Analyzing Lyrical Data from Historically Popular Songs

by: David Ash, Brian Blakely, and Jackson Conrad
even though we won't admit to it later

Goals

I. Web scrape song data from Billboards

- Analyze lyrical data from the top 100 songs from 1959 - 2019

3. Use natural language processing (NLP) to help with analysis

- Make pretty graphs for stuff

D. Design a good presentation for when 2.5/4 of those things fail horrifically

Process

0) Modify a web scraper to get the data set

1. Wait 6 hours
2. Now that you finally have your data set
3. Start analysis
4. Work for 3 hours processing the data
5. Realize your data set has full books instead of lyrics in some cases
6. Try to modify your data set to work correctly and not have books in it
7. Run a NLP to reprocess the data
8. Wait 2.5 hours
9. Watch in despair as it throws an error on the last song
10. Give up and start a new project
11. Wait 2.5 hours
12. Get locked outside
13. Give up on doing the new project and go back to the old project which works now

Summary: Progress made during the first 10 hours:

[this slide intentionally left blank]

After that

The very first data we successfully processed:

Data is from approximately bottom 50 of 2000s - 2019

Idk why there are so many single quotes either

cont.

14. Realize it only works because the frequency measure gives arbitrary values for our purposes

15. Skew the data by only looking at the graphs that look nice

16. Make this presentation

17. Give the presentation

19. ???

