

Scraping and Analysing Song Lyrics and Metadata

F.D. Brunet de Rochebrune
R.K. Dijkstra
J.E. Van der Lei

nice
fbe690
rda630
jli530

VU – Data Wrangling – Abraca-data
Wednesday 29 January 2020



Research Question

What are the characteristics of the genres, most popular songs and artists on www.genius.com? And how do they differ?

Genres:

Rap

R&B

Pop

Rock

Country

Data Acquisition

- Public API with Authentication
- Release Date
- Page Views
- Pyongs
- Hot or Not
- Verified / Meme Verified

Data Acquisition








- Public API with Authentication
- Release Date
- Page Views
- Pyongs
- Hot or Not
- Verified / Meme Verified
- Needs Song ID
- No Lyrics
- No Top Charts

Data Acquisition

- Public API with Authentication
- Release Date
- Page Views
- Pyongs
- Hot or Not
- Verified / Meme Verified
- Needs Song ID
- No Lyrics
- No Top Charts
- Scrape the Website


Genius.com Top Charts

SONGS / RAP / ALL TIME ▾

1		Rap God LYRICS	Eminem	👁 15.5M
2		HUMBLE. LYRICS	Kendrick Lamar	👁 10.4M
3		Man's Not Hot LYRICS	Big Shaq	👁 8.4M
4		Bad and Boujee LYRICS	Migos	👁 8.4M
5		God's Plan LYRICS	Drake	👁 8.2M
6		Bodak Yellow LYRICS	Cardi B	👁 7.7M
				

Inspecting the Webpage

1




Rap God LYRICS

Eminem

15.5M

2




HUMBLE. LYRICS

Kendrick Lamar

10.4M

3



Man's Not Hot LYRICS

Bia Sharq

8.4M

Elements

Console

Sources

Network

Performance

Memory

Application

Security

Audits

Filter

Hide data URLs

All

XHR

JS

CSS

Img

Media

Font

Doc

WS

Manifest

Other

2000 ms

4000 ms

6000 ms

8000 ms

10000 ms

12000 ms

14000 ms

16000 ms

18000 ms











20000 ms

22000 ms

24000 ms


26000 ms

28000 ms

Name	Status	Type	Initiator	Size	Time	Waterfall
 sessions.bugsnag.com	(blocked:other)	xhr	reactPageVendors.desktop-04d...	0 B	424 ms	
 ?verbose=1&version=1&lib=web&token=77967c52dc38186cc1a...	(blocked:other)	xhr	reactPageVendors.desktop-04d...	0 B	47 ms	
 ?data=eyJldmVudC6lCjtcF9wYWdlX3ZpZXc6lCJwcm9wZXJ0...	(blocked:other)	xhr	reactPageVendors.desktop-04d...	0 B	50 ms	
 ?verbose=1&version=2&lib=web&token=77967c52dc38186...d19...	(blocked:other)	xhr	reactPageVendors.desktop-04d...	0 B	58 ms	
 ?data=eyJldmVudC6lCjtcF9wYWdlX3ZpZXc6lCJwcm9wZXJ0aWVz...	(blocked:other)	xhr	reactPageVendors.desktop-04d...	0 B	56 ms	
 auth	200	xhr	reactPageVendors.desktop-04d...	420 B	773 ms	
 chart?time_period=all_time&chart_genre=all&page=1&per_page...	200	fetch	reactPageVendors.desktop-04d...	3.0 KB	1.08 s	
 metrics	(blocked:other)	fetch	reactPageVendors.desktop-04d...	0 B	5 ms	
 chart?time_period=all_time&chart_genre=rap&page=1&per_pag...	200	fetch	reactPageVendors.desktop-04d...	2.9 KB	771 ms	
 metrics	(blocked:other)	fetch	reactPageVendors.desktop-04d...	0 B	7 ms	


10 / 49 requests | 6.4 KB / 110 KB transferred | 27.4 KB / 1.5 MB resources | Finish: 25.40 s | DOMContentLoaded: 4.04 s | Load: 4.27 s

Huh, what's that?



Name	Status	Type	Initiator
<input type="checkbox"/> sessions.bugsnap.com	(blocked:oth...	xhr	reactPageVendors.desk...
<input type="checkbox"/> ?verbose=1&version=1&lib=web&token=77967c...	(blocked:oth...	xhr	reactPageVendors.desk...
<input type="checkbox"/> ?data=eyJldmVudCI6ICJtcF9wYWdlIX3ZpZXciLC...	(blocked:oth...	xhr	reactPageVendors.desk...
<input type="checkbox"/> ?verbose=1&version=2&lib=web&token=77967c...	(blocked:oth...	xhr	reactPageVendors.desk...
<input type="checkbox"/> ?data=eyJldmVudCI6ICJob21lOmxyYWQiLCJwc...	(blocked:oth...	xhr	reactPageVendors.desk...
<input type="checkbox"/> auth	200	xhr	reactPageVendors.desk...
<input type="checkbox"/> chart?time_period=all_time&chart_genre=all&pa...	200	fetch	reactPageVendors.desk...
<input type="checkbox"/> metrics	(blocked:oth...	fetch	reactPageVendors.desk...
<input type="checkbox"/> chart?time_period=all_time&chart_genre=rap&pa...	200	fetch	reactPageVendors.desk...
<input type="checkbox"/> metrics	(blocked:oth...	fetch	reactPageVendors.desk...

Huh, what's that?



Name	Status	Type	Initiator
<input type="checkbox"/> sessions.bugsnap.com	(blocked:oth...	xhr	reactPageVendors.desk...
<input type="checkbox"/> ?verbose=1&version=1&lib=web&token=77967c...	(blocked:oth...	xhr	reactPageVendors.desk...
<input type="checkbox"/> ?data=eyJldmVudCI6ICJtcF9wYWdlIX3ZpZXciLC...	(blocked:oth...	xhr	reactPageVendors.desk...
<input type="checkbox"/> ?verbose=1&version=2&lib=web&token=77967c...	(blocked:oth...	xhr	reactPageVendors.desk...
<input type="checkbox"/> ?data=eyJldmVudCI6ICJob21lOmxxvYWQiLCJwc...	(blocked:oth...	xhr	reactPageVendors.desk...
<input type="checkbox"/> auth	200	xhr	reactPageVendors.desk...
<input type="checkbox"/> chart?time_period=all_time&chart_genre=all&pa...	200	fetch	reactPageVendors.desk...
<input type="checkbox"/> metrics	(blocked:oth...	fetch	reactPageVendors.desk...
<input type="checkbox"/> chart?time_period=all_time&chart_genre=rap&pa...	200	fetch	reactPageVendors.desk...
<input type="checkbox"/> metrics	(blocked:oth...	fetch	reactPageVendors.desk...

