Report: Song Lyric Data Through the Wire

Letitia Lew
M.S. Computer Science, Stanford University

Background

I'm a big fan of hip hop music and the free-flowing language that rap artists use to express themselves. My favourite artists include Drake, Eminem, Nicki Minaj, Childish Gambino and Ludacris. Several of my good friends, who have grown up with deep roots in east coast rap, often joke that rappers mostly 'spit' about designer cars, alcohol, jewelry and women, but other conventional wisdom suggests that most songs universally talk about the same things (such as love, courtship, regret, having a good time, etc). Thus I became curious about whether there was statistical evidence that rap songs actually discussed significantly different topics than the broader corpus of music.

Hypothesis

The most frequently used words in English-language hip hop and rap music talk significantly more about cars, money, and insulting people than any other genre of music.

Dataset

One of the most prominent music-related datasets is the Million Song Dataset[1] (MSD), which provides metadata on a million tracks. An auxiliary "musiXmatch" dataset[2] provides the top word frequencies in about 770,000 songs (the words are stemmed, so "causes" and "causing" all map to "caus"). I could not find a dataset to associate songs with genres, so instead I obtained 'artist_term.db'[3], a SQLite database mapping artists to their genres.

Procedure

My original stoplist contained pronouns, conjunctions and prepositions, as well as common words like "most", "other", "did". I added more verbs and words that were common and meaningless (e.g. "gonna", "let", "get", "want", "time", "need"), and similar words in other languages, e.g. "ich", "que", "un", "da". I made sure that these stopwords would not affect my results by running a test sample of 5000 tracks looking for the subset of genres listed below. If a common word appeared in the top 20 of almost all these genres, then I removed it because it didn't tell us anything interesting about the differences between genres. (See "mapper.py" for my final stoplist.)

I had to clean and integrate the data from the three different datasets mentioned above:

Track -> Lyrics	from musiXmatch dataset
Track -> Artist	from Million Song Dataset

Artist -> Genres   from “artist_term.db”, an inverse index of the MSD

My Map function traversed these databases to obtain the desired relationship between Track -> Genres, then emitted the key-value pair:

```
foreach genre associated with Track:
    emit (genre, <list of word frequencies in track>)
```

My Reduce function then aggregated the word counts for each genre and printed out the top 50 words of each genre.

Results

Top 24 most common words for Genres of Interest		
Pop	**Blues**	**Heavy metal**
love	love	love
babi	babi	away
day	day	die
away	heart	day
caus	night	world
heart	away	eye
look	tell	live
thing	caus	night
tell	well	babi
night	look	look
think	thing	heart
girl	littl	mind
world	girl	caus
keep	think	thing
eye	good	think
tri	keep	onli
well	said	tri
onli	world	tell
good	eye	dream
live	long	lie
little	tri	fall
die	onli	turn
said	around	noth
around	live	end

Motown	**Hip hop**	**Gangster rap**
love	love	nigga
babi	babi	shit
girl	caus	caus
caus	day	fuck
tell	girl	bitch
thing	tell	yo
heart	look	love
day	die	em
good	away	niggaz
ooh	thing	keep
away	think	babi
night	heart	money
look	keep	wit
hey	night	ass
keep	world	gotta
think	tri	hit
little	live	yall
world	nigga	girl
well	eye	real
tri	good	die
onli	yo	tell
long	onli	hoe
hold	mind	day
said	said	look

My results found that mainstream genres of music such as Pop, Blues, even Heavy metal and Hip hop tended to have very similar top words such as “love”, “babi”, “girl” and “heart”, which suggested that the subject matter was predominantly love and courtship.

After digging deeper into more specific genres of hip hop, we notice that in Gangster rap and West coast rap, more physical and active themes dominate, such as “fuck”, “hit”, “rock” and “hard”. Compare this with the prevalence of “think” and “dream” in the other genres. But because of the allure of “money” and “hoes”, there is an emphasis in the hip hop community on authenticity, “keep”ing it “real”. On the subject of physicality, it is also interesting to note that Motown songs talked about movement a lot with words such as “dance”, “shake”, “hold” (from the 26th to 50th words).

Regarding the statistical validity of these results, it must be noted that the narrower genres of Gangsta, West coast hip hop, etc contain fewer songs than Pop, Classic rock or Hip hop. In future I will note the number of tracks in that genre, but as a rough guide the most common word in the narrower genres occurred 15,000 times vs 90,000 of the most common word in the broader genres.

Another issue is the degree to which the MSD corpus of songs are representative of all music. All we know is that they are "contemporary popular music tracks"[1], but the selection will certainly affect our validity.

Conclusions

We have concluded that hardcore genres of Rap and Hip hop discuss significantly more materialistic and physical themes than other genres of music. The abundance of swear words suggest that rap artists do demean people more in their songs. However, the results of our limited study do not indicate any preference for the material topics of cars, jewelry and clothes, possibly because our current methodology does not count the specific mentions of these items in songs. For example, "Chandon" and "Patron" both refer to alcohol, and "whip" and "Lambo" both reference cars but these words are not aggregated in our results.

For future work, I would be interested in slicing the hip hop and rap songs in the dataset by decade and seeing whether the content of the songs has changed over the years. Another possible hypothesis is that hip hop songs were 'harder' back in the 80s when rap was about gaining respect, than they are today where they are influenced more by mainstream pop. I would also like to run a deeper search (filtering out more stopwords and getting more than the top 50 words) and compare several narrower genres against one another to see more entertaining results.

References

1. Million Song Dataset. http://labrosa.ee.columbia.edu/millionsong/
2. musiXmatch dataset: http://labrosa.ee.columbia.edu/millionsong/musixmatch
3. artist_term.db: http://www.ee.columbia.edu/~thierry/artist_term.db