Billboard Lyrics Analysis

This project is co-developed with Will Hwang, H.I. Park, and Whan Lee as a part of CMSC 12200: Computer Science with Applications II at The University of Chicago. All relevant files are in the mysite folder.

This project involves following steps:

- Raw Billboard data processing: data cleaning into analysis-friendly format + finding "top n" artists
- Web crawler: scrapes lyrics of chart-in songs from websites
- Main task: given user input, compares it to the lyrics data and returns an artist name / time period (in decade) of which lyrics style is most similar to the input
- Side task: lyrics style comparison between artists, most positive and negative songs of a given artist or a decade, as well as analysis of word frequency change over time

Also, this project used several packages including:

- Pandas (data storing & processing)
- Urllib3, certify (web crawling)
- Genism (word2vec), NLTK (lyrics sentiment analysis)
- Unidecode (Unicode conversion of irregular characters in lyrics data, etc.)

Below are the descriptions of several key files of the project.

bill_board.csv: original Billboard data with weeks, artist names, and titles of historical top 100 songs up until the end of 2019 (from data.world | posted by Sean Miller)

lyrics_crawler.py: processes Billboard data, generates list of "top n" artists, and extracts their song lyrics from websites

artist_lyrics.txt: saved result of lyrics_crawler.py (has artist name, first chart-in year, title, and lyrics)

main_task.py: does the main task of quantitative comparison between two lyrics + conducts side task analyses

manage.py: executes Django web interface

To run the web interface, on a unix-based terminal, do the following:

1. Go to the directory where these files are stored
2. Go into the mysite sub-directory
3. Execute: "python3 manage.py runserver"
4. Access http://127.0.0.1:8000/create/ on your web browser

For more detailed description about the project, please refer to the final project presentation included in the repository.