Human Language Technologies Assignment 1 Report

Daniel Conroy

Introduction

In this report, I will compare the styles of language used in two different music genres, pop and hip hop (also known as rap), in the 1980s. I choose to restrict my comparison to a single decade as the styles of language used in music genres may significantly change across time and I wanted to compare contemporaneous songs. I will also mention how some of the techniques of comparison I use or propose could be used for tasks in the domain of computer science, specifically machine learning tasks such as classifying the verse forms and the metres of poems or lyrics and classifying genres of songs based on their lyrics in text format, phonemic transcriptions of their lyrics and the audio format of the songs themselves.

I selected seven '80s pop songs and seven '80s hip hop songs to compare. I compare them based on their lyrics in text format. I performed natural language processing (NLP) tasks on their lyrics in text format using Python and Natural Language Toolkit (NLTK) to aid in comparing them. Each song's lyrics are stored in a separate text file and are processed individually. The format of each song's lyrics is akin to the format of poems. Each song's lyrics spread across multiple lines and contain punctuation.

I programmatically read the lines of stored song lyrics using the following method.

```
def get_song_lines(song_file_name):
    with open(song_file_name, encoding='utf8') as song_file:
        return song_file.readlines()
```

Preprocessing

I preprocess song lyrics as follows. First, I use the following method to remove all punctuation marks except for spaces between words and hyphens in words.

```
def remove_unneeded_chars(string):
    return ''.join([c for c in string if c.isalnum() or c == ' ' or c == '-'])
```

Then I use NLTK to tokenize song lyrics. Then I convert all uppercase letters to lowercase. I do this after tokenizing song lyrics since NLTK can sometimes use the information of whether a letter is uppercase or lowercase to help tokenize texts[1].

This preprocessing supports later NLP tasks by ensuring strings like "Girls", "girls" and "girls," (note the comma) don't get counted as distinct words. This also applies to strings like "girl's" and "girls", even though they have different meanings. But I decided this didn't matter since "girl's" and "girls" sound the same and that's what's important when analyzing language style in songs.

Comparing Songs by their Word Counts and Lexical Diversities

Counting the number of tokens NLTK computed for a song results in a word count for that song.

Word Counts for Songs

Songs	1	2	3	4	5	6	7	Average
Pop	289	467	256	240	321	297	241	302
Hip Hop	684	1063	425	674	342	3050	1220	1065

[1] https://www.nltk.org/_modules/nltk/tokenize/punkt.html

The hip hop songs have a far higher average numbers of words than the pop songs. Only pop song 2 (Billie Jean, Michael Jackson) has a higher word count (467) than any of the hip hop songs, having more words than hip hop songs 3 (It's Tricky, Run-D.M.C., 425 words) and 5 (Push it, Salt-n-Pepa, 342 words). Hip hop song 6 (Rapper's Delight, The Sugarhill Gang) is an outlier with 3050 words, but even when this song is excluded, the average word count for hip hop songs is 735, which is still far higher than the average word count for pop songs.

So, a guess that a song with a high word count is a hip hop song rather than a pop song would be highly accurate. However, such a guess may not be highly accurate when comparing hip hop songs to songs of another genre that isn't pop.

The lexical diversity of a song is the number of unique words in the song divided by the number of words in the song.

Lexical Diversities of Songs

Songs	1	2	3	4	5	6	7	Average
Pop	0.391	0.278	0.23	0.392	0.349	0.357	0.311	0.33
Hip Hop	0.465	0.347	0.376	0.374	0.316	0.227	0.426	0.362

The hip hop songs have a slightly higher average lexical diversity than the pop songs, but the distributions of lexical diversities for the hip hop songs and the pop songs have a lot of overlap.

So, in contrast to the word count of a song, the lexical diversity of a song is not a highly accurate predictor of whether the genre of the song is hip hop or pop.

Comparing Songs by their Frequent Words

The top 10 most frequent words in both the hip hop songs and the pop songs convey little to no more information about the topics of the songs than the songs' titles.

After using NLTK to filter out stop words from the list of tokens for each song, the top 10 most frequent non-stop words can be computed for each song. These are much more informative for the pop songs but not for the hip hop songs. For the pop songs, they often convey much more information about the topic and emotions of the song. Here are some examples to show this. Only the last song in this table is a hip hop song.

Top 10 Most Frequent Words

Songs	Including Stop Words	Excluding Stop Words
True Colors, Cyndi Lauper	you, true, colors, i, your, and, be, see, dont, to	true, colors, see, dont, show, shining, thats, love, afraid, let
Papa Don't Preach, Madonna	im, dont, papa, preach, i, my, baby, in, but, me	im, dont, papa, preach, baby, trouble, deep, ive, losing, sleep
Billie Jean, Michael Jackson	the, my, is, not, she, i, one, am, billie, jean	one, billie, jean, lover, son, kid, says, floor, round, told
Eternal Flame, The Bangles	you, do, your, feel, the, i, me, my, this, close	feel, close, eyes, give, hand, dreaming, burning, eternal, flame, darling
The Symphony, Marley Marl	the, i, a, you, and, to, it, me, my, that	like, im, cause, dont, yo, mic, got, rap, ya, thats

The top 10 most frequent non-stop words in a song could be a highly accurate predictor of whether the genre of the song is hip hop or pop. Some machine learning approaches and algorithms could be used to detect a noticeable theme, topic or sentiment among these words, which would indicate the genre of the song is pop, not hip hop.

