Analyzing Rap Lyrics – Part 1 of 3: Creating a Corpus

Posted on January 17, 2016

My initial interest in rap music was spurred a while back by an online announcement I read about the winners of the 2014 Kantar Information is Beautiful Awards. Among the 2014 awards, the one that peaked my curiosity was Matthew Daniels "Rappers, Sorted by Size of Vocabulary," the gold winner in the "Data Visualization" category.

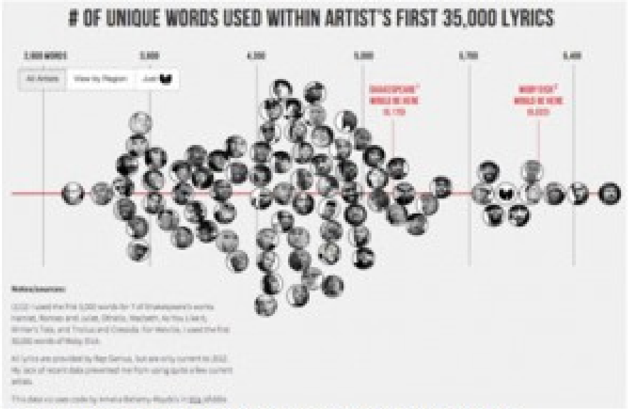


Figure 1. Daniel's Analysis of the Largest Vocabulary in Hip Hop

The visualization, which is shown in Figure 1, actually appeared in a more extensive article entitled, "The Largest Vocabulary in Hip Hop: Rappers, Ranked by the Number of Unique Words Used in their Lyric." The answer to the question about "who's is largest" depends on who's being compared, the specific lyrics being examined, and the meaning of terms "vocabulary" and "largest." Stated succinctly, Daniels used 85 well-known rappers, compared "each rapper's first 35K lyrics" (about 3-5 albums worth), defined vocabulary as the number of unique *words* (actually *tokens*), and created an award winning visualization to display the rankings. By the way, Aesop Rock was the winner (7,392), DMZ was the loser (3,214), and the vocabulary of a number of rappers compared favorably to Shakespeare (5170 unique words in 7 well-known plays) and Melville's *Moby Dick* (6122 unique words in the first 35000 words).

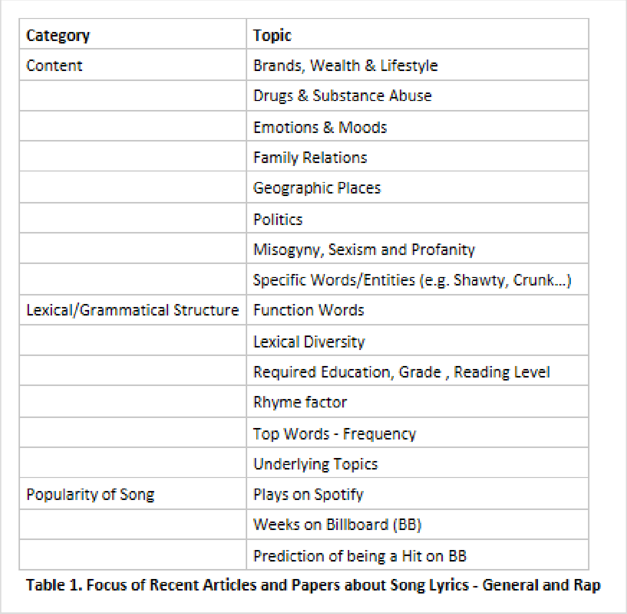
For those folks versed in text analysis, a few thoughts come to mind. First, "Damn, wish I'd thought of that!" Second, Daniels glossed over a number of difficult analytical questions, some of which he acknowledged and others he ignored. For instance, he used tokens not stems, so that things like "pimps and pimp" are treated as two unique words, not one. Similarly, he ignored the commonality of slang terms like "shorty" and "shawty," so that they were also treated as different when in fact that they often have the same meaning. Decisions like these can result in different rankings, although Aesop would win regardless of the choices. In all fairness, this wasn't an academic article and, as Daniels noted, even with the issues "it's still directionally interesting."

The offshoot of my fascination with Daniels' article is that I have spent the past 18 months or more periodically investigating rap lyrics, using various sorts of text analysis and text mining to document and visualize their evolution from earlier years to the present. Like most of the research work I've done in the area of automated text analysis and mining (working with online tweets, blogs, news articles, and product comments, reviews and recommendations), rap lyrics present a number of tough analytical hurdles whose solutions can inform text analysis done in other areas (and vice-versa). Similarly, if you can make a business case for analyzing rap lyrics, you can probably make a business case for doing analysis in other pop culture arenas. I'll talk about the business specifics later.

Text Analysis of Rap Lyrics

Just to reset. In this analysis, the discussion will follow the steps outlined in my last blog entry (excuse the hiatus). The outline includes: 1. Research Question(s); 2. Data Collection; 3. Data Processing & Cleansing; 4. Exploratory Data Analysis; 5. Modeling; 6. Communicating, Visualizing and Reporting; and 7. Packaging. Part 1 of this analysis will cover steps 1-2, Part 2 will be devoted to steps 3-4, and part 3 to 5-7. More specifically, the aim of Part 1 is to arrive at a corpus of lyrics covering a number of years and artists, that can be cleaned, prepared and initially explored in Part 2, and finally analyzed in Part 3.

Step 1 – Research Question(s)



Category	Topic
Content	Brands, Wealth & Lifestyle
	Drugs & Substance Abuse
	Emotions & Moods
	Family Relations
	Geographic Places
	Politics
	Misogyny, Sexism and Profanity
	Specific Words/Entities (e.g. Shawty, Crunk...)
Lexical/Grammatical Structure	Function Words
	Lexical Diversity
	Required Education, Grade , Reading Level
	Rhyme factor
	Top Words - Frequency
	Underlying Topics
Popularity of Song	Plays on Spotify
	Weeks on Billboard (BB)
	Prediction of being a Hit on BB

Table 1. Focus of Recent Articles and Papers about Song Lyrics - General and Rap

In addition to Daniels' article (and other articles he has written on related subjects), there have been a wide variety of articles and academic papers focusing on the text analysis of lyrics in general and rap and hip hop lyrics in particular. A detailed summary of a number of the more recent articles and papers is provided in Appendix 1. Table 1 provides an overview of the primary focus of those articles and papers.

