Proposal: Lyrics Web Scraping and Text Mining Analysis

Problem:

To study hot 100 songs on billboard year-end from year 1959 to 2018.

Dataset:

1. Wiki – Billboard year-end 100:
   https://en.wikipedia.org/wiki/Billboard_Year-End
2. Lyrics:
   https://www.azlyrics.com/
   http://www.metrolyrics.com/
3. Genre: Google search

We would like to do web scraping and data cleaning from these websites, and then do text mining analysis.

Proposed Solution and Real world Application:

Our proposed solution is to do web scraping and data cleaning from these websites. For web scraping, many websites have formatted HTML interface, so it is not hard to extract basic information about hot 100 songs from wiki, and even some lyrics and genre. After that, we will do data cleaning by using regular expression in Python. We plan to set up many variables, such as song name, artist name, lyrics, genre, rank, ranking year and lyrics text length, etc. Furthermore, we would like to do exploratory data analysis and text mining based on some dimensions, like style, decades and rank. For metrics, we plan to focus on word use (frequency, repetition, diversity), top artists, sentimental analysis and other text mining strategies.

The real world applications of this solution is that we can know what factors (e.g. genre, length of lyrics, repetition of some particular words, etc) contribute to a popular song, which could be applied for popular songs prediction or music appreciation.

Project steps:

Step	Estimated completion time	Person(s) in charge (among the group of 3)
1. Extracting data	One week	(Zhaoyuan He, Anwesan Pal)
2. Cleaning up data	One week	(Qinyan Li, Yihua Yang)
3. Exploratory data analysis	Three weeks	(Zhaoyuan He, Anwesan Pal,

& text mining (by style, rank, decades, etc.)		Qinyan Li, Yihua Yang)