It seems like an API...

https://genius.com/api/songs/chart?time_period=all_time&chart_genre=rap&page=1&per_page=10

```
{"meta":{"status":200},"response":{"chart_items":[{"_type":"chart_item","type":"song","item":{"_type":"song","annotation_count":111,"api_path":"/songs/235729","full_title":"Rap God by Eminem","header_image_thumbnail_url":"https://images.genius.com/245592efd1ce48ebbca6d832106ac04f.300x169x1.jpg","header_image_url":"https://images.genius.com/...
```

It seems like an API... but it's undocumented

`https://genius.com/api/songs/chart?time_period=all_time&chart_genre=rap&page=1&per_page=10`

- Undocumented API
- No API Credentials Needed

Let's have some fun

`https://genius.com/api/songs/chart?time_period=all_time&chart_genre=rap&page=1&per_page=100`

Yes.. but not quite yet

```
https://genius.com/api/songs/chart?time_period=all_time&chart_genre=rap&page=1&per_page=100
```

```
{"meta":{"status":422,"message":"Invalid per_page param.  
Must be between 1 and 50. You supplied:  
100"},"response":{"error":"invalid_per_page"}}
```

Data Acquisition, Merging

```
https://genius.com/api/songs/chart?time_period=all_time&chart_genre=rap&page=1&per_page=50
```

Data Acquisition, Merging

`https://genius.com/api/songs/chart?time_period=all_time&chart_genre=rap&page=1&per_page=50`

- Loop over it
- 150 Records per genre
- Country only has 84 records
- Song Chart Position
- Song Name
- Song Artist
- Song ID
- No Lyrics

Data Acquisition, Merging

Genius Public and Undocumented API

- Release Date
- Page Views
- Pyongs
- Hot or Not
- Verified / Meme Verified
- Chart Position
- Name
- Artist
- ID (on Genius)

Genius Scraping

- Beautiful Soup
- Scrape Lyrics Part Based on Song ID

Deezer API

- Search on title+artist, First Result
- URL Encode (`urllib` package)
- Transliterate Russian (`cyrtranslit` package)
- Explicit
- Duration
- BPM

Data Cleaning

- **Huge NLP Task!**
- Removing Punctuation, Brackets
 - Regex

[Chorus]

I tell her man's not hot

I tell her man's not hot

The girl told me, "Take off your jacket"

I said, "Babes, man's not hot" (Never hot)

I tell her man's not hot (Never hot)

I tell her man's not hot (Never hot)

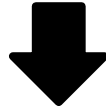
The girl told me, "Take off your jacket"

I said, "Babes, man's not hot" (Never hot)

Data Cleaning and Obtaining Additional Information

- Tokenizing words (TokTok Tokenizer)
 - Maintaining slang words
- Stemming of Words
- Removal of English Stopwords
 - On, a, the, an, to etc.
- Vader Sentiment Analysis
- Marking words as profane (better_profanity package)

This is an example of: "Tokenization"



[this, is, an, example, of, tokenization]

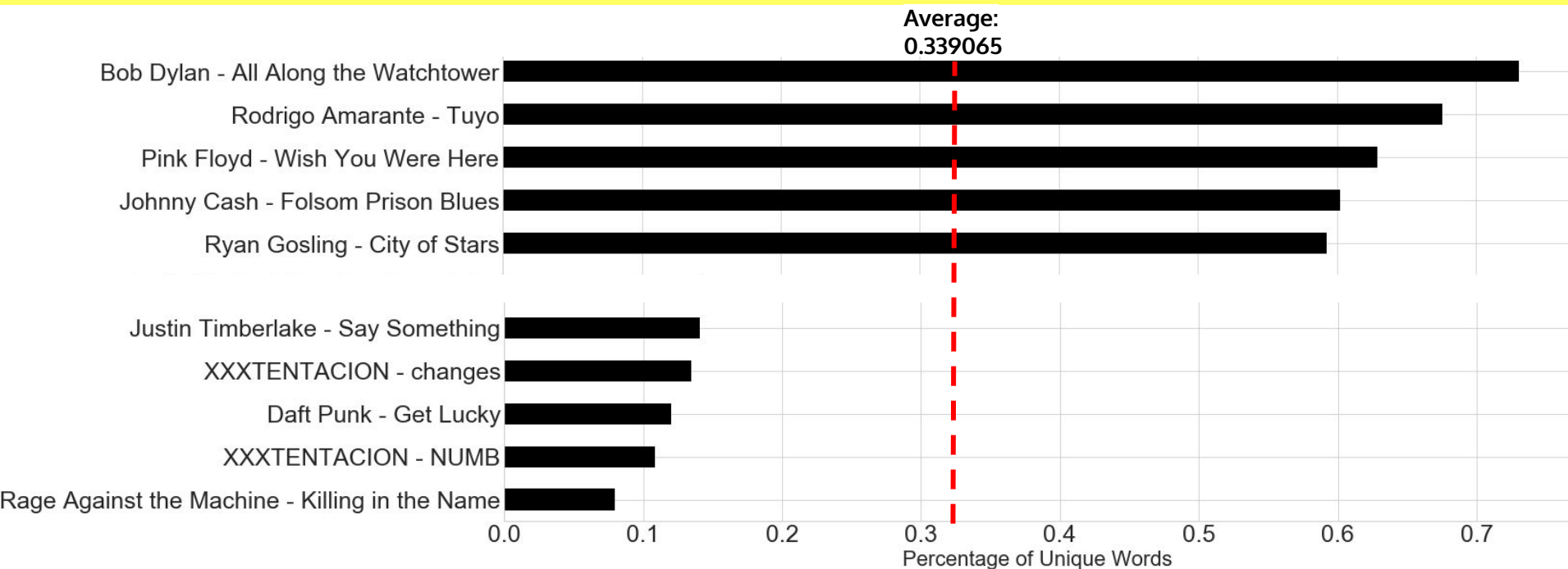
Stemming

adjustable → adjust
formality → formaliti
formaliti → formal



Analysis

Highest&Lowest Percentage of Unique Words



Song with Least Amount of Unique Words

Killing in the name of
Some of those that work forces, are the same that burn crosses
Some of those that work forces, are the same that burn crosses
Some of those that work forces, are the same that burn crosses
Some of those that work forces, are the same that burn crosses
Ugh
Killing in the name of
Killing in the name of
Now you do what they told ya
Now you do what they told ya
Now you do what they told ya
Now you do what they told ya
And now you do what they told ya
And now you do what they told ya
And now you do what they told ya
And now you do what they told ya
And now you do what they told ya
And now you do what they told ya
But now you do what they told ya
Well now you do what they told ya
Those who died are justified
For wearing the badge, they're the chosen whites
You justify those that died
By wearing the badge, they're the chosen whites
Those who died are justified
For wearing the badge, they're the chosen whites
You justify those that died
By wearing the badge, they're the chosen whites