Here, my primary focus is on using machine learning algorithms and data visualization to determine: Whether there have been major shifts in topic content and lyric style across time? If so, what have they been and have they been aligned with the shifts from one genre of rap to the next?

In this respect, this analysis follows in the footsteps of a variety of earlier studies including:

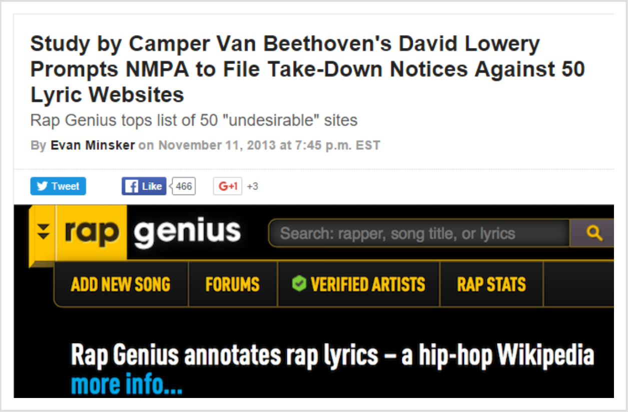
- Mauch et al.'s (2015) study of the audio evolution of popular music from 1960-2010
- Thompson's (2015) study of the lyric evolution of popular music from 1960-2010.
- Sterckx' (2013) analysis of the Million Song Dataset.
- Fell and Sporleder's (2014) lyrics-based analysis and classification of music.

In particular the analysis in this discussion will follow two strains: (1). topic mining which relies on an algorithm known as *Non-negative Matrix Factorization;* and (2) classification based on the underlying *style* of a lyric. At the end of this three part discussion (step 7), I'll reflect on the practical uses of the research being done on rap lyrics.

Step 2 – Data Collection

The foundation of any text analysis is the text corpus on which it is built. By text corpus I simply mean "a large and structured set of texts." In this case the set of texts is a collection of lyrics for a sample of rap song titles or tracks.

In the past, textual analysis based on larger collections of song lyrics were virtually impossible. These earlier studies relied on "close reading" and manual coding of the various words and phrases in the lyrics of interest – a mind numbing process at best. Today, there are enumerable Web sites providing very large collections of song lyrics of various genres. There are so many of these sites that there are special directories devoted to maintaining lists of and links to them (e.g. allrecordlabels.com). A number of earlier studies have relied on these sites for their data (see Appendix 1).



Study by Camper Van Beethoven's David Lowery Prompts NMPA to File Take-Down Notices Against 50 Lyric Websites

Rap Genius tops list of 50 "undesirable" sites

By **Evan Minsker** on November 11, 2013 at 7:45 p.m. EST

Tweet
Like 466
G+1 +3
rap genius
Search: rapper, song title, or lyrics
ADD NEW SONG
FORUMS
VERIFIED ARTISTS
RAP STATS
Rap Genius annotates rap lyrics – a hip-hop Wikipedia
more info...

Be wary of “crowds bearing gifts.” As you might expect, there is a lot of overlap in song titles from one site to the next, although no site has a complete nor perfect archive. At a number of the larger lyric sites, their collections are built through crowdsourcing.

Submissions from the laity are prone to typos, omissions and other errors. Like Wikipedia, not only do these sites rely on end users for submissions but also for editing. Apparently, when it comes to lyrics, their oversight is less than sterling. MacRae and Dixon (2012) did a thorough job of comparing the similarities among the same song lyrics at different sites. I’ll spare you the details. The punch line, they found an average accuracy of less than 40% which indicates substantial variability among the various copies of a lyric. Bottom line, before you start grabbing lyrics from one site or another, you need to get a sense of how much accuracy you require, and how “accurate” and complete the lyrics are at the site or sites you intend to use. And yes, you have to be careful about copyright and licensing issues, if you plan to distribute the lyrics.

Building a Sample of Rap Artist Names and Song Titles

A couple of caveats. I know there’s a difference between the terms “hip hop” and “rap,” but in this paper I’ll usually use the term “rap.” Similarly, I use the term “artist” to cover both individual musicians as well as groups.

In collecting a large sample of rap lyrics, one of the first challenges is creating the list of songs whose lyrics will be included in the sample. Practically speaking, you can’t go to Google or Bing and search for something like “Give me a list of rap songs from 1979 to the present.” For the uninitiated, 1979 is considered by many to be the year the first commercially successful rap song was released (Sugarhill Gang’s *Rapper’s Delight*). Obviously, you can physically make the search request, but what you’ll get back is a series of sites providing lists of their top songs for a given year (say 100), or maybe their top song for each year starting with 1979, or the top songs of all time. What you won’t find is a site providing a large, up-to-date laundry list of rap or hip-hop songs titles.

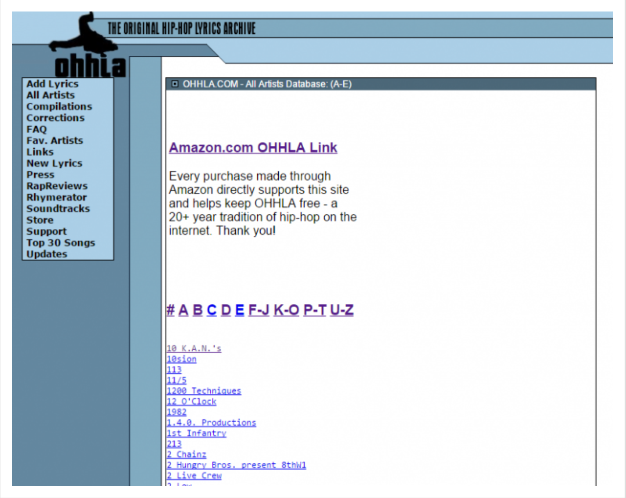
So, how do you build a list of songs? Currently, there seems to be three ways to do this:

1. Create your own list from scratch.
2. Use a preexisting list.
3. Use a preexisting database or dataset of songs which is based on a specified list.

