

Vocabulário dos Artistas Musicais Brasileiros

Paula K. Miyashita,
Tamiris G. S. Lira,
Felipe R. Yoshimura,
Humberto V. T. Correa



The Largest Vocabulary In Hip Hop

Rappers, ranked by the number of unique words used in their lyrics

By Matt Daniels

Updated on January 21, 2019

with 75 new rappers including Brookhampton, Death
Grips, Lil Uzi Vert, Travis Scott, and Migos

This project was originally published in 2014 and recently updated in **January 2019** with newer lyrics data and 75 additional artists, including Lil Uzi Vert, Lil Yachty, Migos, and 21 Savage.

It compares the number of unique words used by some of the most famous artists in hip hop (that is, an example of a quantitative view of lyricism, once proposed by Tahir Hemphill). I used each artist's first 35,000 lyrics. This way, prolific artists, such as Jay-Z, can be compared to newer artists, such as Drake.

“The Pudding explains ideas debated in culture with visual essays. By wielding original datasets, primary research, and interactivity, we try to thoroughly explore complex topics.”

BIO



Matt Daniels is a Journalist-Engineer and Business lead/CEO at The Pudding. He first experienced Internet fame in 2014 and has been chasing that feeling ever since.

🐦 @matthew_daniels

✉️ matt@pudding.cool

STORIES



Are Men Singing Higher in Pop Music?

Men's voices in pop music seem really high. When was vocal register the highest?



Best Year in Music

A journey through every Billboard top 5 hit to find music's greatest era



A People Map of the UK

Where city names are replaced by their most Wikipedia'ed resident.

❖ **Pesquisadores em Música**

- A cultural semantics of string arrangement for recorded Popular music: A model for analysis and practice

❖ **Pesquisadores em Tecnologia**

- Deep Rapping: Character Level Neural Models for Automated Rap Lyrics Composition
- Sistemas de Recomendação?

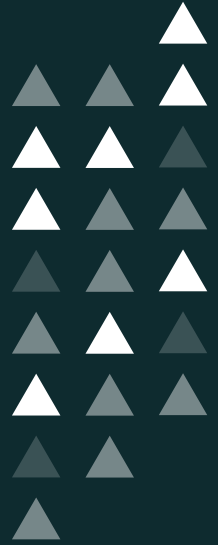
❖ **Pesquisadores em Ciências Sociais/Línguas**

- After Words: The American Language Using Interpretive Design to Provoke Critical Reflection on Language Privilege
- Die politische Dimension von Popmusik: Theoretische Zugänge, empirische Befunde und Potenzial der politikwissenschaftlichen Analyse

❖ **Curiosos**

Público

Metodologia



- ❖ Coleta de Dados
- ❖ Limpeza e Preparação dos Dados
- ❖ Cálculo do Vocabulário
- ❖ Análise dos Dados

Metodologia

❖ Vagalume

- Top 100 Nacional (julho/19)
- Top 25 de cada artista

Coleta dos Dados

❖ **Remoção**

- Coletâneas (ex: Hinos)
- Artistas com menos de 25 músicas
- Trechos de códigos
- Símbolos, pontuação

❖ **Transformação:** letras minúsculas

❖ **Tradução**

Limpeza dos Dados



❖ Tradução

- Função utilizando a biblioteca langid, que utiliza a vizinhança da palavra para determinar o idioma.
- Se o idioma da frase for diferente de português, realizamos a tradução da palavra utilizando a segunda biblioteca, que acessa o Google Tradutor.
- Dicionário com as palavras e suas traduções, desta maneira uma palavra só utilizaria a requisição de tradução uma única vez
- **Complexidade:**
n: palavras

$O(n)$

Limpeza dos Dados



❖ **Número total de palavras**

- com e sem stopwords

❖ **Vocabulário (letra traduzida ou não)**

- Vocabulário sem nenhum processamento
- Vocabulário sem stopwords
- Vocabulário após stemming
- Vocabulário após stemming, sem stopwords

❖ **Complexidade: $O(n)$**

- n = número de palavras

Cálculo de Vocabulário



```
def orengo(song):  
    st = RSLPStemmer()  
    words = filter(lambda x: len(x)>0, song)  
    word_list = list(map(lambda x: st.stem(x), words))  
    return word_list
```

```
def remove_stopwords(doc):  
    stopwords = nltk.corpus.stopwords.words('portuguese')  
    new_doc = list(filter(lambda x: x not in stopwords, doc))  
    return new_doc
```

