LYRICFIND

NUS
National University
of Singapore

http://lyrics.smcnus.org

This README provides some explanation about the data made available.

- **cleaner (folder)**
    - A Python script which calls five separate "rules" files, each of which handles a set of cleaning operations performed on the lyric in its original form: hyphenation, contractions, misspellings, variant spellings, etc.

- **LF_SUBTLEX_merged_20150719**

    66,975 rows, with each row providing the document frequency (DF) and corpus frequency (CF) associated with a given lyric word. DF and CF values are provided both for the LyricFind corpus of 275,905 unique songs, as well as values from the SUBTLEX-US corpus (http://subtlexus.lexique.org/) which were used to calculate our lexical novelty score.

    - For the SUBTLEX-US DF and CF values, the "all" reflects the fact that DF and CF values refer to counts of words in both upper case and lower case.
    - In a handful of cases, the set of cleaning rules used to process the lyrics yielded an orthographic representation of a given word (e.g., "mrs" transformed to "missus") that was *uncommon* in the SUBTLEX-corpus (e.g., "missus" has a DF = 151/8388; "mrs" has a DF = 3415/8388) The column **mapped_to** indicates the "re-corrected" form of the word, and the two **_adj** columns the associated DF values of the mapped_to word.

- **LFID_WordIDs**
    - 275,905 lyrics in bag-of-words format.
    - Column1 is the LyricFind ID (LyricID).
    - Column2 contains a vector representation of all the word instances (including repeated words) in the lyric. Numbers refer to the "wordID" index location in LF_SUBTLEX_merged_20150719.

- **LFID_cross_reference_columns**
    - Some of the 275,905 lyrics were recorded by more than one artist or appear on more than one album, thus giving them a unique LyricID (per the convention of LyricFind).
    - Column1 is the original LyricID.
    - Column2 is the LyricID which is a member of the set of 275,905 distinct lyrics.

- **metadata_plus_LNS_360919_lyrics_20150707**
    - Metadata comprises song title, artist name, album title, and the associated LyricFind IDs for each.
    - Other fields:
        - unique_words: the number of unique words in the lyric.
        - total_words: the total number of words (including repeated words).
        - IDF_trimean: the trimean of the inverse document frequencies of the unique words in this lyric.
        - lyric_LNS: the Lexical Novelty Score for the lyric.

- **billboard_top_100_all_time_songs**
    - http://www.billboard.com/articles/list/2155531/the-hot-100-all-time-top-songs
    - In a handful of cases, the song could be considered a "duet" between two artists; ArtistID_1 and Artist ID_2 provide the associated LyricFind artist IDs.
    - LNS is the lyric-level lexical novelty score.
    - IDF_trimean: the trimean of the inverse document frequencies of the unique words in this lyric.
    - *Note*: lyrics for 5 songs were not present in the LyricFind corpus, and have "NaN" for their LyricID.

- **billboard_top_100_all_time_artists**
    - http://www.billboard.com/articles/columns/chart-beat/5557800/hot-100-55th-anniversary-by-the-numbers-top-100-artists-most-no
    - artist_ID: the LyricFind artist ID.
    - artist_LNS_TM: the trimean of the lexical novelty scores for lyrics uniquely associated with this artist.
    - unique_N: the number of lyrics uniquely associated with this artist.
    - *Note*: Too few lyrics in the LyricFind corpus were available for Bryan Adams and Captain & Tennille.