Which one is preferable? It really depends on the question you're trying to answer and how much work you're willing to do. In my case, whatever approach is used needs to yield a sample of songs that is large enough and variable enough to adequately trace the shifts in lyrics from the early years to the present (assuming there have been shifts).

Create your own list artists and songs from scratch

While most lyric sites don't have an organized list of songs, they do have a list of artists. If you go to lyric archive site and (physically or programmatically) click on the name of an artist, up pops a list of their songs or albums which lead to the songs. With a substantial list of artists and a little programming effort, you can extract a sizable sample of song lyrics. Most of the larger and more popular lyric sites (e.g. genius.com, azlyrics.com, or metrolyrics.com) don't categorize artists and their songs by genre. While you can get a list of artists from these sites, there's no easy or direct way to ferret out the rap artists from the others.



OHHLA.COM - All Artists Database: (A-E)

Amazon.com OHHLA Link

Every purchase made through Amazon directly supports this site and helps keep OHHLA free - a 20+ year tradition of hip-hop on the internet. Thank you!

A B C D E F-J K-O P-T U-Z

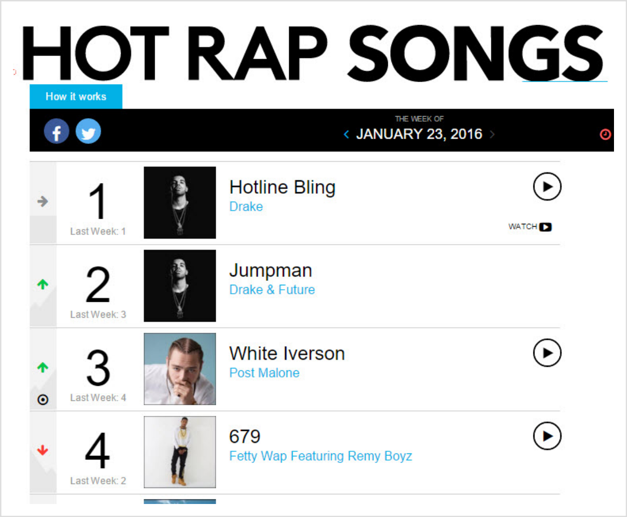
10 K.A.N.'s
10sion
113
11/5
1200 Techniques
12 O'Clock
1982
1.4.0. Productions
1st Infantry
213
2 Chainz
2 Hungry Bros. present 8thW1
2 Live Crew

Fortunately, there is a site that is solely devoted to hip hop and rap. It's OHHLA.com, "the Original Hip-Hop Lyrics Archive." OHHLA provides a list of artists that comes in alphabetical chunks that are divided into five separate pages (e.g. A-E, F-J, etc.). A little manual copying, cutting and pasting produces a list of over 3K hip hop artists. Another source is Wikipedia which has two lists that are pertinent – one for hip hop musicians and one for hip hop groups. Both of these lists are each presented on a single page that is divided into segments based on the first number or letter of the musician's or group's name (so two pages with 27 segments). Again, a little manual effort results in a list of 1671 hip hop or rap artists. Certainly, you could write a program to do this work, but I'd only resort to programming if the list was separated into a much larger number of segments (like an individual page for each of the 27 segments).

Now, with one of these list of rap and hip hop artists names in hand, you can use it to search one or more of the lyric sites for the list of songs associated with each of the names. Sometimes it's a simple list (like you find on Genius.com) and other times it's a list of albums leading to a list of songs (like Azlyrics.com). These lists can be quite long (sometimes in the 1000s) and can include songs where the artist is the sole singer/rapper, songs sung with other artists, songs where they are the primary and others are featured, songs where another artist is primary and he or she is featured … you get the general idea. Generally, you only want the songs where the artist is sole or primary whether others are featured or not. Unlike the creation of the list of artists, there's no way you can do this manually unless you have an inordinate amount of spare time on your hands (more about this later).

Use a Preexisting List of Artists and Songs

Most of the previous analysis of song lyrics do not create their lists of artists and songs from scratch. Instead, they use someone else's list. A lot of researchers have built their lists from various segments of Billboard (BB) Magazine's "Hot Rap" song charts.



HOT RAP SONGS

THE WEEK OF
JANUARY 23, 2016

1	Last Week: 1	Hotline Bling Drake
2	Last Week: 3	Jumpman Drake & Future
3	Last Week: 4	White Iverson Post Malone
4	Last Week: 2	679 Fetty Wap Featuring Remy Boyz

 Billboard has been charting "Hot Rap" Songs since March 11, 1989. Originally, the were reported every other week but toward the end of 1989 with the November 4th chart they began reporting weekly. Actually, there are a series of charts, but the three most often used for the analysis of Rap Songs are:

1. The 15 Hot Rap Songs released weekly — billboard.com/charts/rap-song/YYYY-MM-DD (this chart is published on Saturdays and requires the exact dates of those Saturdays in order to query the charts);
2. The Rap Song at the top of the weekly Hot Rap Songs chart — billboard.com/archive/charts/YYYY/rap-song
3. Top 100 Rap Songs released every year — billboard.com/charts/year-end/YYYY/hot-rap-songs starting in 2006.

About 6 or 7 months ago, I started looking in detail at the BB charts. My initial BB examination started with the Rap songs at the top of the weekly lists (#2). I did a pretty thorough text analysis of the 300 top ranked songs that appeared on the weekly lists (the data are provided in the Datasets Section for you analyzing leisure). However, the small sample size on which the analysis was built concerned me, especially when it came to comparing songs released in different eras. So, I decided to look at the Top 15 songs over the same time period (#1).

ROYAL SOCIETY
OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

Cite this article: Mauch M, MacCallum RM, Levy M, Leroi AM. 2015 The evolution of popular music: USA 1960–2010. *R. Soc. open sci.* 2: 150081.
http://dx.doi.org/10.1098/rsos.150081

Received: 17 February 2015
Accepted: 9 April 2015

Subject Category:
Biology (whole organism)

Subject Areas:
evolution/cognition/acoustics

Keywords:
popular music, diversity, stylistic revolutions, cultural evolution

Author for correspondence:
Matthias Mauch
e-mail: m.mauch@qmul.ac.uk

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsos.150081 or via http://rsos.royalsocietypublishing.org.

