysis of the dataset.

2. Clustering 1 Final.ipynb – Contains clustering analysis on artists based on their lyrical usage.

3. Clustering 2 Final.ipynb – Contains clustering analysis on genres based on their lyrical usage.

Note:

1. Please update all NLTK packages when the code runs the nltk downloader. Once they're installed, please close the window to continue.
2. Please install gensim by entering pip install gensim in your console.
3. Since parsing the entire dataset of 380000 records was too intensive for our machines, we just used a subset of it for clustering.