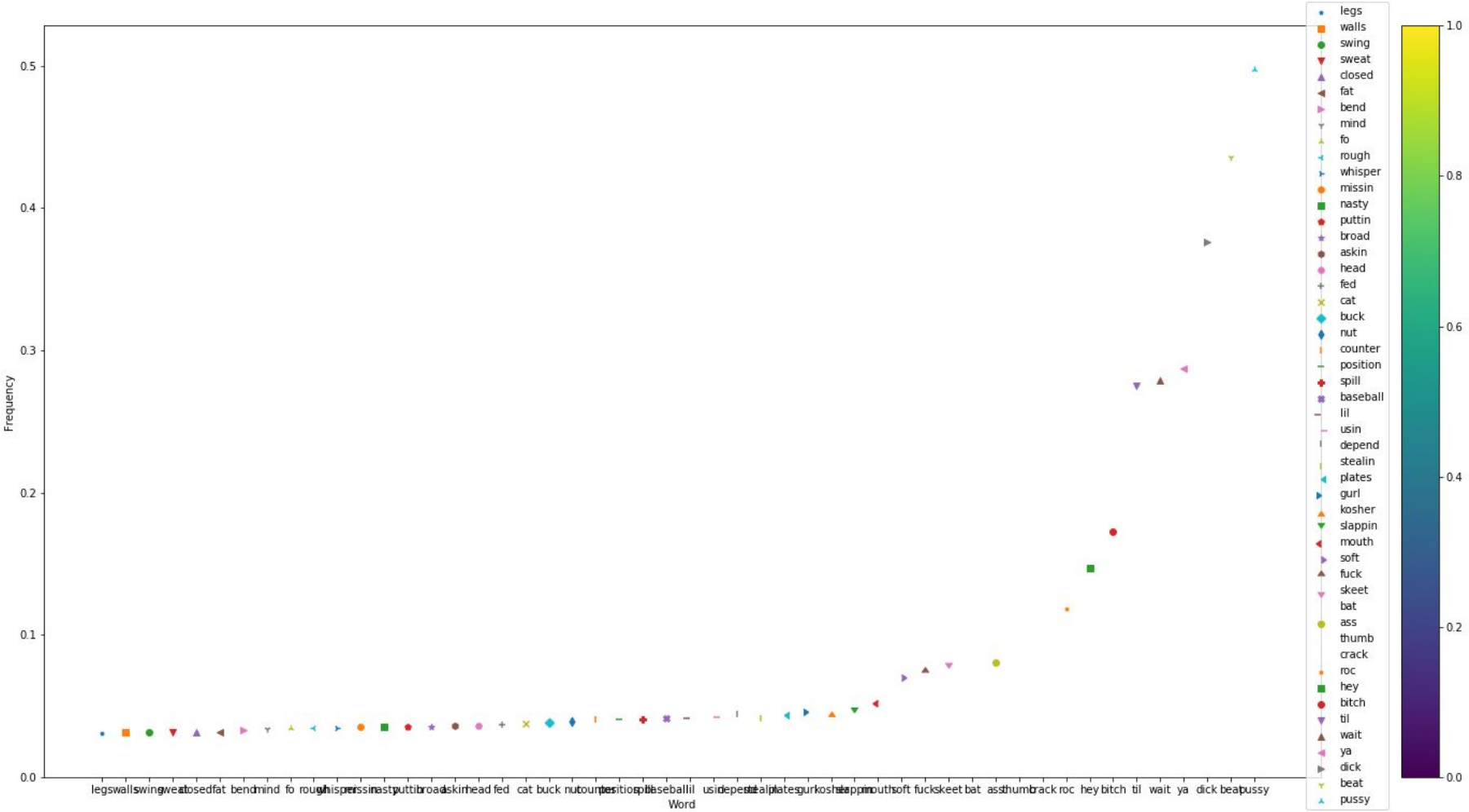
20. Win 1st place

21. Skip class tomorrow because I'm tired and everything hurts

Actual data (not a joke this time) (last time wasn’t a joke either fyi)