THE ROYAL SOCIETY
PUBLISHING

The evolution of popular music: USA 1960–2010

Matthias Mauch[1], Robert M. MacCallum[2], Mark Levy[3] and Armand M. Leroi[2]

[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK
[2]Division of Life Sciences, Imperial College London, London SW7 2AZ, UK
[3]Last.fm, 5-11 Lavingdon Street, London SE1 0NZ, UK

MM, 0000-0002-4352-6809

1. Summary

In modern societies, cultural change seems ceaseless. The flux of fashion is especially obvious for popular music. While much has been written about the origin and evolution of pop, most claims about its history are anecdotal rather than scientific in nature. To rectify this, we investigate the US Billboard Hot 100 between 1960 and 2010. Using music information retrieval and text-mining tools, we analyse the musical properties of approximately 17 000 recordings that appeared in the charts and demonstrate quantitative trends in their harmonic and timbral properties. We then use these properties to produce an audio-based classification of musical styles and study the evolution of musical diversity and disparity, testing, and rejecting, several classical theories of cultural change. Finally, we investigate whether pop musical evolution has been gradual or punctuated. We show that, although pop music has evolved continuously, it did so with particular rapidity during three stylistic 'revolutions' around 1964, 1983 and 1991. We conclude by discussing how our study points the way to a quantitative science of cultural change.

2. Introduction

The history of popular music has long been debated by philosophers, sociologists, journalists, bloggers and pop stars [1–7]. Their accounts, though rich in vivid musical lore and aesthetic judgements, lack what scientists want: rigorous tests of clear hypotheses based on quantitative data and statistics. Economics-minded social scientists studying the history of music have done better, but they are less interested in music than the means by which it is marketed [8–15]. The contrast with evolutionary biology—a historical science rich in quantitative data and models—is striking, the more so because cultural and organismic variety are both considered to be the result of

Clearly, looking at the top 15 BB rap songs charts results in a much larger sample than restricting it to the songs at the very top – about 1900 songs compared to approximately 300. I was just starting to analyze this larger sample when I came across Mauch et al.'s study of "The Evolution of Popular Music: USA 1960–2010" (2015). The data backbone of this study was the US Billboard Hot 100 charts for the designated time span. During this time period over 17,000 songs appeared on the weekly BB Hot 100 chart. While their study utilized "features extracted from audio rather than from scores," their dataset, which is publically available, provides a list of artist names and songs from a wide variety of genre including rap and hip hop that be used to seed a lyric study. In fact, this is exactly what James Thompson did in a follow-up study of the lyrics for the same sample of songs.

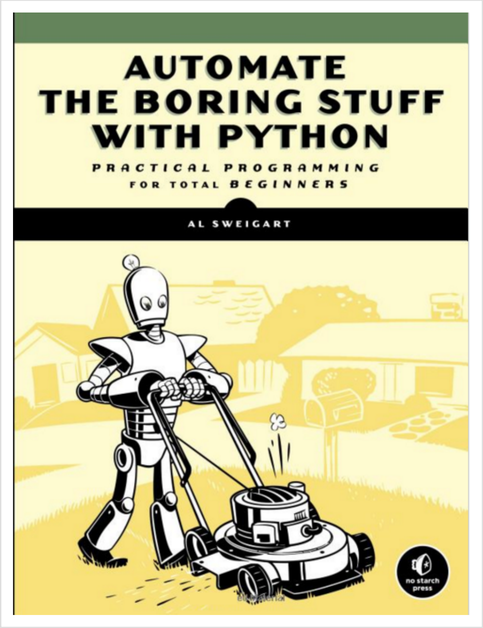
In Thompson's words: "Inspired by this amazing paper … I thought it would be interesting to see if something similar could be done with Pop lyrics. First, I need to get the lyrics for the songs. The authors … kindly published the data from their paper. It contains all the song titles and artists. Using this data and the API for ChartLyrics.com, I wrote a Python script to scrape the lyrics for each song. Unfortunately I wasn't able to find every song, but I found around 80%. This was also an automated process and I haven't thoroughly checked the output, so I would imagine there are a few incorrect songs in there." Thompson's analysis made it easy to follow in his footsteps since he also published his data.

So, inspired by both articles, my next trek down the data rabbit hole was to take their list of songs and metadata (primarily subgenre and years), reduce it to a set of rap and hip hop artists and songs, and, like Thompson, write a Python script to scrape the lyrics for each rap song. Based on their sets, this meant scraping about 2K rap titles. Like Thompson, I also wrote a Python script to scrape the lyrics. Unfortunately, while Thompson found lyrics for 80% of the original list, I only found 1185 of the rap and hip hop songs whose lyrics were available at ChartLyrics.com. A hit rate of 57%, which is pretty poor although 1185 is certainly larger than my original samples of 300 lyrics.

Creating the Corpus – Retrieving and Extracting (Rap) Song Lyrics

Whether you like it or not, if you're going to do text analysis of any reasonable sized corpus, there's virtually no way to avoid programming. If you're relying on Web sources for your data, it's close to impossible to avoid something called *web scraping*.