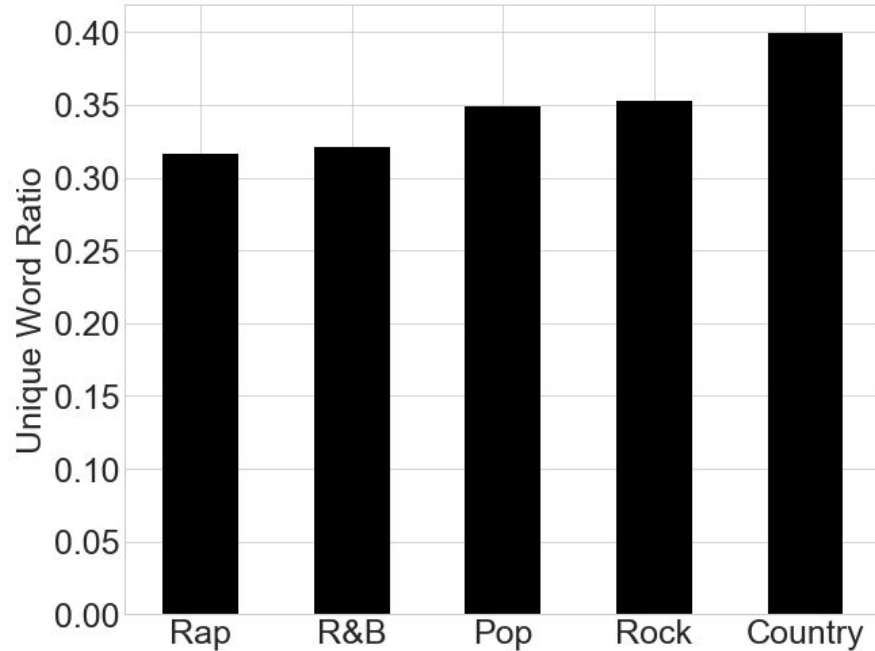
Some of those that work forces, are the same that burn crosses
 Some of those that work forces, are the same that burn crosses
 Some of those that work forces, are the same that burn crosses
 Some of those that work forces, are the same that burn crosses
 Ugh
 Killing in the name of
 Killing in the name of
 Now you do what they told ya
 Now you do what they told ya
 Now you do what they told ya
 Now you do what they told ya
 And now you do what they told ya
 Now you're under control And now you do what they told ya
 Now you're under control And now you do what they told ya
 Now you're under control And now you do what they told ya
 Now you're under control And now you do what they told ya
 Now you're under control And now you do what they told ya
 Now you're under control And now you do what they told ya
 Now you're under control And now you do what they told ya
 Those who died are justified
 For wearing the badge, they're the chosen whites
 You justify those that died
 By wearing the badge, they're the chosen whites
 Those who died are justified
 For wearing the badge, they're the chosen whites
 You justify those that died
 By wearing the badge, they're the chosen whites

[illegible]

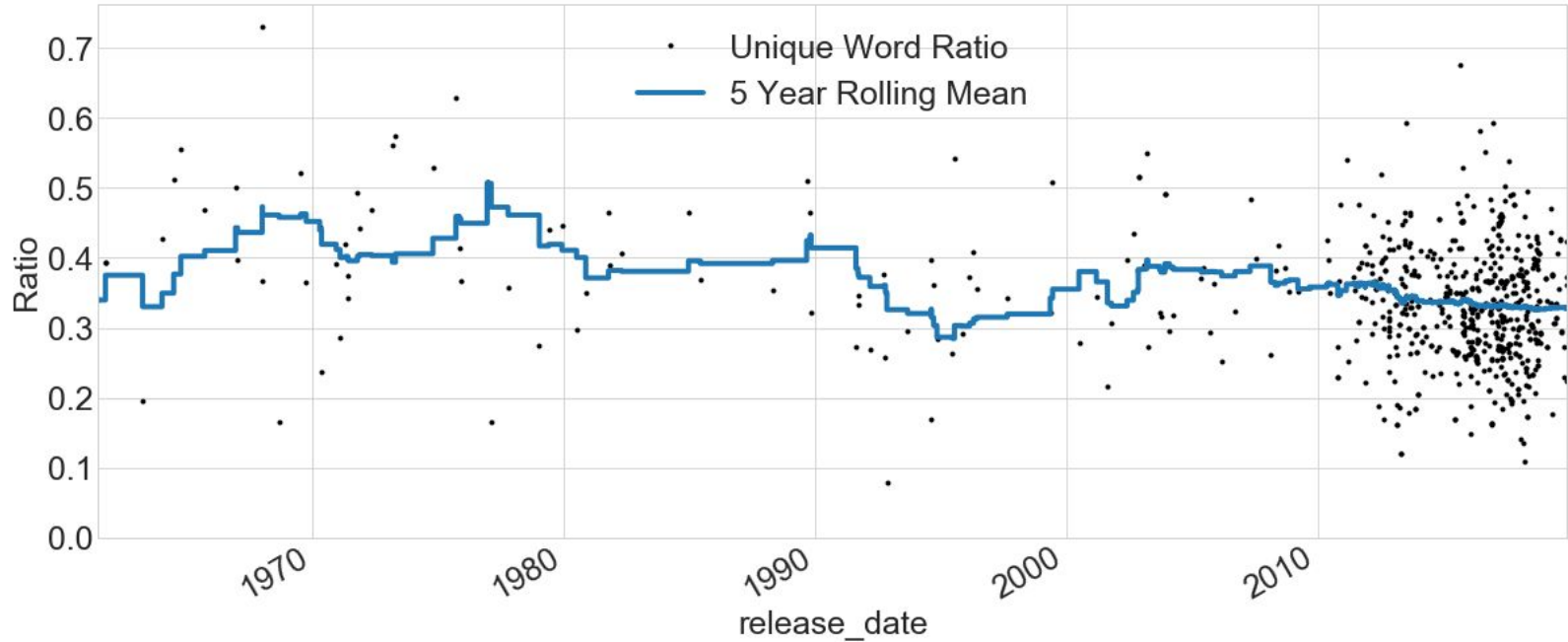
Song with Least Amount of Unique Words

Killing in the name of
Some those that work forces, are same burn crosses
Ugh
Now you do what they told ya
And now
But
Well
Those who died justified
For wearing badge, they're chosen whites
You justify
By you're under control
Come on
Yeah
Fuck you, I won't tell me
Motherfucker

