WELCOME

Macron v. Le Pen

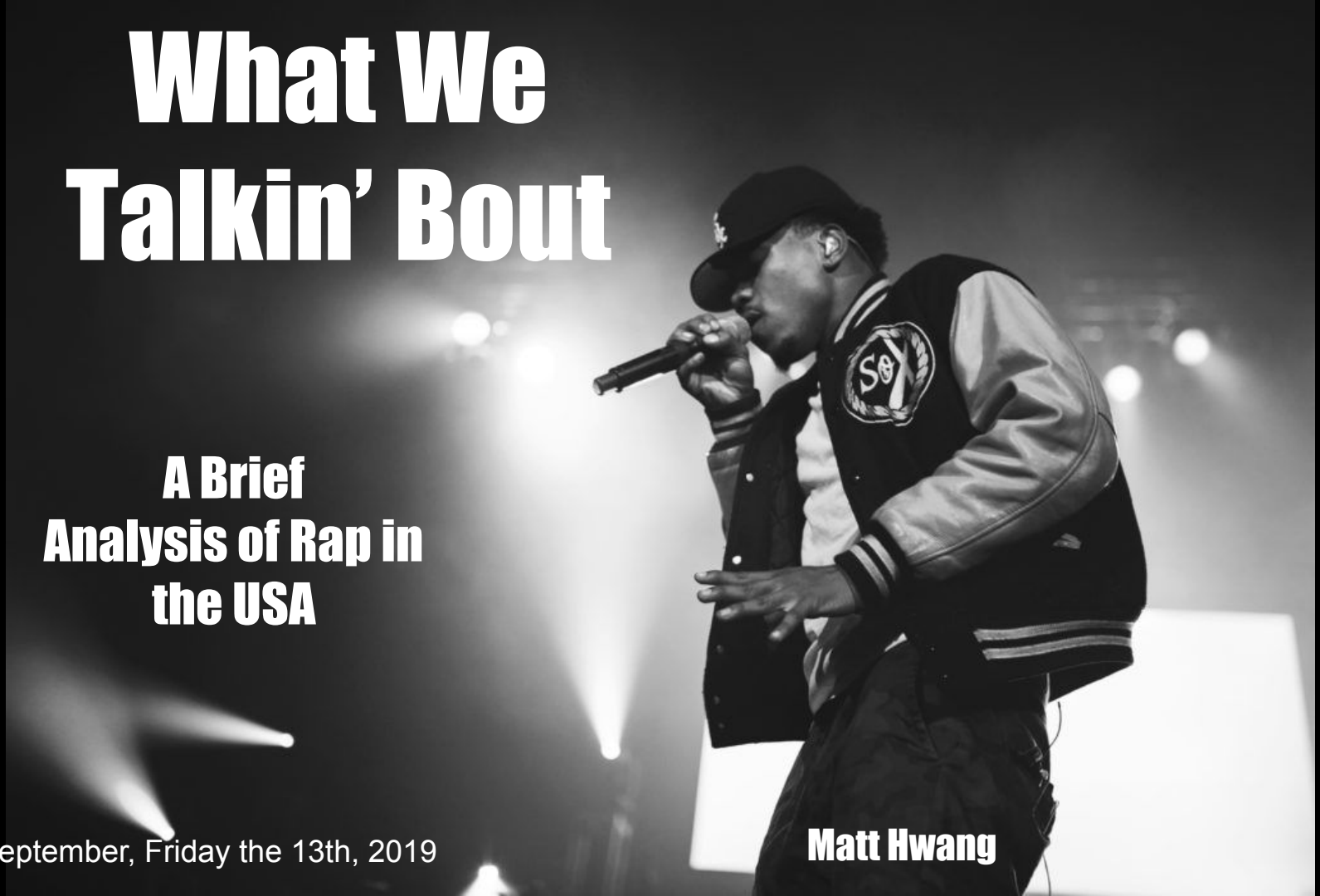
A Data Story

Lea Balido
Adam Ricks
Matt Hwang

8-30-19

What We Talkin' Bout

A Brief Analysis of Rap in the USA

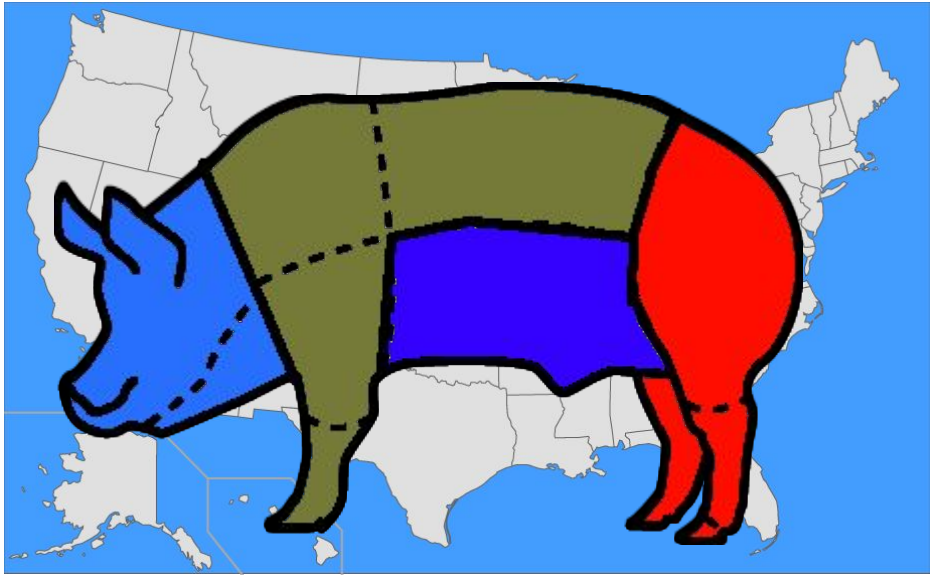


September, Friday the 13th, 2019

Matt Hwang

Rap in the USA

- Although rap music has its origins in NY, it has spread in popularity across the US
- It be generalized and subdivided into four regions
  - West Coast
  - East Coast
  - Mid West
  - Down South