Web Scraping

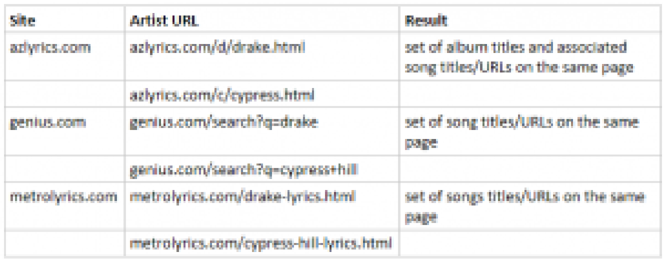

AUTOMATE THE BORING STUFF WITH PYTHON
PRACTICAL PROGRAMMING FOR TOTAL BEGINNERS
AL SWEIGART
no starch press

"Web scraping is the term for using a program to download and process content from the Web" (Sweigart, 2015). Generally, it includes the following tasks (Mitchell, 2015) :

- Retrieving HTML data (a page) from a domain name (URL)
- Parsing the retrieved data/page for target information
- Storing the target information
- Optionally moving to another page to repeat the process

For our purposes, the scraping steps could include one or more of the following:

1. From a chosen lyric site or sites, retrieving the HTML source page(s) associated with a particular artist's name from a lyric site. As noted, the vast majority of the time the retrieved page will contain either a list of album titles associated with the artist, or a list of song titles, or both. Quite often it's a good idea to store each page in a local file and work from this file, especially in the early stages of a project.
2. Next, each source page is parsed to retrieve the titles of the associated artist's albums or songs. Typically, the titles are listed as domain names or URLs that can be used to retrieve either the list of songs associated with the albums or the lyrics of the songs.
3. If the page contains a list of album URLs, each one of these will be retrieved and parsed to obtain the list of song URLs in the album.
4. With the list of song URLs in hand, the final phase is to use these URLs to retrieve the lyric pages. I usually store these pages locally (sans any images) so I don't have to keep visiting the original sites so I don't have to. These pages will form the basis of the corpus of the study.

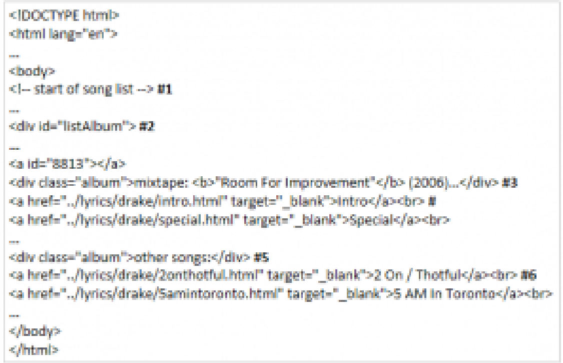


Site	Artist URL	Result
azlyrics.com	azlyrics.com/d/drake.html	set of album titles and associated song titles/URLs on the same page
	azlyrics.com/c/cypress.html	
genius.com	genius.com/search?q=drake	set of song titles/URLs on the same page
	genius.com/search?q=cypress+hill	
metrolyrics.com	metrolyrics.com/drake-lyrics.html	set of songs titles/URLs on the same page
	metrolyrics.com/cypress-hill-lyrics.html	

Table 2. Using an Artist Name to Build a List of Song Titles

As an example, suppose we use the Wikipedia's list of hip hop artists in order to retrieve the song titles and consequently the lyrics from one of the larger sites that it open to scraping activity. In order to write the scraping code to do this, you first have to construct the URL requests need to access the artists albums or songs. Unfortunately, there are no standards when it comes to the way the artists names are listed on the URLs (or anything other request for that matter). This is especially true for artists with two or more words in their names which is the vast majority. Table 2 illustrates some of the differences for the of the sites listed earlier.

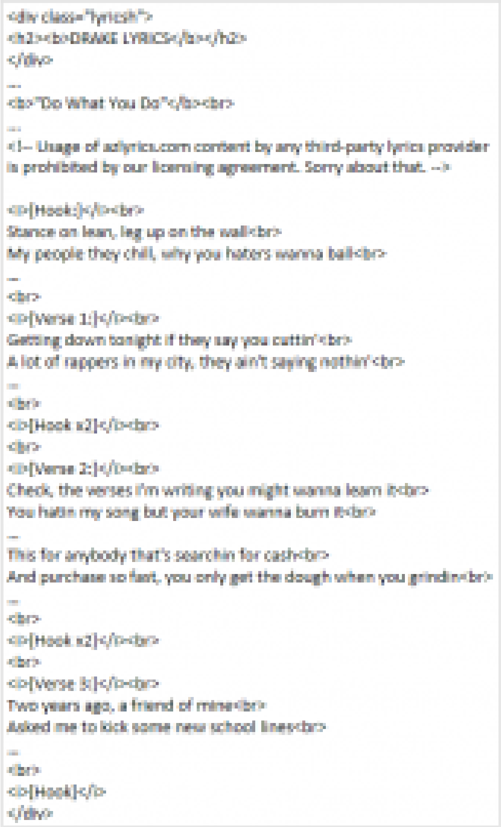
The result for each of the above requests is a single HTML Web page whose underlying source contains a list of songs associated with the specific artist. The source can be scraped to retrieve the list. Again, there are no standards from one site to the next. Obviously, you'll have to preview the underlying source for a few artists on a site in order to understand the html tags that can be used to locate and extract the song lists. Figure 2 shows what the relevant sections of underlying HTML source look like from requesting the page for the rap artist "Drake" at Azlyrics.com (i.e. azlyics.com/d/drake.html):



```
<!DOCTYPE html>
<html lang="en">
...
<body>
<!-- start of song list --> #1
...
<div id="listAlbum"> #2
...
<a id="8813"></a>
<div class="album">mixtape: <b>"Room For Improvement"</b> (2006)...</div> #3
<a href="../lyrics/drake/intro.html" target="_blank">Intro</a><br> #
<a href="../lyrics/drake/special.html" target="_blank">Special</a><br>
...
<div class="album">other songs:</div> #5
<a href="../lyrics/drake/2onthotful.html" target="_blank">2 On / Thotful</a><br> #6
<a href="../lyrics/drake/5amintoronto.html" target="_blank">5 AM In Toronto</a><br>
...
</body>
</html>
```

Figure 2. HTML Page from Azlyrics.com Listing Albums and Songs by Drake

- #1 – Announces the pertinent content with a comment;
- #2 – Declares the overall list of albums with a div(ision) "id" tag;
- #3 & #5 – For each album, it provides the title and year with a div(ision) class;
- #4 & 6 – Each album is followed by a list of song titles each denoted by an anchor reference (URL) that denotes the relative Web address where the song lyrics can be retrieved.