Using NLTK to perform the additional NLP task of stemming the non-stop words of a song ensures the counts for words with the same stem are combined. However, after stemming the non-stop words of songs using either the Lancaster or Porter stemmers, the top 10 most frequent stems of non-stop words in both the hip hop and pop songs convey little to no more information about the songs than the top 10 most frequent non-stop words.

Comparing the Number of Rhymes in Songs

I use the following heuristic to count the number of rhymes in a song. I only count rhymes of the last words of adjacent lines in a song. A pair of adjacent lines in a song are considered rhyming lines if:

1. their last words are alphanumeric and at least 2 characters long,
2. the last two characters of their last words match, and
3. their last words are not the same.

Using this heuristic, the following two lines of "Push It" by Salt-N-Pepa don't count as rhyming lines. Repeated lines are common in songs, but I don't consider them to be rhyming lines.

```
Ooh, baby, baby, baby, baby
Ooh, baby, baby, ba-baby, baby
```

But the following pairs of lines of "It's Tricky" by Run-D.M.C. count as rhyming lines.

```
I'm not braggin', people naggin' cos they think I'm a star
Always tearin' what im wearin', I think they're goin' too far
...
They even bother my poor father cos he's down with me
It's Tricky to rock a rhyme, to rock a rhyme that's right on time
```

This heuristic has the flaw of sometimes considering non-rhyming words (such as me and time) as rhyming, but this happens rarely.

I use the following methods to count the number of rhymes in songs.

```
N = 2
def adj_lines_rhyme(song_lines, first):
    return (len(song_lines[first]) >= N
            and len(song_lines[first + 1]) >= N
            and song_lines[first][-N:].isalnum()
            and song_lines[first][-N:] == song_lines[first + 1][-N:]
            and song_lines[first][song_lines[first].rfind(' ') + 1:]
                != song_lines[first + 1][song_lines[first + 1].rfind(' ') + 1:])
```

```
def num_rhyming_lines_in_song(song_lines):
    count = 0
    for i in range(len(song_lines) - 1):
        if adj_lines_rhyme(song_lines, i):
            count += 1
    return count
```

Number of Rhymes in Songs

Songs	1	2	3	4	5	6	7	Average
Pop	0	3	6	9	8	8	0	5
Hip Hop	15	19	8	18	1	45	19	18

The hip hop songs have a far higher average number of rhymes than the pop songs. So, the number of rhymes in a song is a highly accurate predictor of whether the genre of the song is hip hop or pop, with a higher number of rhymes indicating the genre of the song is much more likely to be hip hop rather than pop (or perhaps any other music genre).

Flaws in the Approach to Counting Rhymes

The approach I used to detect rhymes in songs fails to detect some classes of rhymes.

Some rhyming lines have last words whose last 2 characters don't match. For example, the words son and one in the following lines of "Billie Jean" by Michael Jackson.

```
Billie Jean is not my lover
She's just a girl who claims that I am the one
But the kid is not my son
```

Rhymes like this could be detected using a phonemic transcription of song lyrics. Detecting rhymes using a phonemic transcription of a song's lyrics also avoids the problem of sometimes considering non-rhyming words (such as me and time) as rhyming.

However, singers sometimes mispronounce words to force a rhyme. For example, in "Fuck tha Police" by N.W.A., the singer mispronounces narcotics as narcotucts to make it rhyme with product.

```
Searchin' my car, lookin' for the product
Thinkin' every nigga is sellin' narcotics
```

These problems can be avoided by detecting phonemes in the audio format of the songs. Machine learning could be used to do this.

Some rhymes don't sound exactly the same but have similar patterns of phonemes. For example, the rhymes in the following lines of "The Symphony" by Marley Marl.

```
So just acknowledge the way that I kicked it
'Cause if rap was a house, you'd be evicted
And dismissed from the microphone
Chokin' on a bone, 'cause daddy's home
```

Such rhymes could be detected using pattern recognition among the phonemes of the song's lyrics.

In some songs, adjacent lines don't rhyme, but pairs of lines separated by one line in between them rhyme. This happens in the first four lines of "Africa" by Toto.

```
I hear the drums echoing tonight
But she hears only whispers of some quiet conversation
She's coming in, 12:30 flight
The moonlit wings reflect the stars that guide me towards salvation
```

In other songs, words within a line of the song rhyme. This happens in “Rapper’s Delight” by The Sugarhill Gang.

```
I said the age of one, my life begun
At the age of two i was doin’ the do
At the age of three, it was you and me
```

A more complex algorithm than the one I used would be needed to detect rhymes such as those in these two examples.

An algorithm that could detect patterns in both the structures of song lyrics or poems and phonemic transcriptions of them could be used to classify their verse forms and metres.