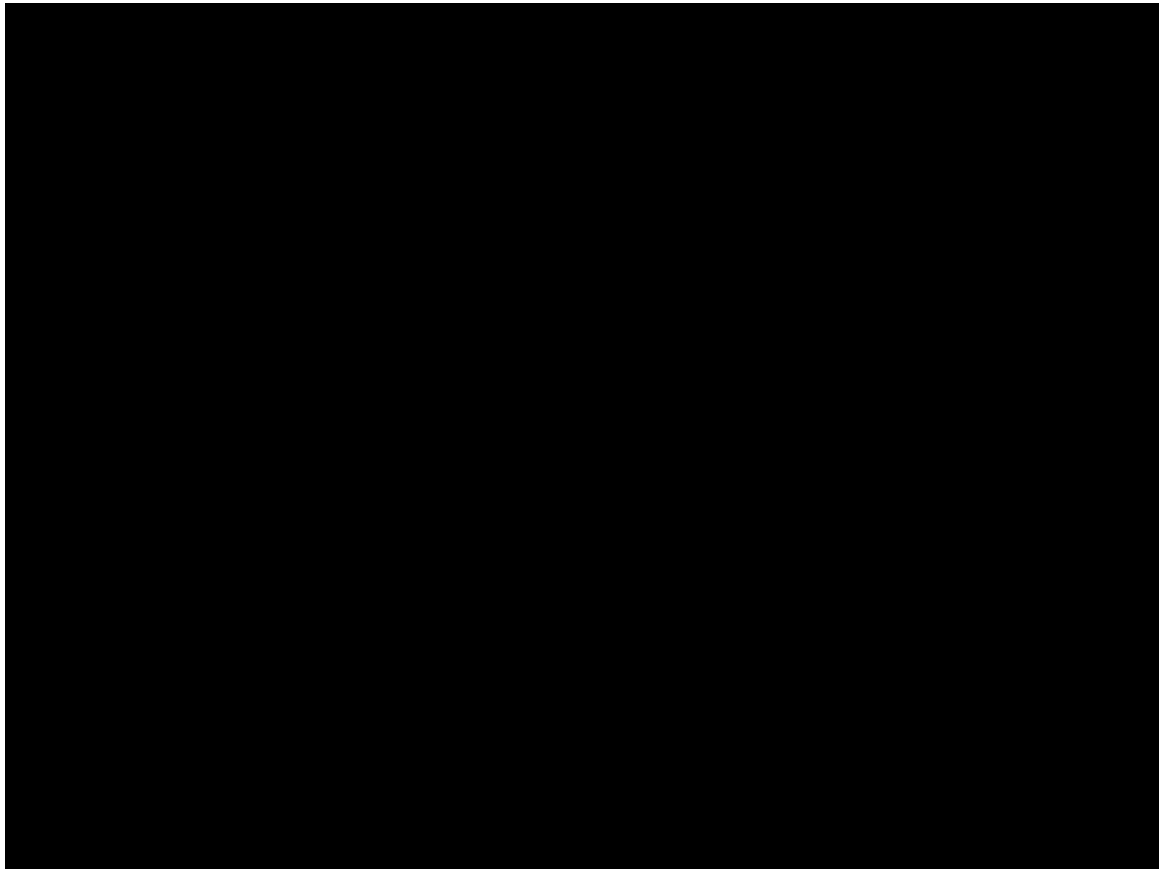
Ranker.com

-Ranker is a public platform where users can vote and “rank” their opinions on favorite topics

-Used to determine top representatives per region for analysis

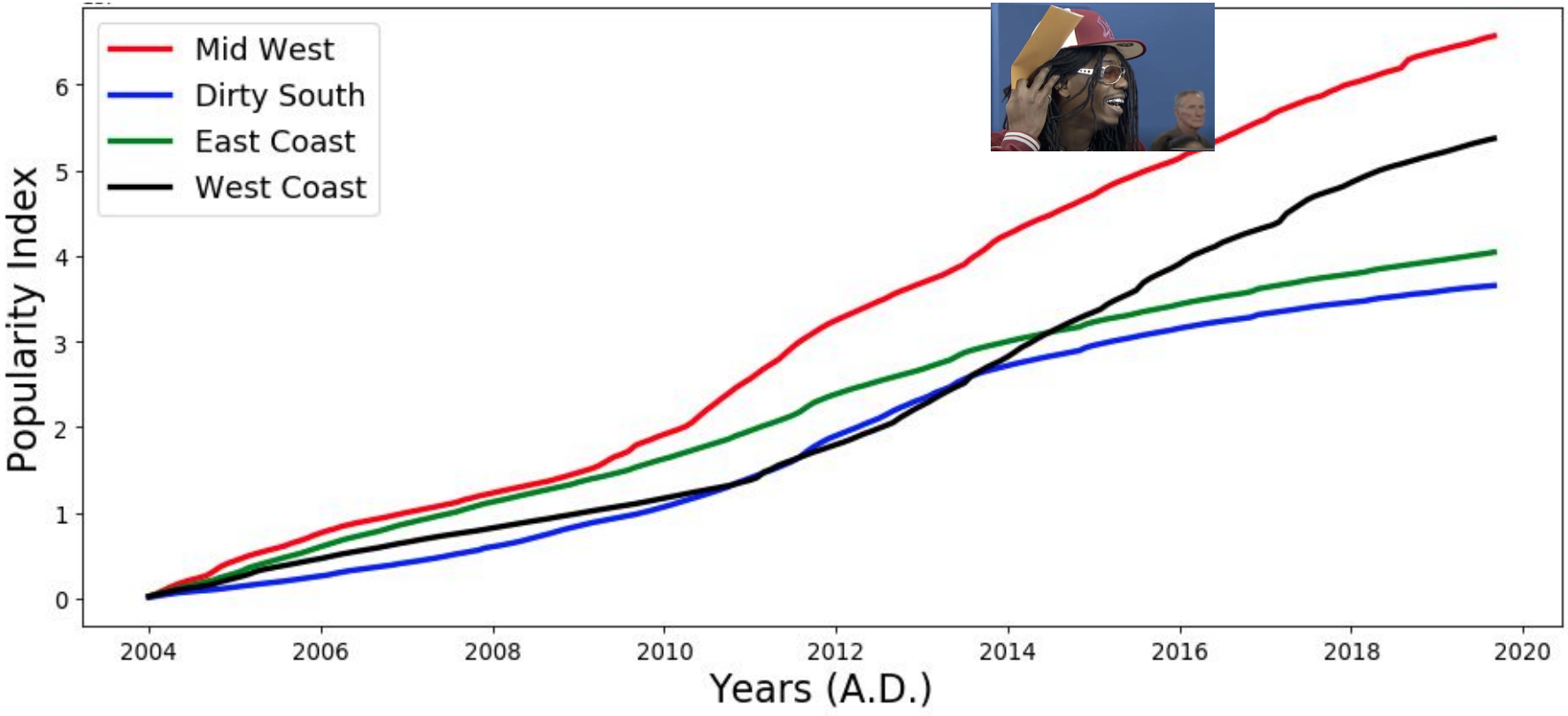
-Chose top 25 artists per region and gathered data via a python webscraping script