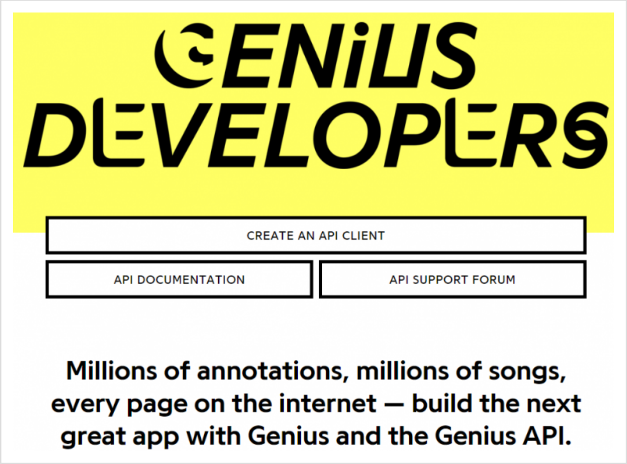
```
<div class="lyricsh">
<h2><b>DRAKE LYRICS</b></h2>
</div>
...
<b>"Do What You Do"</b><br>
...
<!-- Usage of azlyrics.com content by any third-party lyrics provider
is prohibited by our licensing agreement. Sorry about that. -->

<i>[Hook:]</i><br>
Stance on lean, leg up on the wall<br>
My people they chill, why you haters wanna ball<br>
...
<br>
<i>[Verse 1:]</i><br>
Getting down tonight if they say you cuttin'<br>
A lot of rappers in my city, they ain't saying nothin'<br>
...
<br>
<i>[Hook x2]</i><br>
<br>
<i>[Verse 2:]</i><br>
Check, the verses I'm writing you might wanna learn it<br>
You hatin my song but your wife wanna burn it<br>
...
This for anybody that's searchin for cash<br>
And purchase so fast, you only get the dough when you grindin<br>
...
<br>
<i>[Hook x2]</i><br>
<br>
<i>[Verse 3:]</i><br>
Two years ago, a friend of mine<br>
Asked me to kick some new school lines<br>
...
<br>
<i>[Hook]</i>
</div>
```

Figure 3. HTML Page from Azlyrics.com for the Lyrics to Drake's "Do What You Do" from "Room for Improvement"

For instance, if we follow the implicit anchor "…/lyrics/drake/dowhatyoudo.html" for the song "Do What You Do" at Azlyrics.com, this will take us to a page with the lyrics for that song. The underlying source HTML is shown in Figure 3. It's these underlying lyric pages that can form the corpus for a text analysis of rap lyrics.

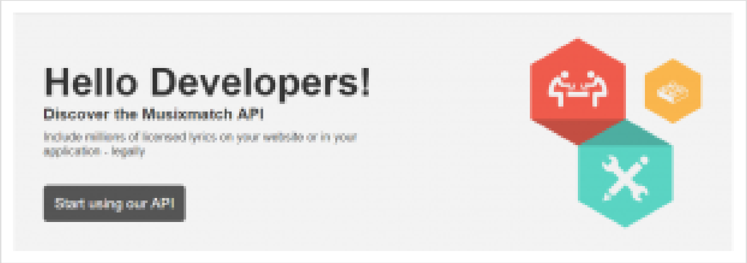
Regardless of the specific lyric site, once you understand the underlying HTML structure, it's a straightforward task to create a software program to retrieve and scrape the pages of interest and to extract and save the lyrics from the source to a local file. This can be done in a programming language like Python, using some combination of the Beautiful Soup module, regular expression matching (yes I understand all the debates about this), and custom code. See Sweigart 2015, Mitchell 2015, or Lawson 2015 for details.

A Word about APIs



Millions of annotations, millions of songs, every page on the internet — build the next great app with Genius and the Genius API.

Various lyric sites have created application programming interfaces (APIs) to support the integration of their platforms with 3rd party commercial applications. Two of the better known APIs for music lyrics are from Genius.com and Musixmatch.com. The Genius API primarily supports programmatic access for creating, managing and viewing of annotations, i.e. interpretations of various lyric phrases and lines made by the site's end users. While their API provides access to data and metadata about particular artists and lyrics, the application developer needs to supply Genius' proprietary numerical IDs for artists and songs in order to gain access to pertinent data about them. Because there is no consolidated list of the IDs available to the developer, it would be virtually impossible to use the API to develop a large corpus of lyrics.

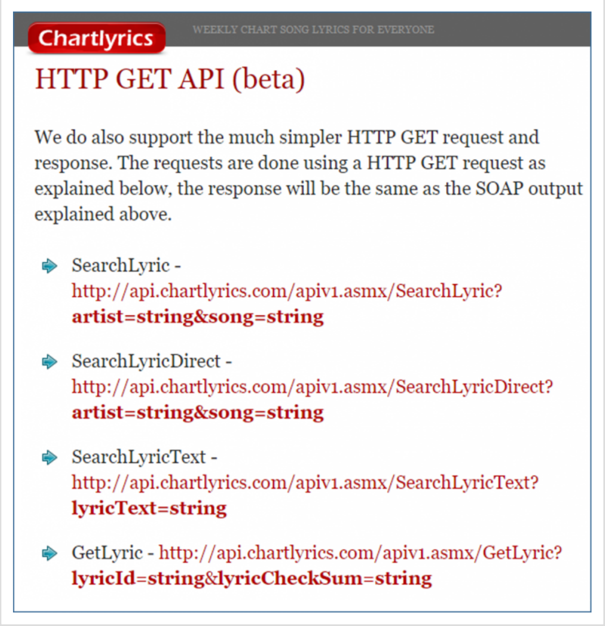


Hello Developers!

Discover the Musixmatch API

Include millions of licensed lyrics on your website or in your application - legally

Like Genius, Musixmatch's API is also aimed at commercial rather than research use. Musixmatch was started in 2010 and focuses on allowing "users to scan their music library and streaming playlist to retrieve lyrics via Musixmatch apps." This includes 3rd party apps as well. The site has a lyrics catalog of 7 million lyrics and information about 5 million artists. On the surface, it's ideal for creating a large sample of lyrics for research purposes. The problem is that access to their catalog costs a minimum of $25K, a bit prohibitive for the average researcher. The reason for the charge revolves around lyric copyright issues and licensing restrictions which Musixmatch has to pay in order *to display* the lyrics on their apps (the same is true for other sites — see Minsker (2013) for the problems Genius.com has encountered). To their credit, Musixmatch recognizes the value of this information for research purposes and has provided a special "bag-of-words" dataset available for non-commercial uses.