Cálculo de Vocabulário

#removes stopwords and creates new column

```
df1['no_stopwords'] = df1['lyrics'].apply(lambda x: remove_stopwords(x.split(' ')))
```

#applies stemmer and creates new column

```
df1['orengo'] = df1['lyrics'].apply(lambda x: orengo(x.split(' ')))
```

#total number of words

```
df1['words'] = df1['lyrics'].apply(lambda x: len(x.split(' ')))
```

#total number of words without considering stopwords

```
df1['words_ns'] = df1['no_stopwords'].apply(lambda x: len(x))
```

#removes stopwords from translated text and creates new column

```
df1['ns_pt'] = df1['translated'].apply(lambda x: remove_stopwords(x.split(' ')))
```

#vocabulary

```
df1['unique'] = df1['lyrics'].apply(lambda x: len(set(x.split(' '))))
```

#vocabulary after stemming

```
df1['orengo_unique'] = df1['orengo'].apply(lambda x: len(set(x)))
```

#vocabulary without considering stopwords

```
df1['ns_unique'] = df1['no_stopwords'].apply(lambda x: len(set(x)))
```

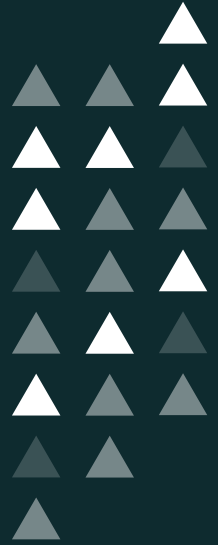
Cálculo de Vocabulário

- ❖ 10 artistas com maior e menor número de palavras
- ❖ 10 artistas com maior e menor taxa (vocabulário)/(número total de palavras)
- ❖ 10 maiores e menores vocabulários, de acordo com diferentes análises
- ❖ Vocabulário dos artistas mais populares
- ❖ Dados estatísticos

Análise dos Dados



Resultados





	words	words_ns	unique	orengo_unique	ns_unique	ns_orengo_unique	unique_pt	orengo_pt	ns_uni_pt	ns_or_pt
count	87.000000	87.000000	87.000000	87.000000	87.000000	87.000000	87.000000	87.000000	87.000000	87.000000
mean	4905.045977	2721.229885	1060.436782	802.471264	966.080460	740.333333	1051.885057	803.712644	959.620690	744.482759
std	2509.384887	1511.020529	498.190650	344.787998	491.355842	344.489601	494.278813	343.986983	487.942411	343.851741
min	2460.000000	1306.000000	482.000000	404.000000	406.000000	347.000000	485.000000	409.000000	409.000000	351.000000
25%	3635.000000	1976.500000	814.000000	617.000000	729.500000	555.000000	813.000000	622.500000	728.000000	565.000000
50%	4120.000000	2306.000000	915.000000	688.000000	826.000000	624.000000	911.000000	692.000000	816.000000	634.000000
75%	5069.500000	2784.500000	1107.500000	857.500000	1006.500000	792.000000	1104.000000	863.500000	1002.500000	800.500000
max	20712.000000	12075.000000	4085.000000	2786.000000	3962.000000	2723.000000	4078.000000	2823.000000	3961.000000	2766.000000

describe()

	artist	words_ns
6	/racionais-mcs/	12075
32	/hungria/	6894
78	/emicida/	6873
16	/projota/	6830
12	/anitta/	5901
71	/tribo-da-periferia/	5685
21	/ludmilla/	4146
75	/vanilda-bordieri/	4096
35	/iza/	4040
10	/charlie-brown-jr/	3794


Maior e menor número de palavras

	artist	words_ns
11	/midian-lima/	1306
22	/raca-negra/	1463
17	/fernandinho/	1668
85	/leandro-leonardo/	1701
80	/amado-batista/	1718
58	/belo/	1727
37	/diante-do-trono/	1761
56	/exalta-samba/	1766
24	/milionario-e-jose-rico/	1782
44	/djavan/	1782

	artist	taxa_palavras
73	/tiao-carreiro-e-pardinho/	0.327984
46	/ze-ramalho/	0.323978
24	/milionario-e-jose-rico/	0.301999
82	/maneua/	0.288627
76	/maria-bethania/	0.288570
39	/chitaozinho-e-xororo/	0.288012
49	/chico-buarque/	0.285634
48	/gilberto-gil/	0.284350
41	/caetano-veloso/	0.283239
44	/djavan/	0.283059