Regional Popularity Over Time



Using a popularity metric based on multiple sources the relative regional popularity can be estimated when overlayed on top of historical Google Search Trending Data

Regional Rap Popularity in America 2004 - Present



What's Going On In The Mid West?

- Why are they more popular over the artists based in the rest of the USA?
- Are there any obvious features that may correlate to their popularity?
- Is this why Kanye West purchased a ranch in Wyoming?

Let's find out with DATA

Lyrical Analysis

- A potential indication of popularity may stem from the **topic** of discussion
- A database of all song lyrics for the top 3 artists per region (from Ranker) was created by scraping a lyrics website (AZ-Lyrics.com) with a python webscraping tool and some VPN magic

felt guilty 😔

West Coast	East Coast	Mid West	Dirty South
- 2Pac	- Jay-Z	- Kanye	- J.Cole
- Snoop Dogg	- 50 Cent	- Eminem	- 2 Chainz
- Kendrick Lamar	- Biggie Smalls	- Big Sean	- Lil' Wayne

Lyrical Analysis Cont.

Web Harvest Yielded:

- 12 Artists
- 3,440 songs
- 1,000,000+ words
- One Love

- Filtered HTML Jargon and removed STOP / filler words from analysis

Results - Started From The Bottom

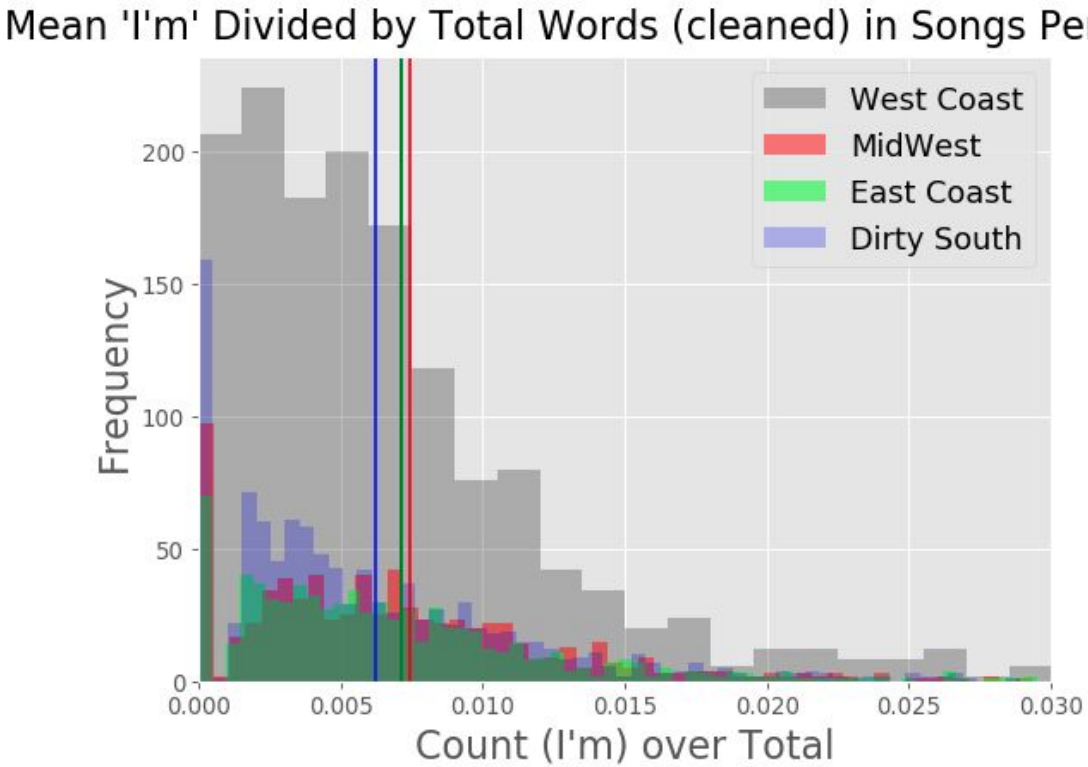
Now we're here

- Top word across all artists and regions is "**I'm**" signaling the artists' tendencies to speak about themselves and their stories and also hobbies and interests.
- Also a large mixture of less than nice words floated to the top.

Is the frequency of talking about themselves or frequency of cursing indicative of popularity?

Word	Count	artist
im	4831	Lil' Wayne
im	3412	Eminem
[redacted]	2691	Lil' Wayne
im	2662	50 Cent
im	2566	Snoop Dogg
im	1980	2Pac
[redacted]	1950	Lil' Wayne
im	1770	2 Chainz
im	1737	Jay-Z
[redacted]	1695	50 Cent
im	1611	Big Sean
yeah	1530	Lil' Wayne
[redacted]	1519	2Pac
[redacted]	1434	Lil' Wayne
money	1427	Lil' Wayne
[redacted]	1425	Snoop Dogg