HTTP GET API (beta)

We do also support the much simpler HTTP GET request and response. The requests are done using a HTTP GET request as explained below, the response will be the same as the SOAP output explained above.

- SearchLyric - http://api.chartlyrics.com/apiv1.asmx/SearchLyric?**artist=string&song=string**
- SearchLyricDirect - http://api.chartlyrics.com/apiv1.asmx/SearchLyricDirect?**artist=string&song=string**
- SearchLyricText - http://api.chartlyrics.com/apiv1.asmx/SearchLyricText?**lyricText=string**
- GetLyric - http://api.chartlyrics.com/apiv1.asmx/GetLyric?**lyricId=string&lyricCheckSum=string**

The last API I'll mention is ChartLyrics.com — the one that Thompson used, and I subsequently used. As the name implies, their API is specifically aimed at supporting the retrieval of lyrics from their catalog of songs which includes all genres. The API has two key requirements. First, it requires the name of the artist and song title in their URL format — this is nothing new. Second, they have a 20 second governor on retrieval requests of all sorts. This is sort of new — a lot of API's have time or frequency restrictions. In Thompson's case, he used two 20 second delays for two separate steps (overall 40 seconds) before the song's lyrics could be retrieved. As he quipped, "This meant it took days to run and my mac kept going to sleep mid-process." Following Thompson's lead, when I used the Top 100 rap song set discussed earlier, I only had to retrieve approximately 3,000 rap songs versus his 17,000 "pop music" songs. Fortunately, it only took part of a day, plus I have a second computer that I use for this sort of task and, thanks to a hint from him, I avoided the "going to sleep" part.

Cheating – Copying a corpus from the Million Song Dataset

Million Song Dataset

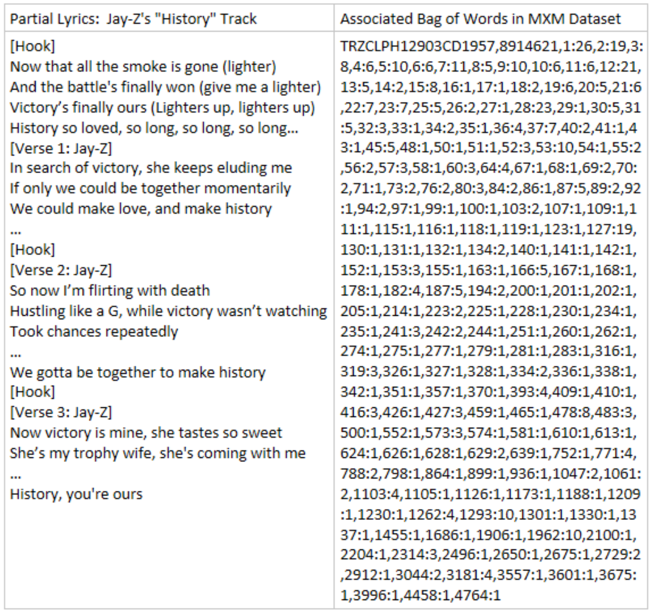
Home

The musiXmatch Dataset

I know all this programming sounds like fun, but you can avoid much of the initial pain thanks to Echo Nest (now part of Spotify), LabROSA (aka Columbia University's Laboratory for the Recognition and Organization of Speech and Audio) and Musixmatch.com. In 2011 Echo

Nest(echonest.com), in conjunction with LabROSA, released the Million Song Dataset (MSD). In their words, it's "a freely-available collection of audio features and metadata for a million contemporary popular music tracks." (Bertin-Mahieux et al. 2011). As the name implies the dataset includes audio features for 1M songs/files along with over metadata about 44K unique artists, 2K musicbrainz tags, 2M artist similarity relationships, and release dates for 515K tracks. However, no lyrics. This is where musiXmatch comes in. Somewhat later, Echo Nest in partnership with musiXmatch released the musicXmatch (MXM) dataset, a collection of lyrics for over 237K tracks. Again, in their words, "Through this dataset, the MSD links audio features, tags, artist similarity, … to lyrics" (see labrosa.ee.columbia.edu/millionsong/musixmatch).

Copyright issues prevent the partnership from distributing the full, original lyrics. As a consequence, the lyrics come in a "bag-of-words" format and are stemmed. So, technically speaking it's a bag-of-stems, more specifically a bag-of-5000-stems (in essence the root forms of the words in the lyrics). For example, this is what a bag-of-words looks like for a sample track:



Partial Lyrics: Jay-Z's "History" Track	Associated Bag of Words in MXM Dataset
[Hook] Now that all the smoke is gone (lighter) And the battle's finally won (give me a lighter) Victory's finally ours (Lighters up, lighters up) History so loved, so long, so long, so long... [Verse 1: Jay-Z] In search of victory, she keeps eluding me If only we could be together momentarily We could make love, and make history ... [Hook] [Verse 2: Jay-Z] So now I'm flirting with death Hustling like a G, while victory wasn't watching Took chances repeatedly ... We gotta be together to make history [Hook] [Verse 3: Jay-Z] Now victory is mine, she tastes so sweet She's my trophy wife, she's coming with me ... History, you're ours	TRZCLPH12903CD1957,8914621,1:26,2:19,3:8,4:6,5:10,6:6,7:11,8:5,9:10,10:6,11:6,12:21,13:5,14:2,15:8,16:1,17:1,18:2,19:6,20:5,21:6,22:7,23:7,25:5,26:2,27:1,28:23,29:1,30:5,31:5,32:3,33:1,34:2,35:1,36:4,37:7,40:2,41:1,43:1,45:5,48:1,50:1,51:1,52:3,53:10,54:1,55:2,56:2,57:3,58:1,60:3,64:4,67:1,68:1,69:2,70:2,71:1,73:2,76:2,80:3,84:2,86:1,87:5,89:2,92:1,94:2,97:1,99:1,100:1,103:2,107:1,109:1,111:1,115:1,116:1,118:1,119:1,123:1,127:19,130:1,131:1,132:1,134:2,140:1,141:1,142:1,152:1,153:3,155:1,163:1,166:5,167:1,168:1,178:1,182:4,187:5,194:2,200:1,201:1,202:1,205:1,214:1,223:2,225:1,228:1,230:1,234:1,235:1,241:3,242:2,244:1,251:1,260:1,262:1,274:1,275:1,277:1,279:1,281:1,283:1,316:1,319:3,326:1,327:1,328:1,334:2,336:1,338:1,342:1,351:1,357:1,370:1,393:4,409:1,410:1,416:3,426:1,427:3,459:1,465:1,478:8,483:3,500:1,552:1,573:3,574:1,581:1,610:1,613:1,624:1,626:1,628:1,629:2,639:1,752:1,771:4,788:2,798:1,864:1,899:1,936:1,1047:2,1061:2,1103:4,1105:1,1126:1,1173:1,1188:1,1209:1,1230:1,1262:4,1293:10,1301:1,1330:1,1337:1,1455:1,1686:1,1906:1,1962:10,2100:1,2204:1,2314:3,2496:1,2650:1,2675:1,2729:2,2912:1,3044:2,3181:4,3557:1,3601:1,3675:1,3996:1,4458:1,4764:1