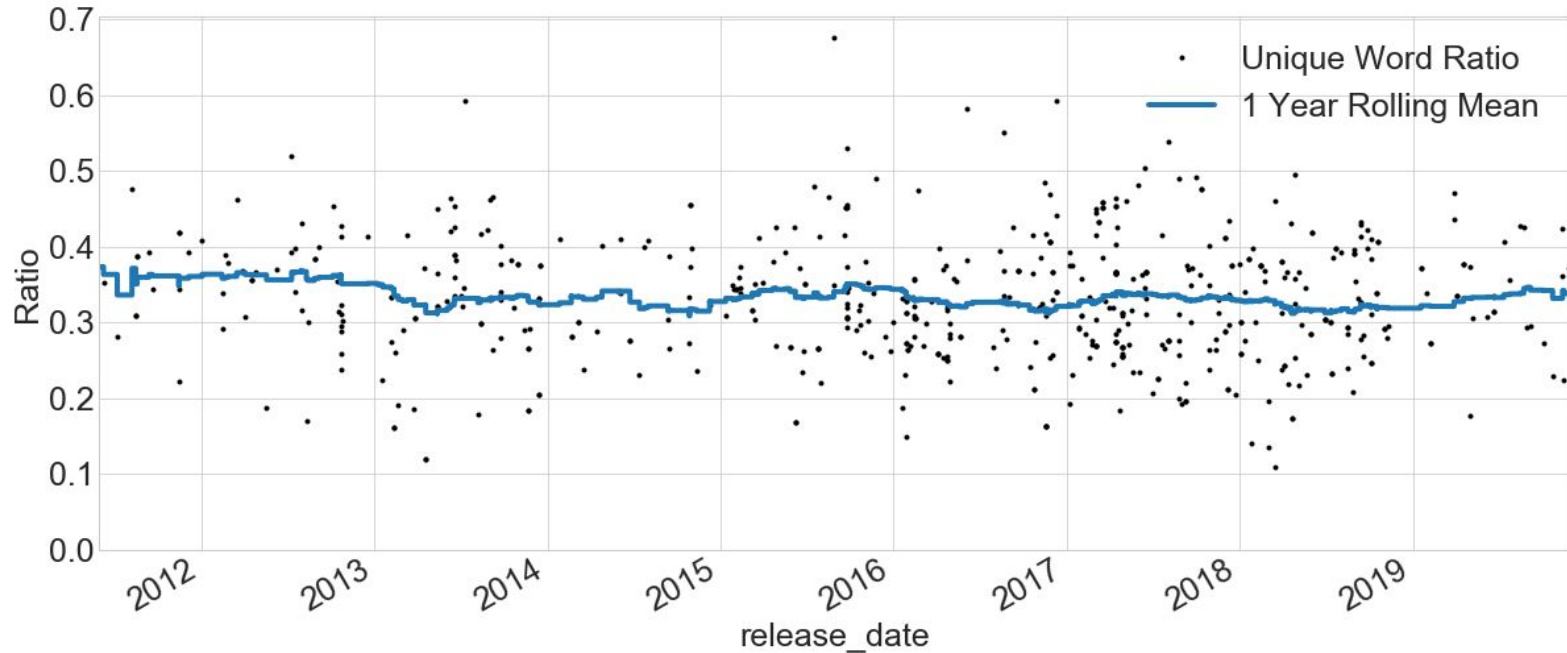
Unique Word Ratio per Genre



Unique Word Ratio over Time

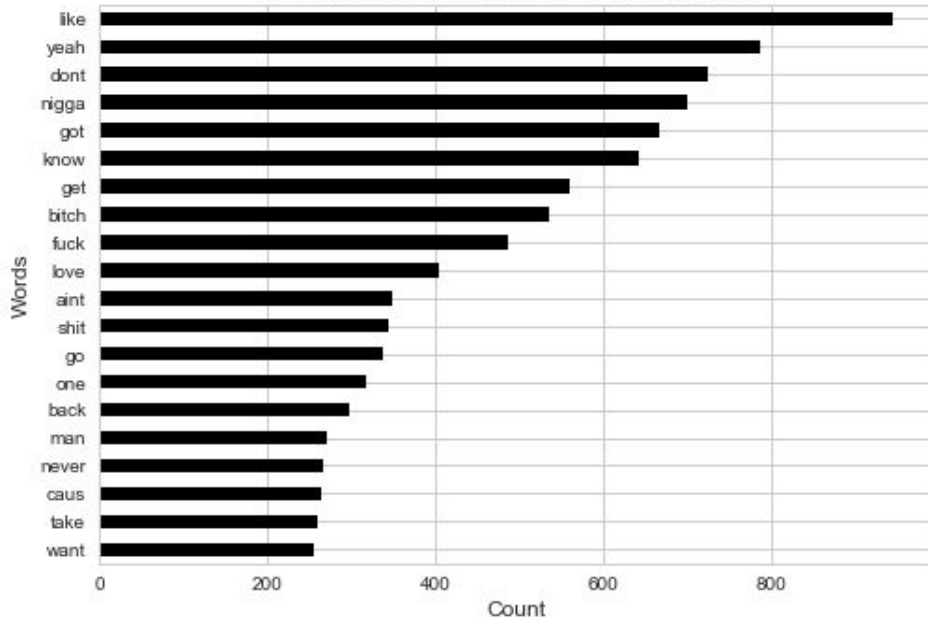


Unique Word Ratio From 2011 Upwards

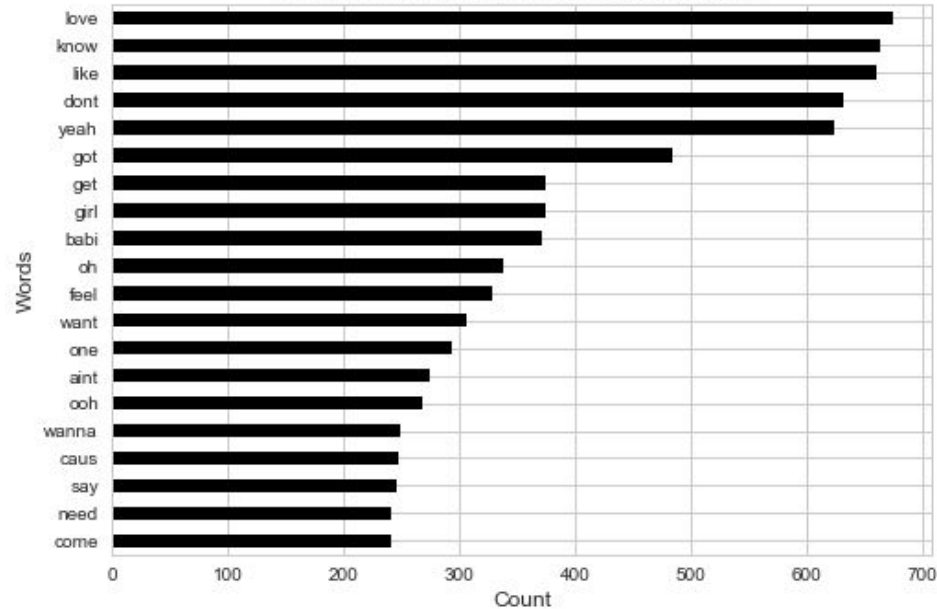


20 Most Frequent Words per Genre

20 Most Frequent Words for Category: Rap



20 Most Frequent Words for Category: Rb



20 Most Frequent Words per Genre

I Like It Cardi B, Bad Bunny & J Balvin

[Chorus: Cardi B]

Diamond district in the chain, chain (*I said I like it like that*)

Certified, you know I'm gang, gang, gang, gang (*I said I like it like—*
woo)

Drop the top and blow the brains, woo (*Woo, I said I like it like*
that)

Oh, he's so handsome, what's his name? Yeah (*Woo, bags, I said I*
like it)

Oh, I need the dollars, cha-ching (*I said I like it like that*)

Beat it up like piñatas (*I said I like it like—; uh*)

Tell the driver, close the curtains (*I said I like it like that, skrrt*)

Bad bitch make you nervous (*I said I like it*)

Cardi B

[Chorus]

Shut your mouth, baby, stand and deliver (*Like a river, like a river*)

Holy hands, will they make me a sinner? (*Like a river, like a river*)

Like a river, like a river

Shut your mouth and run me like a river

Choke this love 'til the veins start to shiver (*Like a river, like a river*)

One last breath 'til the tears start to wither (*Like a river, like a river*)

Like a river, like a river

Shut your mouth and run me like a river

Hold Up Beyoncé

[Chorus]

Hold up, they don't love you like I love you

Slow down, they don't love you like I love you

Back up, they don't love you like I love you

Step down, they don't love you like I love you

Can't you see there's no other man above you?