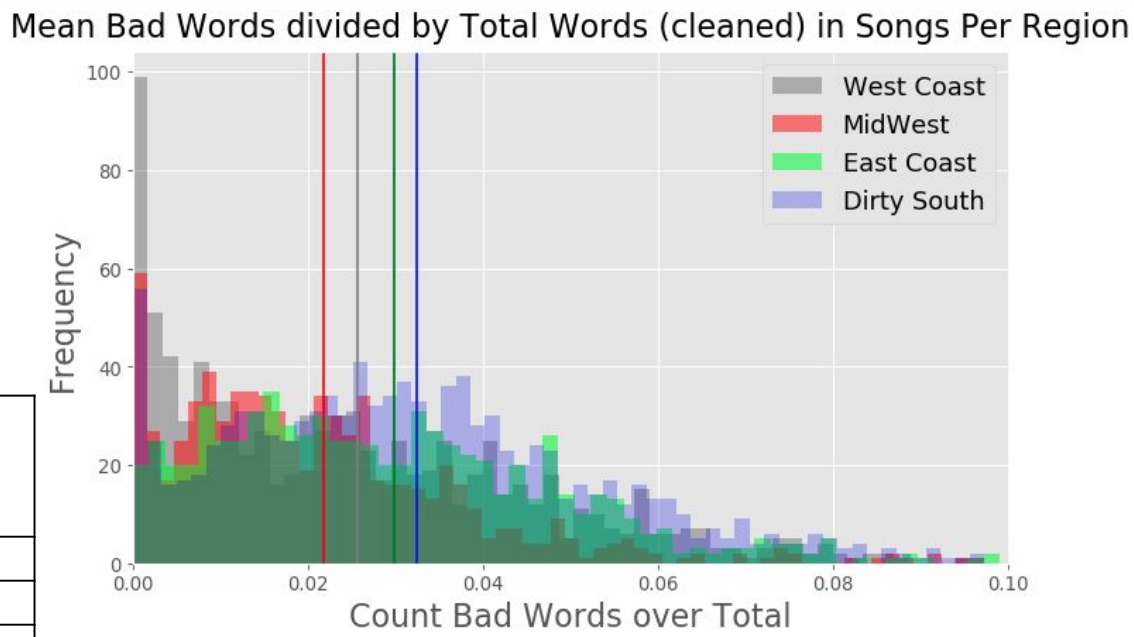
MidWest Vs The Rest



- Artists from the Midwest talk about themselves the most
- Mean self reference per song divided by total words per song
- 0.0075% calculated value for MW on average
  - Once per 135 words

Choice Words 🤫

- Mid West manners run deep
- Calculated by running lyrics through a list of banned Google words



Who Swears the LEAST			
1st Place	Mid West	2.19%	😇
2nd Place	West Coast	2.57%	😮
3rd Place	East Coast	2.99%	😐
Dead Last	Dirty South	3.24%	👿