Figure 4. Sample Bag-of-Words for Song Track from MXM Dataset

The first field value "TRZCLPH12903CD1957" is the MSD track id, while the second value "8914621" is the MXM track id. This is followed by a series of "<word idx>:<cnt>" pairs. For instance, in this track the word "1" appears 26 times ("1:26"), word 2 occurs 19 times ("2:19"), and so on until the last pair which is the 4832nd word which appears 1 time ("4764:1"). The translation of word IDs can be found determined from either the MXM training or testing data – word 1 is "I," word 2 is "the", …, word 4764 is "flirt" even though the word in the song is "flirting" (remember it's stems not words).

The 5000 words were chosen to represent the words that occur most frequently (culled to clear out the junk like foreign symbols, glued together words, etc.). Obviously, there are more than 5000 words in the entire collection of 237K lyrics. In fact according to the official count there are approximately 498K unique words for a total of over 55M occurrences. The 5000 words in the dataset account for close to 51M out of the 55M (over 90%). Of the 5000 words, only those that actually appear in a track are recorded in the dataset. This particular track has 195 out of the 5000 words. For the remainder of the 4805 words there no "<word idx>:0" pair. This is done to reduce the size of the set for storage and transmission purposes.

Using this BOW words certainly simplifies the task of building a corpus of lyrics of song tracks for analysis, eliminating the need to create a list of artists and songs and to retrieve and extract the lyrics for a very large sample of songs. Yet, there is still a bit of housecleaning to be done, and there are some major drawbacks when it comes to text analysis. From a housecleaning perspective, we need to determine which of the songs are rap songs, which of the rap songs have lyrics, what years the songs were released, and which of the tracks in the set are duplicates (of which there are a number). Once we address those issues, we'll have a "clean bag of stems". That's the good news, the bad news is that we'll only have the stems. This limits types of text analysis we can do, even simple things like "n-gram" analysis of frequently occurring sequential pairs or triplets of words. If you understand how to do the housecleaning, don't mind the restricted analysis, and have machine that can crunch data of this size without putting it in cardiac arrest, then it's a good place to start and certainly will result in a large corpus of lyrics.

[A parenthetic note about using someone else's corpus of rap lyrics. In 2011 Tahir Hemphill introduced his *Hip Hop Word Count*, "a searchable ethnographic database manually built from the lyrics of over 50,000 hip-hop songs from 1979 to present day." The massive database is searchable by date, artist, word, word complexity, locality, and a host of other metrics. Originally, my understanding was that the database was going to be publicly available. To my knowledge that hasn't happened, although he has used it for some special youth oriented projects. I wish it were publicly available, it would have saved me a substantial amount of work.]

Prelude to Part 2

In Part 2 of the discussion I'll delve into the details of preparing the corpus of rap lyrics for exploratory analysis, as well as carrying out this initial analysis. The primary datasets I'll use will rest primarily on the corpus I've extracted from the weekly BB Top 100 Rap Songs using the Chartlyrics' API.

Resources

References

Bertin-Mahieux, T. et al. "The Million Song Dataset." *International Society for Music Information Retrieval* (2011).

Fell, M. and Sporleder, C. "Lyrics-Based Analysis and Classification of Music." 25th *International Conference on Computational Linguistics* (2014).

Lawson, R. Web Scraping with Python Paperback. *Packt Publishing* (2015).

Macrae, R. and Dixon, S. "Ranking Lyrics for Online Search." *13th International Society for Music Information Retrieval Conference* (2012).

Mauch, M. The Evolution of Popular Music: USA 1960–2010. *Royal Society of Open Science* (2015).

Minsker, E. “Study by Camper Van Beethoven’s David Lowery Prompts NMPA to File Take-Down Notices Against 50 Lyric Websites: Rap Genius tops list of 50 ‘undesirable’ sites.” (2013).

Mitchell, R. Web Scraping with Python: Collecting Data from the Modern Web 1st Edition. O’Reilly Media (2015).

Sterckx, L. Topic Detection in a Million Songs. Dissertation, *Universiteit of Gent* (2013).

Sweigart, A. Automate the Boring Stuff with Python: Practical Programming for Total Beginners 1st Edition. *No Starch Press* (2015).

Thompson, J. “The Evolution of Popular Lyrics and a Tale of Two LDA’s.” (2015).

People

James Thompson

Matthew Daniels

Tahir Hemphill

Places (Virtual and Real)

Azlyrics.com

Echo Nest

Genius.com

LabROSA

Original Hip Hop Lyrics Archive (OHHLA)

Tools & Modules

Beautiful Soup

ChartLyrics API

Genius API

Musixmatch API

Musixmatch Million Song Dataset

Web Scraping

This entry was posted in **Uncategorized** by **daveking63@gmail.com**. Bookmark the **permalink [http://dataffiti.com/2016/01/17/analyzing-rap-lyrics-part-1-of-3-creating-a-corpus/]** .