What a wicked way to treat the girl that loves you

Hold up, they don't love you like I love you

Oh, down, they don't love you like I love you

[Chorus]

To love, love, yeah

To love, love, yeah

To love, yeah

I needed to hate you to love me, yeah

To love, love, yeah

To love, love, yeah

To love, yeah

I needed to lose you to love me

To love, love, yeah

To love, love, yeah

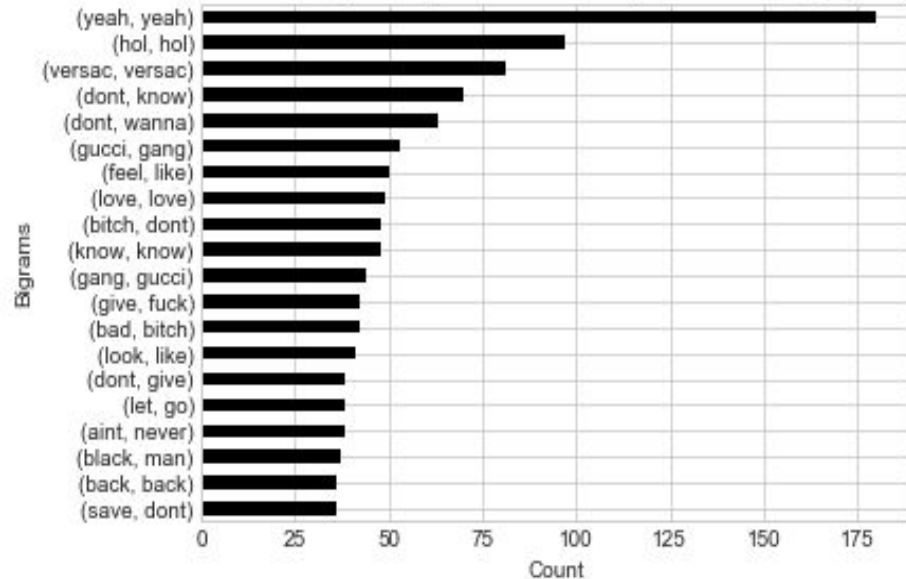
To love, yeah

River Bishop Briggs

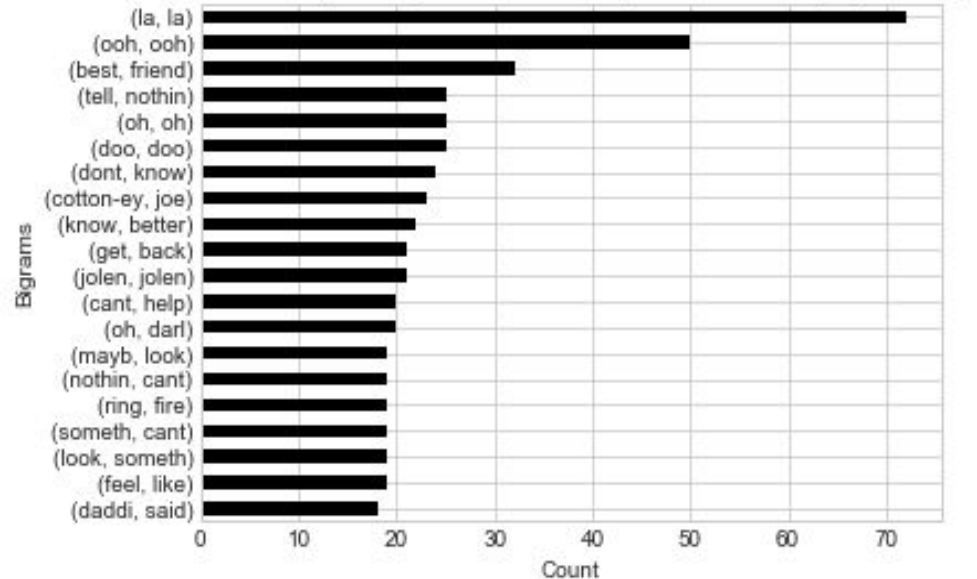
Lose You To Love Me Selena Gomez

20 Most Frequent Bi-Words per Genre

20 Most Frequent Bigrams Without Stopwords for Category: Rap



20 Most Frequent Bigrams Without Stopwords for Category: Country



20 Most Frequent Bi-Words per Genre

Versace
Migos

[Chorus: Quavo]

Versace, Versace, Versace, Versace

Versace, Versace, Versace, Versace

Versace, Versace Versace, Versace Versace

Versace, Versace Versace, Versace Versace

How many times does Lil Pump say "Gucci Gang" in this song?

Lil Pump says "Gucci Gang" a total of 53 times throughout the song.



+239



4 contributors

Cotton Eye Joe
Rednex

If it hadn't been for Cotton-Eyed Joe

I'd been married long time ago

Where did you come from? Where did you go?

Where did you come from, Cotton-Eyed Joe?

If it hadn't been for Cotton-Eyed Joe

I'd been married long time ago

Where did you come from? Where did you go?

Where did you come from, Cotton-Eyed Joe?

[Chorus]

Jolene, Jolene, Jolene, Jolene

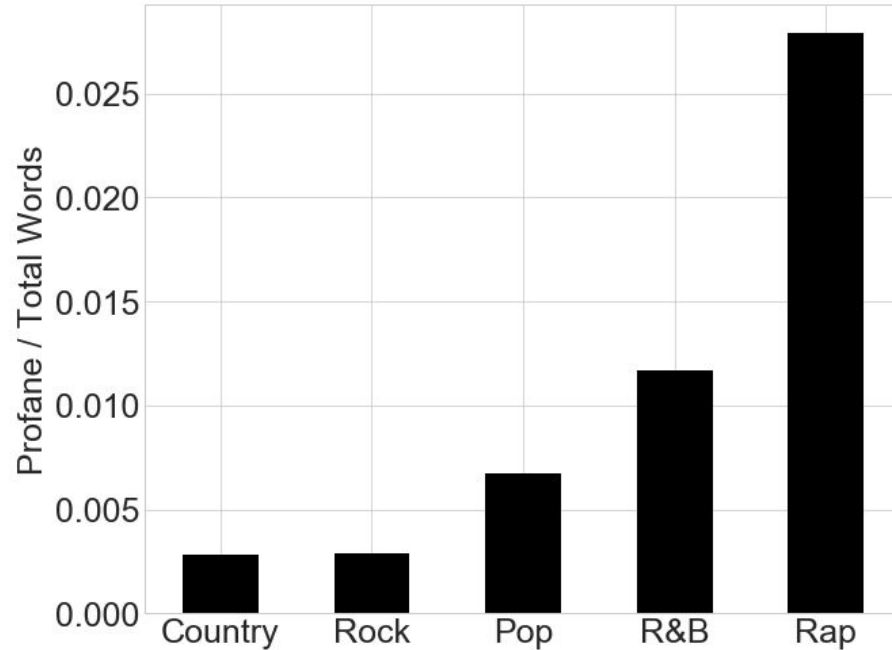
I'm begging of you please don't take my man

Jolene, Jolene, Jolene, Jolene

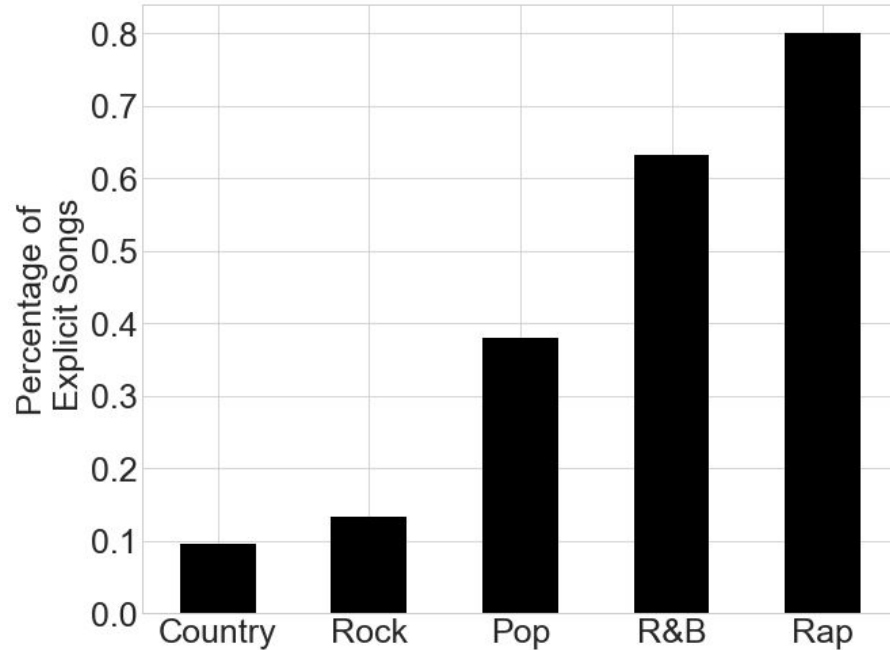
Please don't take him just because you can