(vocabulário / total de palavras)





	artist	words	words_ns	unique	orengo_unique	ns_unique	ns_orengo_unique	unique_pt	orengo_pt	ns_uni_pt	ns_or_pt	taxa_palavras
0	/marilia-mendonca/	3951	2165	817	602	735	545	821	613	740	560	0.206783
1	/roberto-carlos/	5169	2706	1038	727	941	664	1038	742	944	681	0.200813
2	/gusttavo-lima/	4414	2306	844	679	752	613	844	681	754	619	0.191210
3	/jorge-e-mateus/	4120	2162	765	591	681	529	767	603	684	544	0.185680
4	/aline-barros/	3826	2081	747	581	650	520	750	585	657	529	0.195243
5	/zeze-di-camargo-e-luciano/	4046	2123	836	622	745	561	834	625	742	565	0.206624
6	/racionais-mcs/	20712	12075	4085	2786	3962	2723	4078	2823	3961	2766	0.197229
7	/anavitoria/	4885	2614	851	673	766	613	853	689	771	632	0.174207
8	/legiao-urbana/	6255	3243	1524	1088	1403	1022	1525	1094	1411	1033	0.243645
9	/ze-neto-e-cristiano/	3482	1856	767	583	687	524	773	594	693	538	0.220276

*Vocabulário dos 10
artistas mais populares*

	artist	ns_or_pt
6	/racionais-mcs/	2766
78	/emicida/	2097
32	/hungria/	1613
16	/projota/	1454
71	/tribo-da-periferia/	1399
49	/chico-buarque/	1099
73	/tiao-carreiro-e-pardinho/	1094
45	/kamaitachi/	1070
41	/caetano-veloso/	1034
8	/legiao-urbana/	1033

Top 10: Artistas com maior vocabulário
após tradução, stemming e
desconsiderando stopwords.

	artist	ns_or_pt
11	/midian-lima/	351
22	/raca-negra/	390
17	/fernandinho/	391
25	/gabriela-rocha/	450
79	/andre-valadao/	483
69	/padre-marcelo-rossi/	490
37	/diante-do-trono/	517
4	/aline-barros/	529
84	/eyshila/	530
9	/ze-neto-e-cristiano/	538

Top 10: Artistas com menor vocabulário
após tradução, stemming e
desconsiderando stopwords.

Maiores e menores vocabulários

- ❖ Foram feitas análises de idade para os 10 artistas com maior vocabulário e para os 10 com menor vocabulário.
- ❖ Metodologia utilizada:
 - Ano do primeiro álbum do artista subtraído do ano atual.

Idade dos Artistas

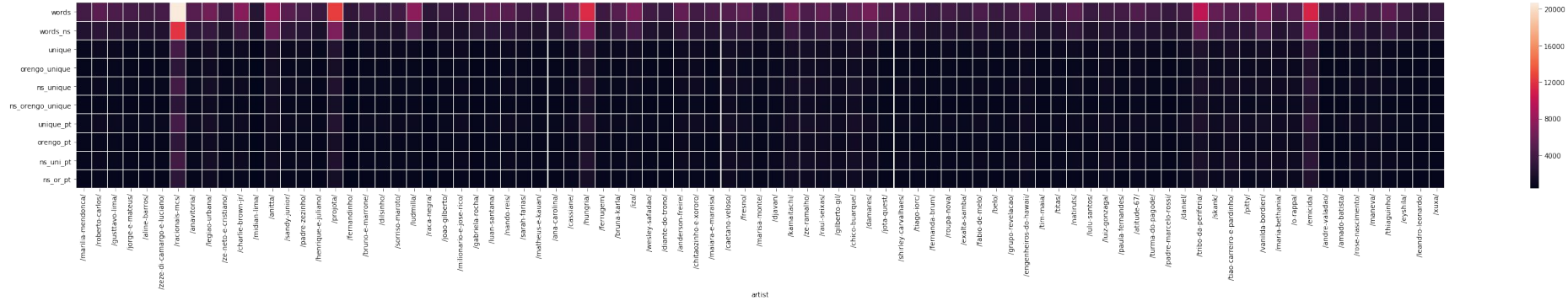
Artistas	Estimativa
/racionais-mcs/	29
/emicida/	10
/hungria/	10
/projota/	10
/tribo-da-periferia/	16
/chico-buarque/	53
/tiao-carreiro-e-pardinho/	18 65
/kamaitachi/	2
/caetano-veloso/	52
/legiao-urbana/	34

Top 10: Artistas com maior vocabulário
após tradução, stemming e
desconsiderando stopwords.