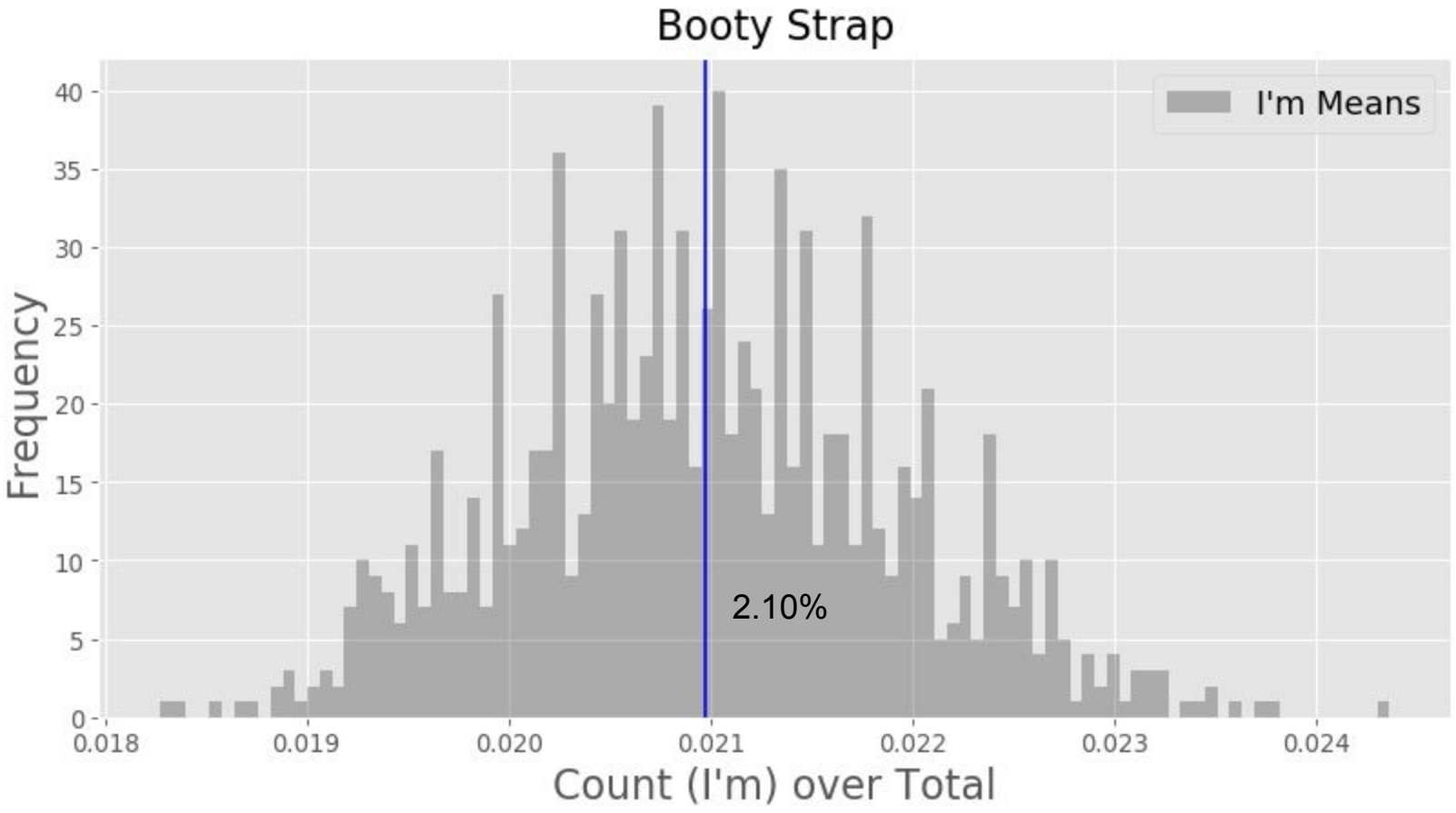
Boot Strap

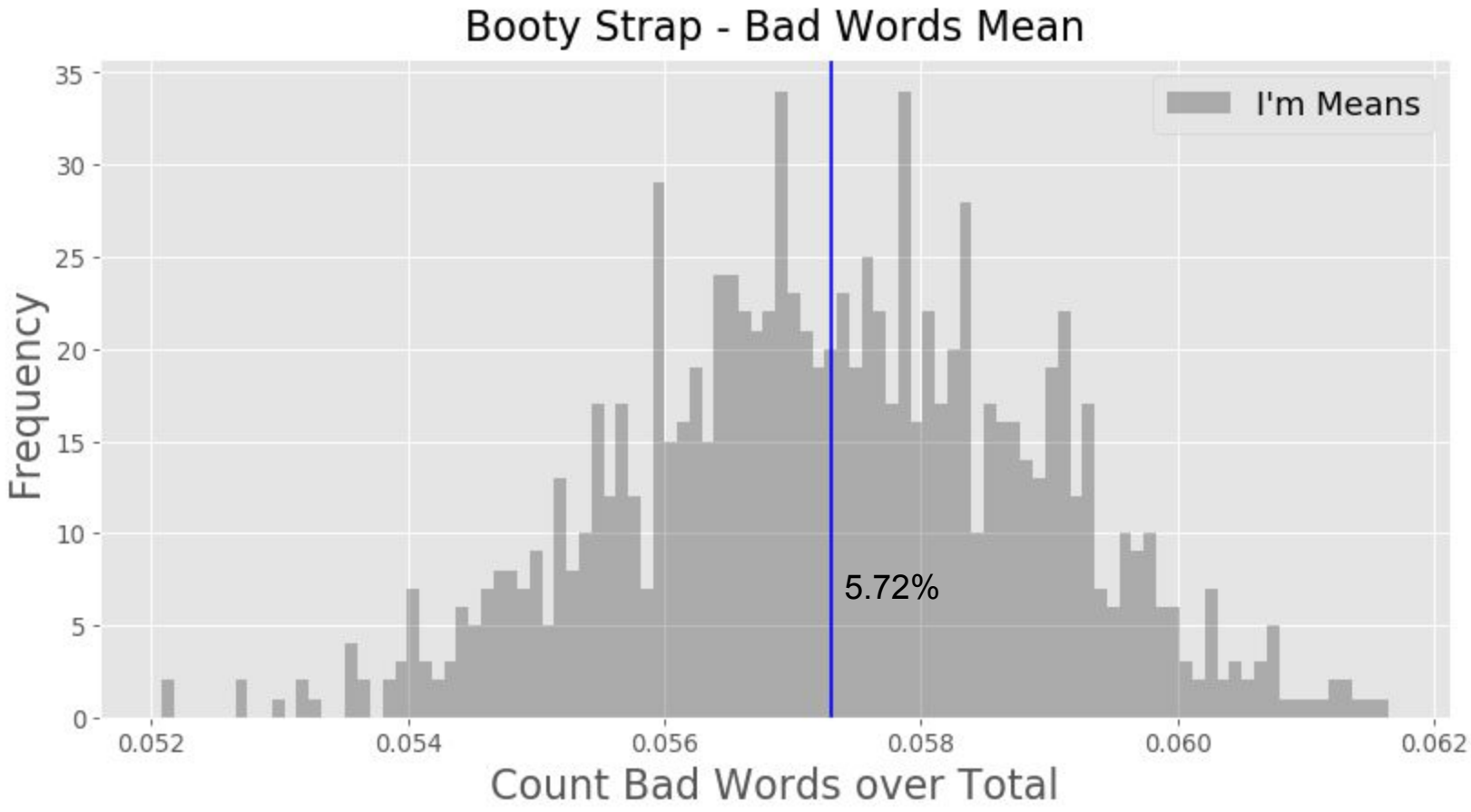
Let’s increase improve the quality of our sampling with Bootstrapping!

Originally I took top 3 per region and analyzed their discography

What could be more representative of the population is taking random 4 songs from top 25 artist per region as an initial sample (stratified random sampling) then running it 1000x times to determine mean values.

DISCLAIMER - I did not finish and only bootstrapped the MidWest and also took a different approach that produced different numbers


Booty Strap
I'm Means
Frequency
40
35
30
25
20
15
10
5
0
2.10%
0.018
0.019
0.020
0.021
0.022
0.023
0.024
Count (I'm) over Total


Booty Strap - Bad Words Mean
I'm Means
5.72%
Frequency
Count Bad Words over Total
0
5
10
15
20
25
30
35
0.052
0.054
0.056
0.058
0.060
0.062