Jolene
Dolly Parton

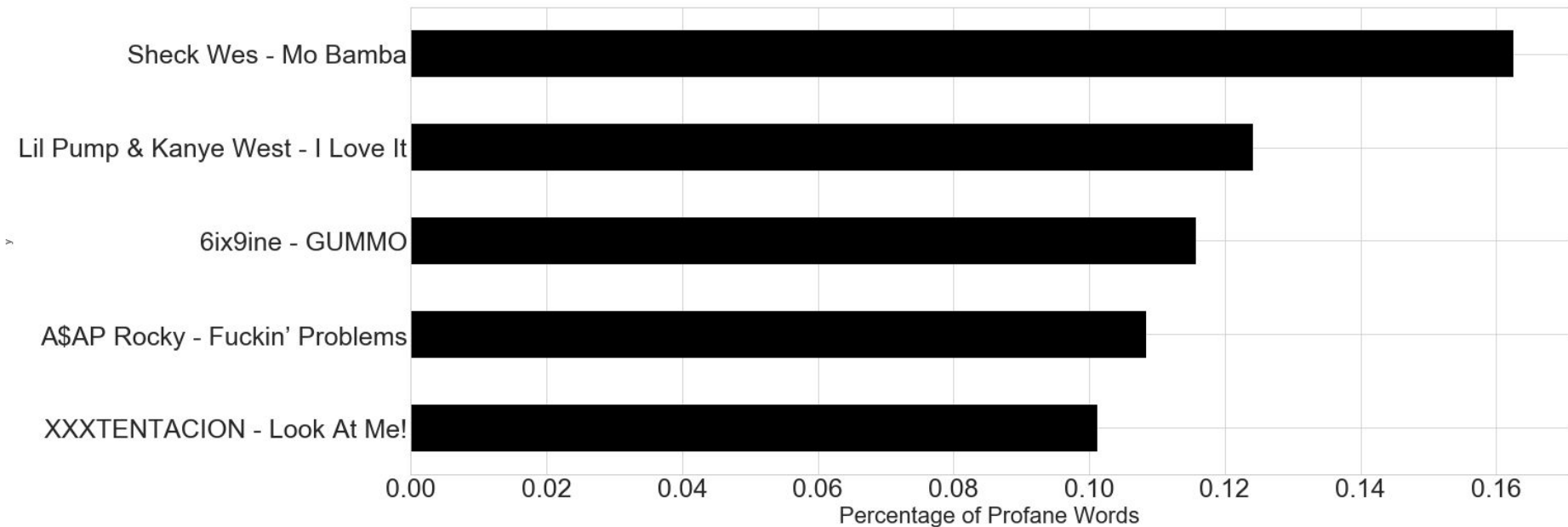
Percentage of Profane Words per Genre



Percentage of Explicit Songs per Genre

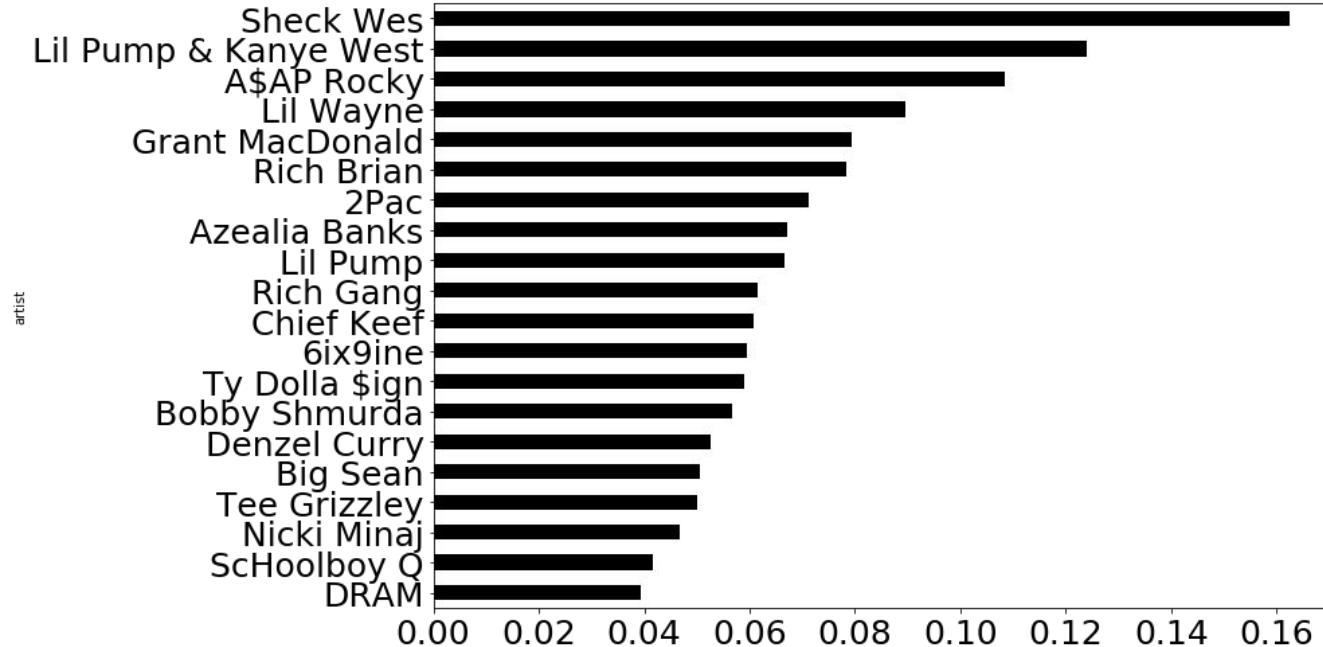


Top 5 Most Explicit Songs

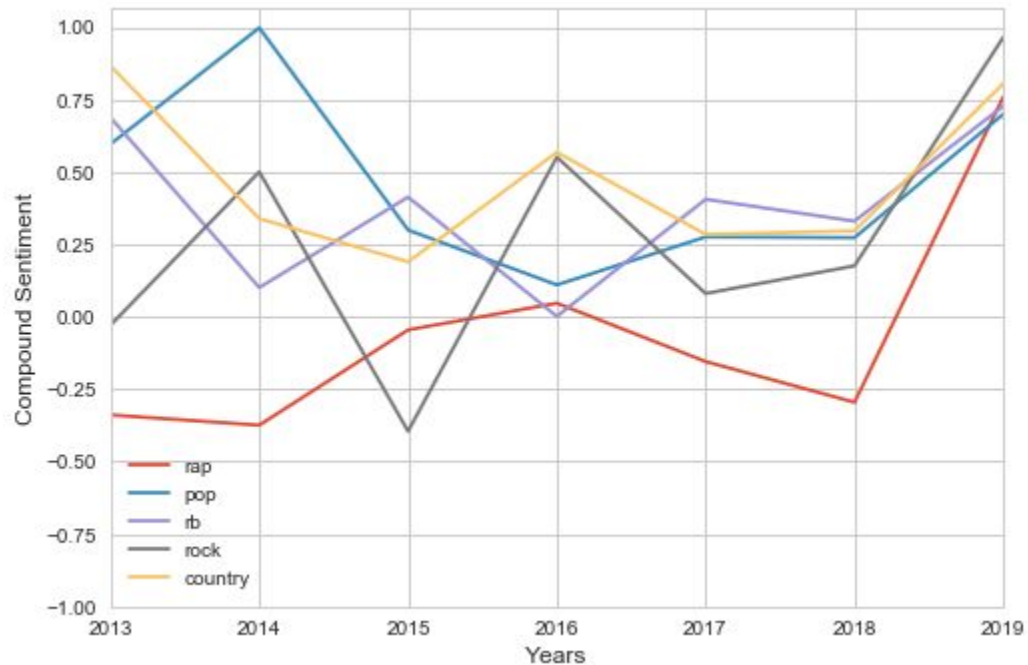


mean: 0.011126

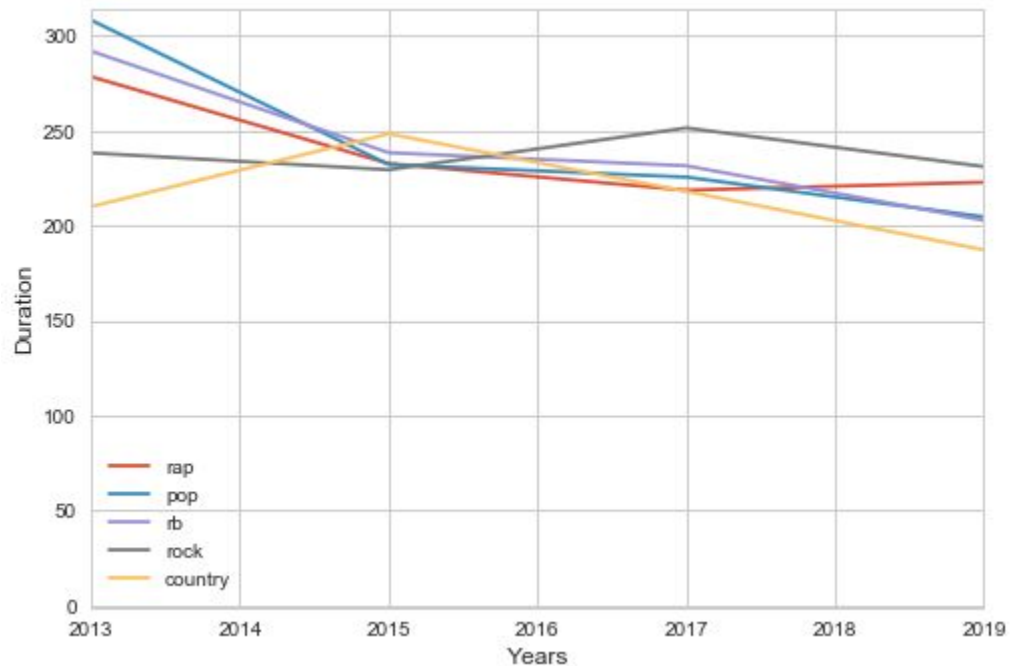
Top 20 Most Explicit Artists



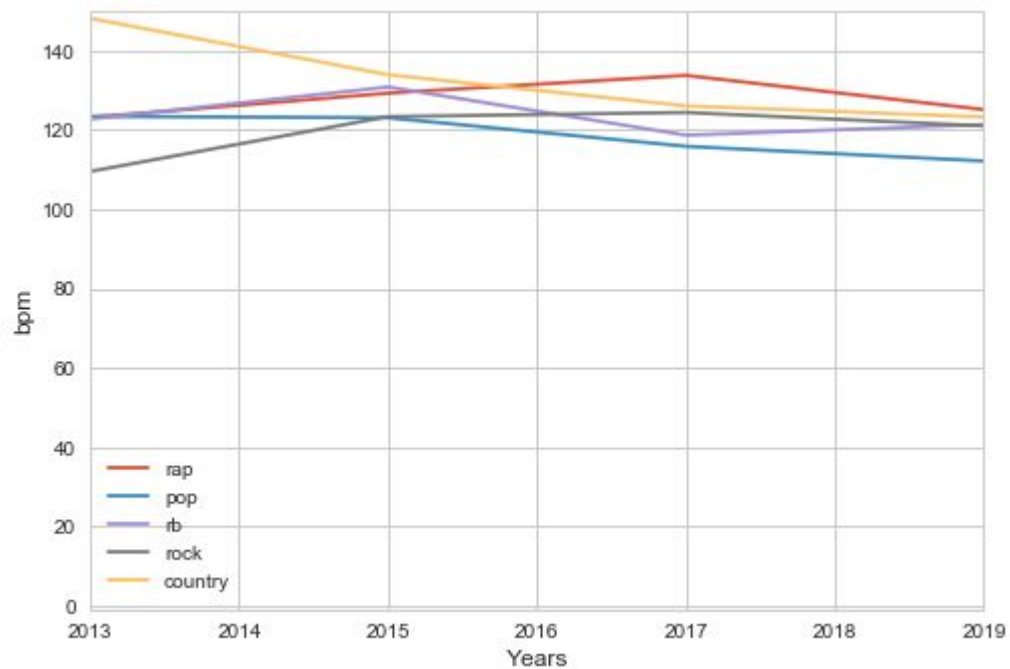
Song Sentiment



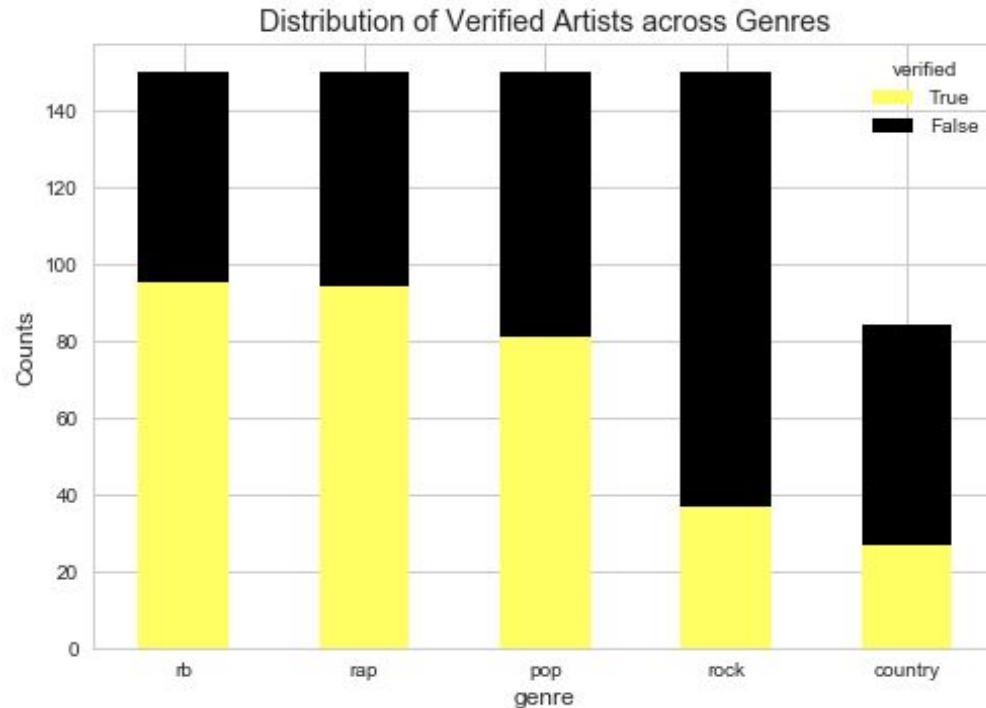