Idade dos Artistas

Artistas	Estimativa
/midian-lima/	0 10
/raca-negra/	23
/fernandinho/	18
/gabriela-rocha/	7
/andre-valadao/	15
/padre-marcelo-rossi/	21
/diate-do-trono/	21
/aline-barros/	21
/eyshila/	23
/ze-neto-e-cristiano/	5

Top 10: Artistas com menor vocabulário
após tradução, stemming e
desconsiderando stopwords.



Vida Loka

(Part. w/ Racionais MC's)

Nada como um dia após o outro dia

Primeira total, mais um ano se passando aí
Grupos e Deus a gente tá com aquele "ah", "hauu"? Com certeza
Muita coisidade na quebrada
Dinheiro no bolso, sem miséria
É "loka", "vamo" brincar a dia de hoje
O amor tá só, perdendo a Deus
A vida é "loka"

Deixa eu falar pro "tê"
Tudo, tudo, tudo vai, tudo é fase imbuído
Logo mais "vamo" amparar no mundo
De corréio de elite, tá quites
Põe no pulso, logo brelling

Que tal, tá bom?

De tudo bausch e bomb, bomba branca e vinho
Champagne para o ar, que é pra abrir nossos caminhos
Poder e o dia tá e eu só tá a sobrevivência
Poder tá, rit, mas não desencana lá não

É só questão de tempo, o fim do sofrimento
Um arde por guerras, Zé povinho eu lamento
Vermes que só faz peso na terra

Tira o "zelo"

Tira o "zelo", vê se me ensina
Eu durmo pronto pra guerra
E eu não era assim, eu tenho dódo
É sei que é mau pra mim
Fazer o que se tá assim
Vida "loka" cobalado
O cheiro é de pólvora
E eu prefiro tocas

E eu que... e eu que

Sempre sou um lugar
Gratidão e tempo, assim verde como o mar
Cercas brancas, uma seringueira com balanço
Distanciando pipa cercado de crianças

Now... how brown

Acorda sangue bom
Aqui é Chapão Redondo "tu"
Não Pokemon
Zona sul é invés, é estresse concentrado
Um coração ferido pra metro quadrado

Quanto mais tempo eu vou resistir
Por que eu vi meu lado sou na LULA
Meu ano do perdão foi bom
Mas tá fraco
Culpas dos mundos do espírito opaco

Eu queria ter pra testar e ver
Um milide, com glória, fama
Introduzido em Jacaré
Se é isso que "loka" quer
Vem pegar

Jogar num rio de menta e ver vários "pau"
Dinheiro é foda
No mão de favelado, é "mô guri"
No crime, vários pedra na venta, eu fuma

Eu vou jogar pra ganhar

O meu nome, vai e vem
Poder quem tem, tem
Não como o "loka" em ninguém
O que tem que ter
Seri meu
O mundo não estrai
Vai recomeçar com Deus

Imagina "loka" de Audi
Ou de Citroën
Indo aqui, indo ali
Se "loka"

De vai e vem
No lado do Aquino, vou color
No pedreiro do São Bento
No fundo, no pédo
Sendo feroz

Develo sair
O luar representa
Ouvindo Casiano, na
Do "gueto" não "gueto"
É, mas se não der

"Neger"
O que é que tem?
O impulsiona e não aqui
Junto ano quem é
É o caminho
Os Nervos não existe
É uma trilha extrema
É em meio a uma trilha

Quanto se paga
Eu vi sua vida agora
É nunca mais vê seu paiete
Achando
É o caso, dá o caso
Uma glória e uma foi
Sóde negro da janela
Me e com "negro"

Quanto é mi "gru"
O que é o guerreiro da
O primeiro tá em homem
Deus é a luz

Inquieto Zé povinho
Adebrava a cruz
Um coração ferido
Cuidado com Deus

Wé

Assi do do segundo amparado
Sóde e perdido
É Dinis o bandido
É Dinis o bandido

É Dinis o bandido
Ameco na hora
É

Dinis, primeiro veio "loka" da história

Eu digo
glória... glória
Se que Deus tá aqui

É só quem é
Só quem é o sermão

É meu "gueto" de fé
Quero suar, quero suar
É meu "gueto" de fé
Quero suar, imbuído

Programado pra nome "loka"
Certo X... certo X... Certo X... Certo X... Certo X... Certo X...

