Classification of Korean Lyrics by Genre, Artist, and Release Date

20140507_JunSeonyoung
20170149_JaewooKim
20150860_TaeyoonKim
20150608_Lee,JunHyeong

1 Problem Description

From given lyrics of a Korean song, estimate information about the artist and song, such as age/gender/genre etc.

2 Data

The data will be acquired through web crawling. We will crawl websites that provide Korean lyrics along with the name of the artist. Since such websites do not provide details of the artist, we will separately search the name of the artists to find their personal information such as birthdate and gender. This will result in data in (lyrics, artist's name, gender, age, genre, release date) format.

3 Methodology/Algorithm

We will use Korean lyric corpus for our project. The reason we're doing so is to think about how algorithms we've learnt in class can be used in other languages such as Korean. We think basic methodologies can be applied similarly, but there might be some differences due to the differences in structure of the language.

We will first preprocess given lyrics by tokenizing and lemmatizing the words. We will use Korean natural language processing libraries such as konlpy(reference) or soynlp(reference). We will extract important keywords from each song. After that, we can use Google's archive of trained word2vec key-value dictionaries to vectorize given words so we can analyze them in ~300 dimensional vector space (reference). From the pool of lyrics we obtain, we will extract a code dictionary using k-means clustering to get k vectors that represent the corpus.

Each word vector from a lyric will be compared to those representative vectors, and it will be mapped to the closest vector in Euclidean distance. Using this methodology, we can generate a histogram for each lyric with k entries. We will train an SVM model to classify these histograms for each feature (age/gender of the artist, genre, release date etc). We will analyze more fields that we have not thought of as we study the data, if necessary.

4 Related Work

Predicting genre of a song with word2vec and gradient boost classifier

word2vec 을 이용한 거리 기반의 음악 가사 클러스터링 기법.

5 Evaluation Plan

We will evaluate the results in terms of percentage of correctly classified lyrics. As a result of SVM classification, each lyric will be classified for various fields such as genre, artist, release date etc. We could then evaluate the correctness of the classifications.