CSCI 544: Applied Natural Language Processing

Course Project Proposals

1. **Project Name**
   Collection and Classification of Lyrics Information

2. **Project Inspiration**
   Music lives forever. Along with audio, lyrics convey the spirits of artists, in a way that they can better reflect the unique style and characteristics of the author. Imagine this, after being exposed to a large set of lyrics examples, a machine can learn to write lyrics just like John Lennon. It allows us to reproduce or even revive the spirits of the artists we love. Also, if trained appropriately, it is possible for the machine to automatically generate lyrics for different sections of a song, e.g. verse, chorus, hook, or even an entire song of a specific genre. This can provide inspiration for human song writers.

   After some research, we found that most of the existing projects similar to ours were based on bag-of-words lyrics. Very intuitively, the structure of a song is lost if the lyrics are scrambled. In this project, we introduce an intact lyrics corpus.

   To accomplish such tasks, we need a dataset that includes the original lyrics of all the songs, and labels that classify the songs, such as author, genre, and section, etc. Apart from generation/synthesis tasks, this dataset is suitable for classification tasks as well if incorporated with other labels. For example, we can also collect release years, genders of the authors such that gender classification and release-time interval classification are possible. These information can be further applied to similarity detection or recommendation systems.

3. **Data Collection and Labeling**
   In this project, we need:
   - Lyrics of the songs in their original forms (not bag-of-word)
   - Genres
   - Release years
   - Gender of the authors
   - Song structures

   Existing lyrics datasets, such as MusiXmatch (https://labrosa.ee.columbia.edu/millionsong/musixmatch) are mostly outdated and do not provide lyrics of orginal forms for analysis. Therefore, we will have to retrieve lyrics of more recent songs from some websites (e.g. https://genius.com/) using a web crawler, and then integrate the lyrics with existing datasets. Since lyrics and all their labels

usually cannot be obtained from one single source, our collection process will be in three steps (highly simplified):

1. Obtain names, authors, release dates of songs of various genres. We have two approaches in mind:
    a. Use a web crawler to scrape some famous music content providers, such as Amazon Music, Apple Music, and Spotify, etc.
    b. Use iTunes API (link below) to request metadata of music media on itunes, and extract information including genres, artists, composer, album, release dates.
2. Use a web crawler to retrieve lyrics in original forms and song structure information from the site: https://genius.com.
3. Use a web crawler to catch Personal Pronouns of artists from Wikipedia, and these information to judge their genders. E.g., “he” indicates Male. “She” indicates Female. We will not label gender if such information cannot be obtained.

4. **Targeted Dataset Size**

Since our method does not require hand-annotation, we expect our database size to be reasonably larger. Technically, we can obtain all lyrics and labels as long as they can be scraped from the web. However, we need to put more effort into data cleaning since all our data is collected using a web crawler. This might be the biggest constraint on the size. We expect to collect a dataset of 3,000 data points.

5. **Team Members**

- Su, Chang: csu272@usc.edu
- Chen, Tao: taochen@usc.edu
- Yang, Mu: yangmu@usc.edu
- Yang, Zhe: zheyang@usc.edu

6. **Additional Information**

iTunes API:
https://developer.apple.com/library/archive/documentation/AudioVideo/Conceptual/iTuneSearchAPI/index.html#//apple_ref/doc/uid/TP40017632-CH3-SW1