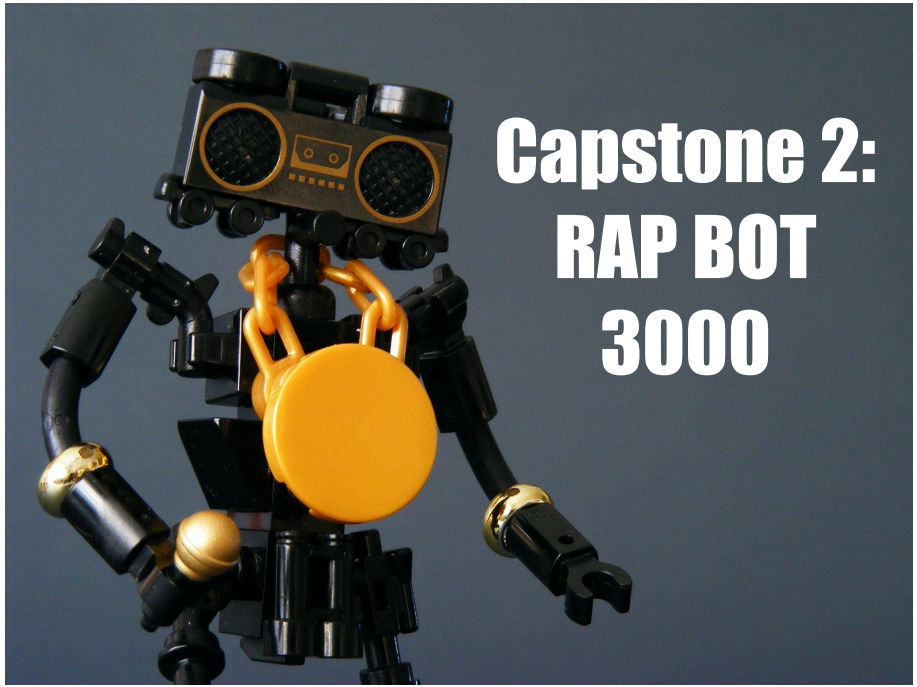
Small Fish Big Pond

Without context it is hard to determine a definitive strong positive correlation

However it is interesting to take note of these trends

Would love to run further analysis on the gathered lyrical datasets with NLP

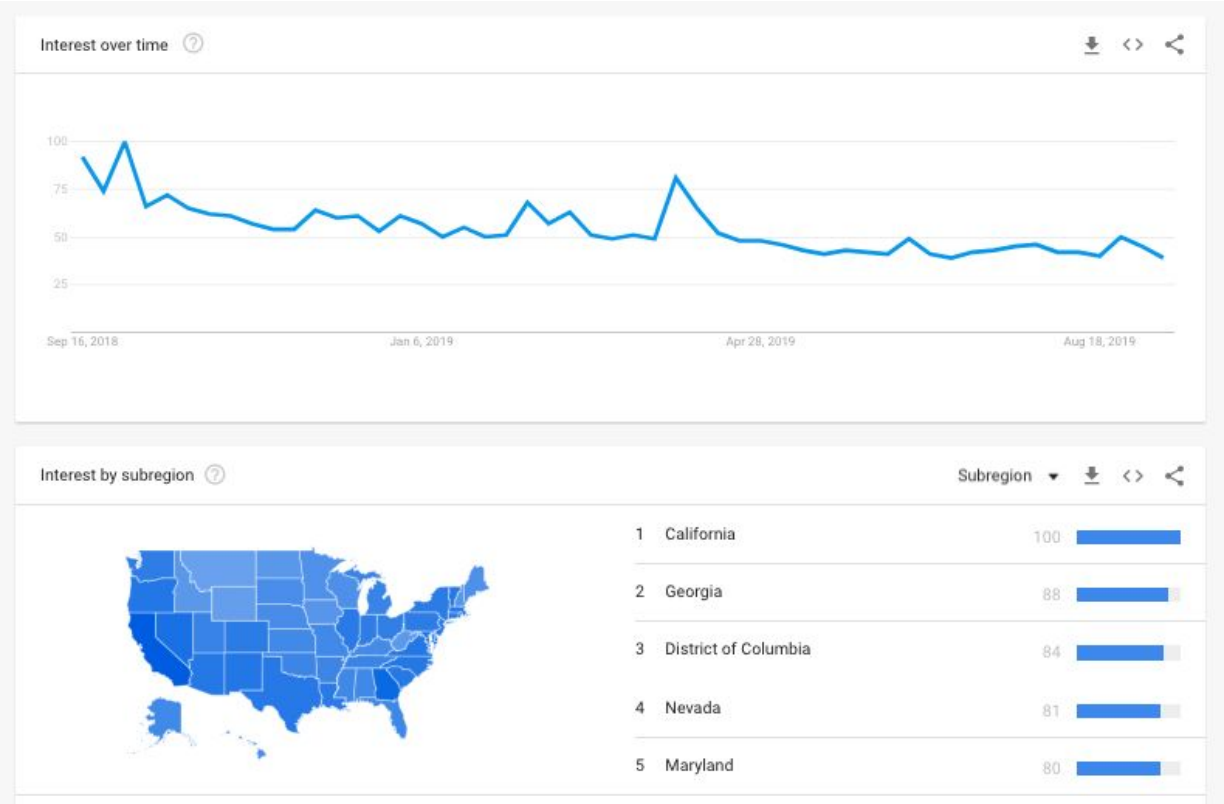
Potential for Machine Learning



The End

Thanks for coming to my TED Talk 🙏

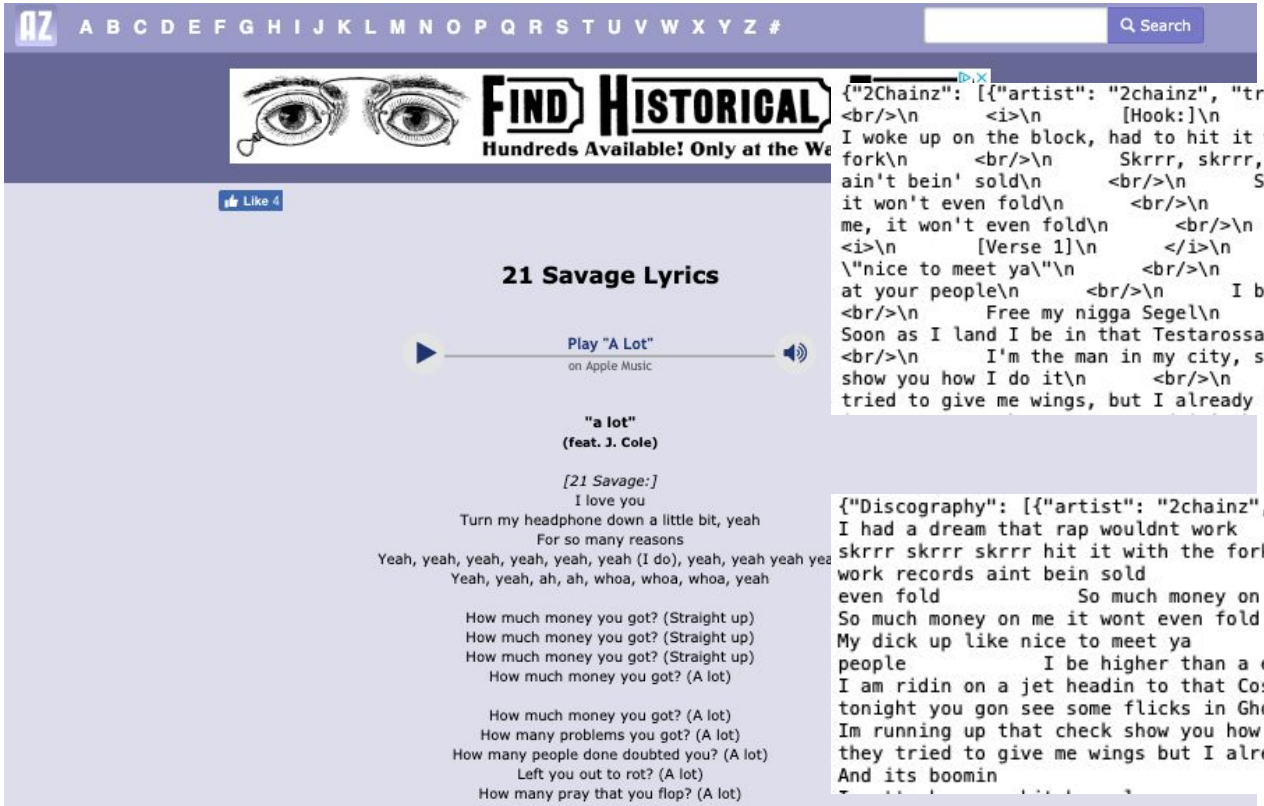
Extra Slides - Google Trends



Category: All categories	
Month	Scarface: (Unit
2004-01	38
2004-02	32
2004-03	28
2004-04	25
2004-05	21
2004-06	19
2004-07	19
2004-08	22
2004-09	24
2004-10	25
2004-11	23
2004-12	24
2005-01	24
2005-02	24
2005-03	25
2005-04	23
2005-05	27
2005-06	25
2005-07	22
2005-08	19
2005-09	20
2005-10	29
2005-11	31

	MidWest_Spotify_Totals	Dirty_South_Spotify_Totals	West_Coast_Spotify_Totals	East_Coast_Spotify_Totals	date
0	246620	122764	215113	233593	2004-01-01
1	535356	279395	435139	483267	2004-02-01
2	920519	404058	621445	719941	2004-03-01
3	1317393	537931	798758	933802	2004-04-01
4	1656751	650154	982855	1173643	2004-05-01
5	1950342	735140	1123947	1381209	2004-06-01
6	2194538	806722	1252312	1552101	2004-07-01
7	2405427	872273	1362015	1744674	2004-08-01
8	2659478	942744	1496787	1938362	2004-09-01
9	3145252	1023320	1669152	2146739	2004-10-01
10	3712830	1111326	1935171	2456835	2004-11-01
11	4084066	1220215	2145827	2698273	2004-12-01

Extra Slides - AZ-Lyrics




```
{"2Chainz": [{"artist": "2chainz", "track": "fork", "lyrics": "\n      <i>\n       [Intro
<br/>\n      <i>\n       [Hook:]\n      </i>\n      <br/>\n      I had a dream that rap w
I woke up on the block, had to hit it with the fork\n      <br/>\n      Skrrr, skrrr, skr
fork\n      <br/>\n      Skrrr, skrrr, skrrr, skrrr: hit it with the fork\n      <br/>\n
ain't bein' sold\n      <br/>\n      So much money on me, it won't even fold\n      <br/>