Song Duration



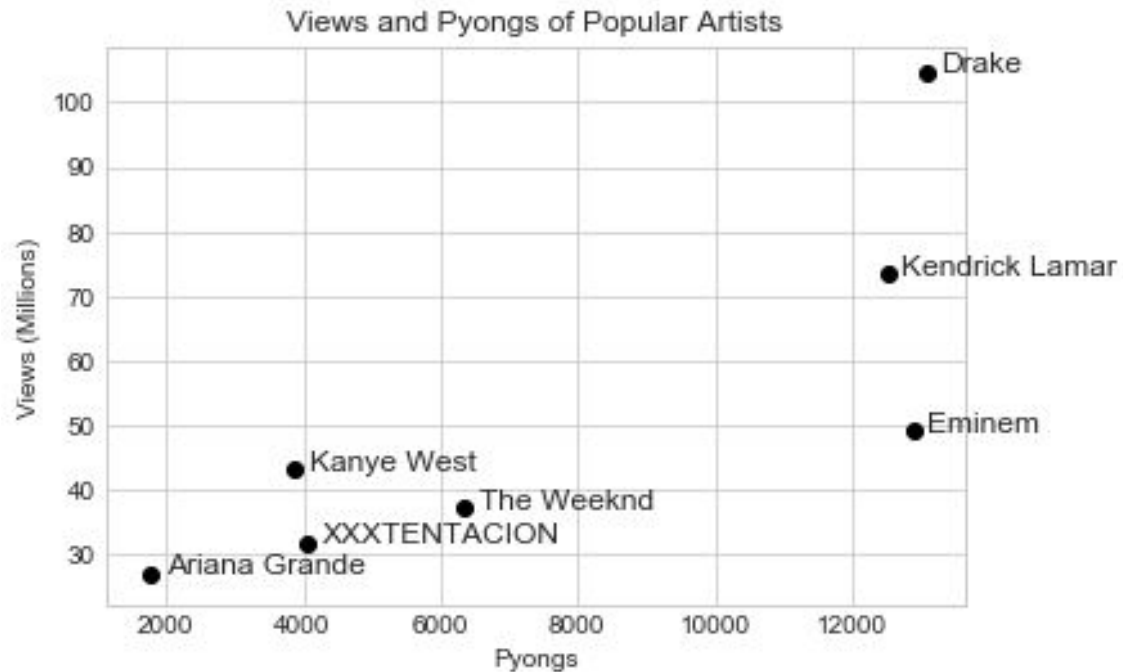
Song BPM



Platform Use by Artists

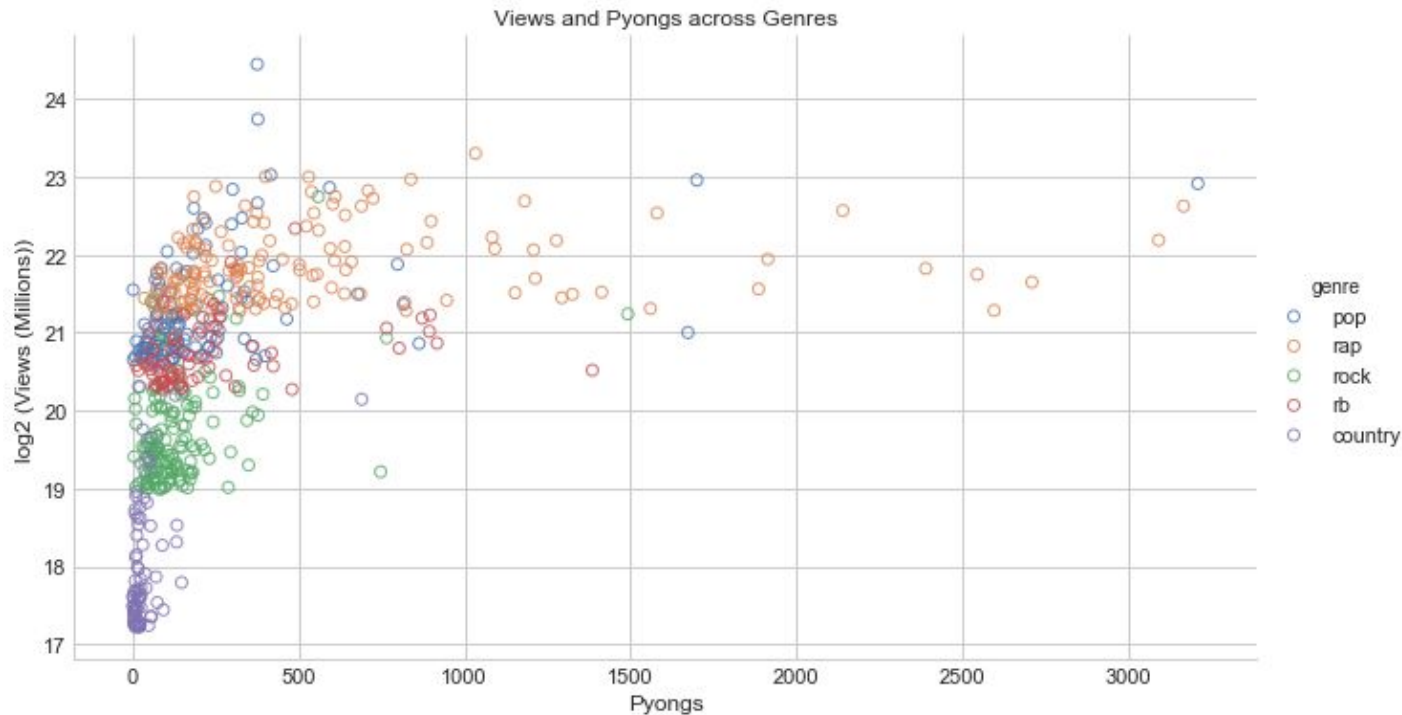


Platform Use by Users



Nr 8:
Original Broadway Cast of Hamilton

Platform Use by Users



Conclusion

- What are the characteristics of the genres, most popular songs and artists on www.genius.com? And how do they differ?

Genres

Most Views:	Rap	Least Views:	Country
Most Unique:	Country	Least Unique:	Rap
Most Explicit:	Rap	Least Explicit:	Country
Most Verified Artists:	Rap + R&B	Least Verified Artists:	Rock + Country
Most Negative Sentiment:	Rap		

Conclusion

Worduse

All genres except Rap largely same

Rap more explicit

'Like' most common single word

'Yeah Yeah' most common biword

Artists

Most liked (pyongs/views)

Top 3 most viewed

Eminem

1. Drake

2. Kendrick

3. Eminem

Over Time

 Unique word ratio

 Sentiment

 BPM

 Duration



Questions?

