Group

Mudit Bhargava (mbhargava@umass.edu)
Priyadarshi Rath (priyadarshir@umass.edu)

Steps to run the file

- Please download the dataset from https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics. Keep the dataset in the same folder as the midterm.py file.
- NLTK download and installation will be required. Please follow the steps given below
  - Run the python command line
  - >>> import nltk
  - >>> nltk.download()
  - Check if nltk has been downloaded properly
    - >>> from nltk.corpus import brown
    - >>> brown.words();

Run from command line

```
bokeh serve ––show midterm.py
```

Kindly note the loading, parsing, cleaning, transformation and topic modeling on this huge dataset takes time. It may take around 120 seconds for the notebook to show up in the browser.

Direct Data Download link
https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics/downloads/380000-lyrics-from-metrolyrics.zip