it won't even fold\n      <br/>\n      So much money on me, it won't even fold\n      <br
me, it won't even fold\n      <br/>\n      So much money on me, it won't even fold\n
<i>\n       [Verse 1]\n      </i>\n      <br/>\n      I got Medusa on my sneakers\n
\"nice to meet ya\"\n      <br/>\n      100K for a feature, hundred K's at my leisure\n
at your people\n      <br/>\n      I be higher than a eagle\n      <br/>\n      When I'm
<br/>\n      Free my nigga Segel\n      <br/>\n      I am ridin' on a jet, headin' to tha
Soon as I land I be in that Testarossa\n      <br/>\n      If I die tonight, you gon' see
<br/>\n      I'm the man in my city, same thing in South Dakota\n      <br/>\n      Man I
show you how I do it\n      <br/>\n      I drink red bitches, I don't drink Red Bulls\n
tried to give me wings, but I already had some\n      <br/>\n      I'm all that and then
```

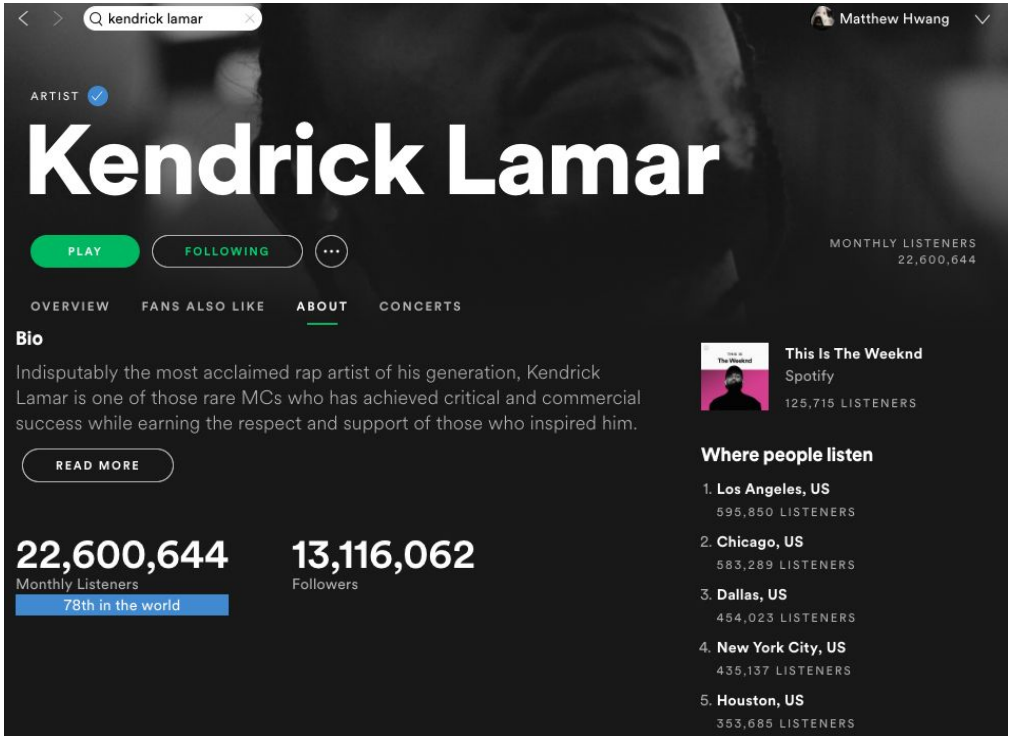
```
{"Discography": [{"artist": "2chainz", "track": "fork", "lyrics": "
I had a dream that rap wouldnt work           I woke up on the block had to hit it wit
skrrr skrrr skrrr hit it with the fork           Skrrr skrrr skrrr skrrr hit it with t
work records aint bein sold           So much money on me it wont even fold
even fold           So much money on me it wont even fold           So much money on
So much money on me it wont even fold                                   I got M
My dick up like nice to meet ya           100K for a feature hundred Ks at my leisure
people           I be higher than a eagle           When Im sipping on that codeine
I am ridin on a jet headin to that Costa           Soon as I land I be in that Testaro
tonight you gon see some flicks in Ghosta           Im the man in my city same thing i
Im running up that check show you how I do it           I drink red bitches I dont dri
they tried to give me wings but I already had some           Im all that and then some
And its boomin
```