Primeira

Não é questão de luxo
Não é questão de cor
É questão que fortuna
Algo a sofrer

Não é questão de presa
Nem cor
A ideia é essa
Módica traz tristeza e vice-versa
Inconscientemente
Vem na minha mente inteira

Uma loja de téis
O olhar do paranoide feliz
De poder comprar
O azul, o vermelho
O branco, o amarelo
O estoque e a moderação

Não importa
Dinheiro é tudo
É abre as "portas"
Dos "casinos" de arca que quiser

Preto e dinheiro
São palavras rivais?
É?
Então mostra pra esses cú
Como é que faz

O seu encontro foi dramático
Como o blues antigo
Mas de estilo
Me perdido de bandido

"Tempo" tá "pensar"
Quer pensar?
Que se quer?
Viver pouco como um rei
Ou então muito, como um Zé?

Às vezes eu acho
Que todo preto como eu
Só quem um tempo no mundo
Só eu

Sem luxo, descalço, nadar num rio
Sem fome, segundo as frutas na cesta

Altruída, é o que eu acho
Quero também
Mas em São Paulo
Deus é uma nota de 100
Vida "loka"

Porque o guerreiro de fé nunca gera
Não agreda o injusto e não amenda
O rei do meu foi traidor e sanguinário nessa terra
Mas morreu como um homem, é o prêmio da guerra
Mas o

Confirmei por se precisar alugar no próprio sangue
Assim será
Nosso espírito é imortal, sangue do meu sangue
Entre o corte do estado e o perfume da rosa
Sem menção honrosa, sem mensagem

A vida é "loka", "negro"
É meu eu-tô-de-pacagem

A Dinis o primeiro
Soude guerreiro

Dinias...

Racionais MC's X Zé Neto e Cristiano

Bebida na Ferida

Zé Neto e Cristiano

Esquece O Mundo La Fora

Parecia um bom negócio

Vc se desfazer de mim

Só não contava com os juros

Que a saudade no futuro

la te cobrar por mim

Agora tá de standby

Vai fazer isso com os outros vai

Tudo que vai volta ai ai ai

Quem abre também fecha a porta

Te perder

Foi a dor mais doída

Que eu senti na vida

Sem você

Joguei bebida na ferida

Que bom que o álcool cicatriza

Comparação de letras

Conclusões



- ❖ A tradução não afetou significativamente o resultado do cálculo de vocabulário
- ❖ Racionais tem o maior vocabulário, mas não aparece no TOP 10 vocabulário/número de palavras
- ❖ Hip Hop/Rap são os gêneros com maior vocabulário (5 primeiros artistas)
- ❖ Gospel possui grande concentração de artistas no ranking de menores vocabulários
- ❖ Artistas mais populares possuem vocabulário abaixo da média

Conclusões

- ❖ Daniels, M. The Largest Vocabulary in Hip Hop. 2014, atualizado pela última vez em 21 de Janeiro, 2019. Disponível em: <https://pudding.cool/projects/vocabulary/index.html>
- ❖ Language Identification tool (LangID). Disponível em : <https://github.com/saffsd/langid.py>.
- ❖ Matplotlib, a Python 2D plotting library. Disponível em: <https://matplotlib.org/>
- ❖ Mayer, Rudolf & Neumayer, Robert & Rauber, Andreas. (2008). Rhyme and Style Features for Musical Genre Classification by Song Lyrics.. ISMIR 2008 - 9th International Conference on Music Information Retrieval. 337-342.
- ❖ Natural Language Toolkit. Disponível em: <https://www.nltk.org/>
- ❖ Numpy. Disponível em: <https://www.numpy.org/>
- ❖ Orengo, V. M; Huyck, C. A Stemming Algorithm for the Portuguese Language. Eighth International Symposium on String Processing and Information Retrieval (Spire 2001), p. 186-193, 2001.
- ❖ Pandas, Python Data Analysis Library. Disponível em: <https://pandas.pydata.org/>
- ❖ Regular expression operations. Disponível em: <https://docs.python.org/3/library/re.html>
- ❖ Rodrigues, MF. E assim a música caminhou pelo Brasil. Revista Mosaico. 2018 Jul./Dez.; 09 (2): 32-34.
- ❖ RSLP Stemmer. Disponível em: https://www.nltk.org/_modules/nltk/stem/rslp.html
- ❖ Seaborn: statistical data visualization based on matplotlib. Disponível em: <https://seaborn.pydata.org/>
- ❖ Textblob, a Python (2 and 3) library for processing textual data. Disponível em : <https://textblob.readthedocs.io/en/dev/contributing.html>
- ❖ Vagalume. Disponível em: <https://www.vagalume.com.br/>

Referências



Scan me

Link do projeto no GitHub:

https://github.com/felipery03/music_nlp_analysis/

Apresentação:

<http://abre.ai/vocab> ou

<https://bit.ly/2TToSv1>

para saber mais