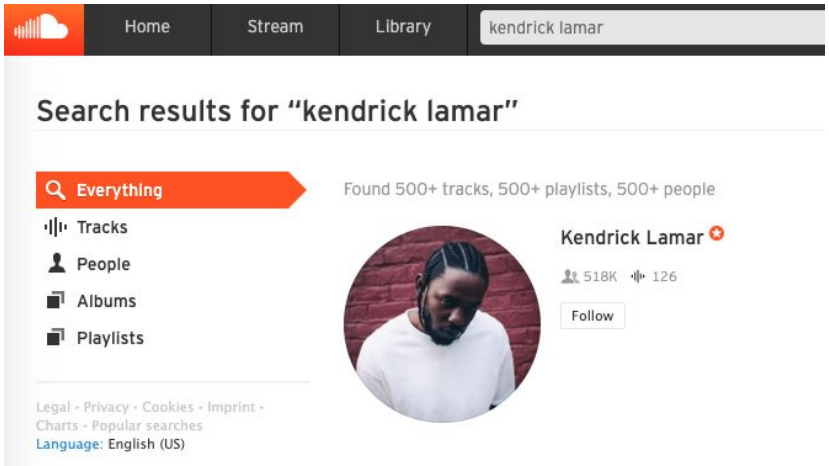
Webscraping Workflow

In [22]:

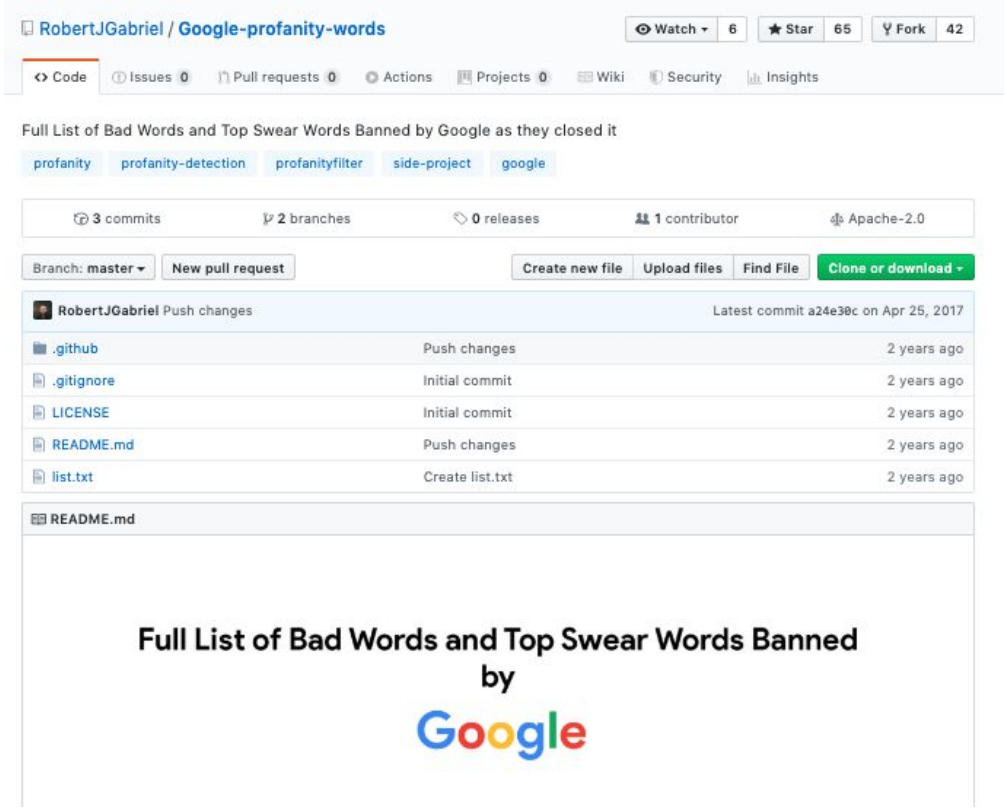
```
#WEB TO LETTER LINKS
def get_letter_links(website_url):
    soup = BeautifulSoup(requests.get(website_url).content, 'html.parser') #get that shit
    links = []
    home_url = "https:"
    for each in soup.find_all('a'):
        links.append(home_url + each.get("href")) #store that shit
    return links[1:28] #gets only nav links, return that shit


#LETTER LINKS TO ARTIST LINKS
def get_artist_links(letter_url):
    soup = BeautifulSoup(requests.get(letter_url).content, 'html.parser') #get that shit
    links = []
    home_url = "https://www.azlyrics.com/"
    for each in soup.find_all('a'):
        links.append(home_url + each.get("href")) #store that shit
    return links[28:-8] #takes out nav links and footer links, return that shit


#ARTIST LINKS TO TRACK LINKS
def get_track_links(artist_home_url):
    name = artist_home_url.split('/')[4].split('.html')[0]
    soup = BeautifulSoup(requests.get(artist_home_url).content, 'html.parser') #get that shit
    links = []
    home_url = "https://www.azlyrics.com"
    for each in soup.find_all('a'):
        links.append(each.get("href").replace("..",home_url)) #store that shit
    return {name:links[30:-8]} #takes out nav links and footer links, return that shit as a dict for JSON reasons
```

Extra Slides - Spotify / Sound Cloud





Extra Slides - Profanity Check



https://github.com/RobertJGabriel/Google-profanity-words