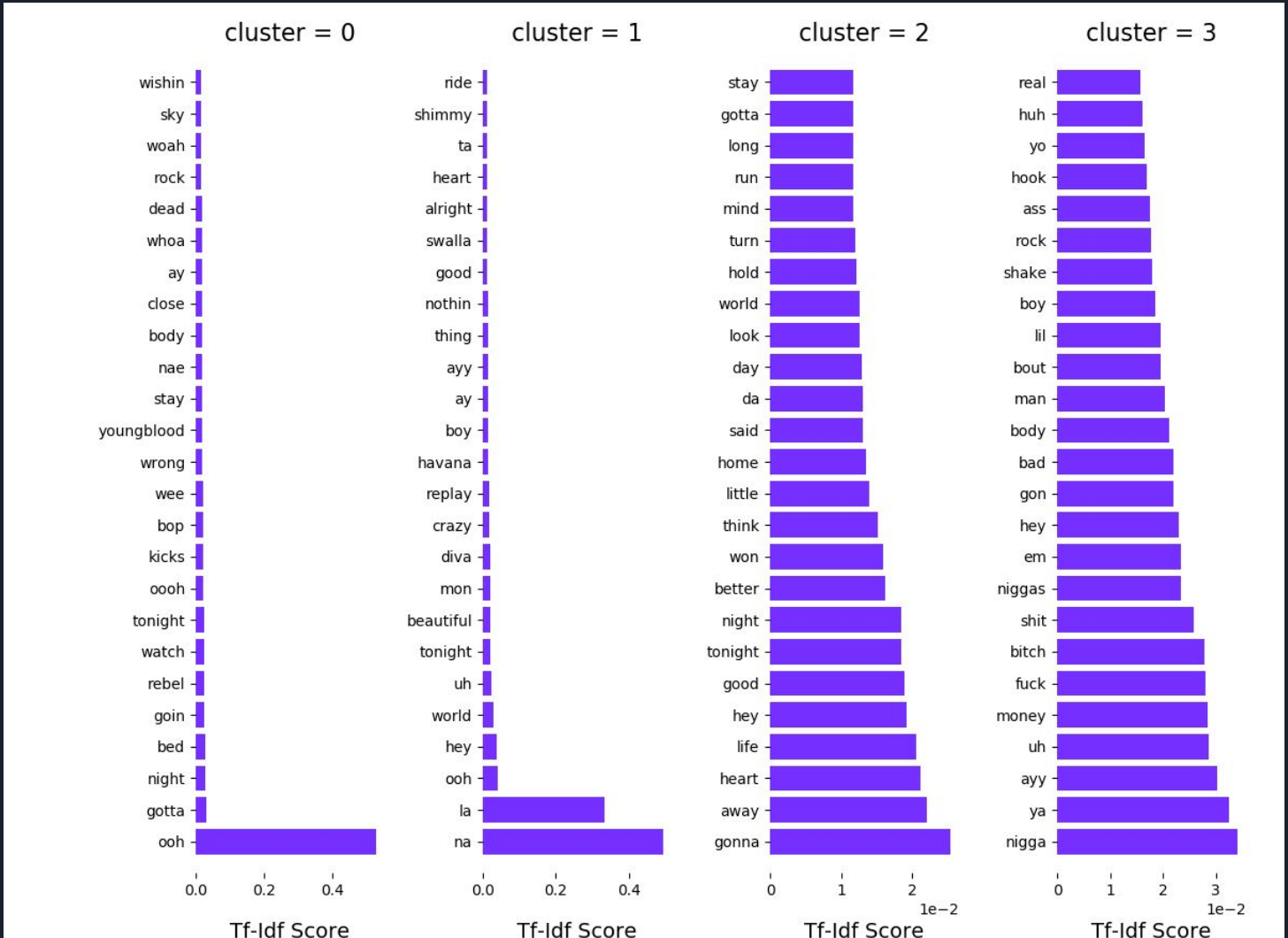
please excuse the profanity

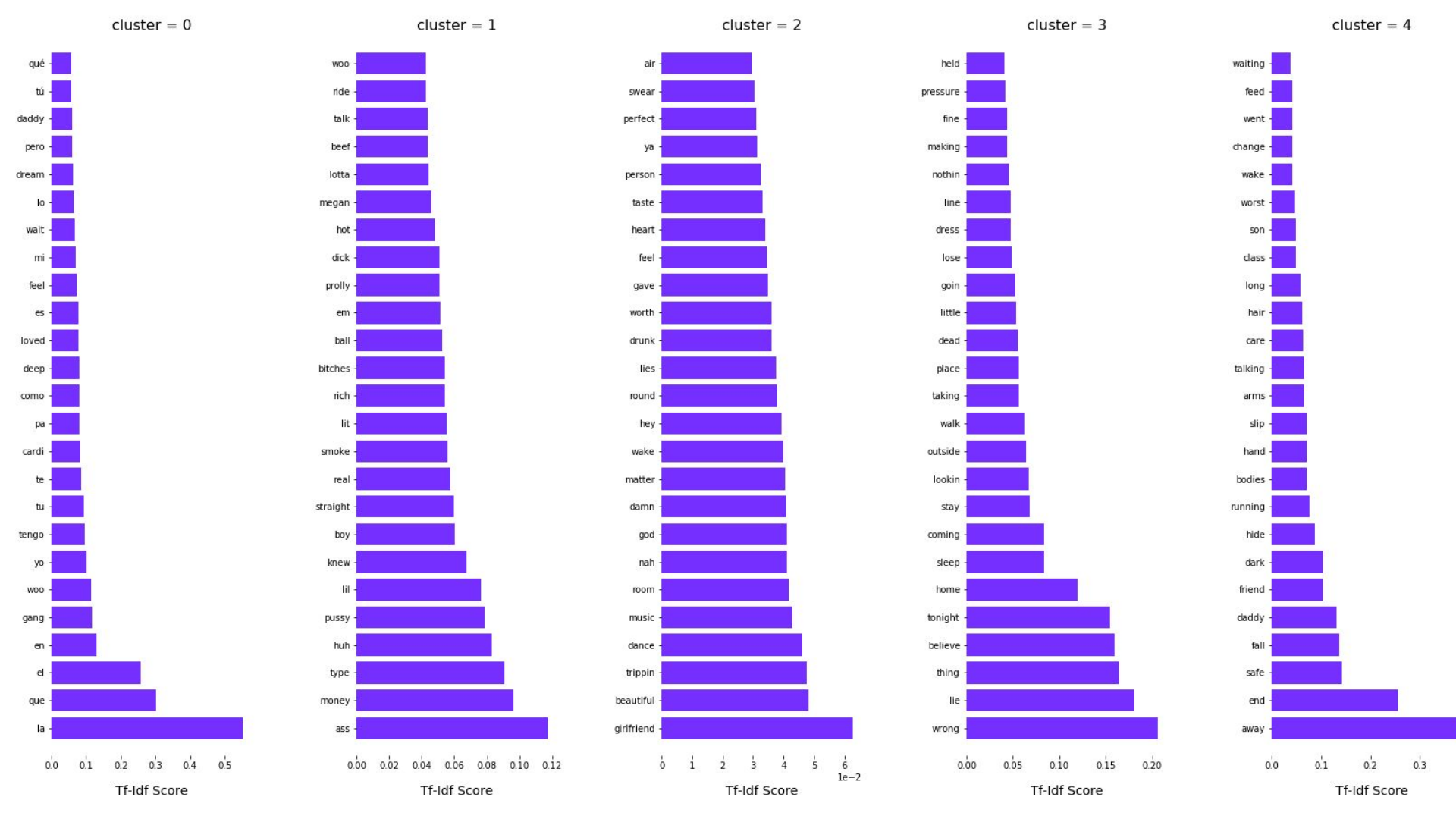
we didn’t want to skew the data

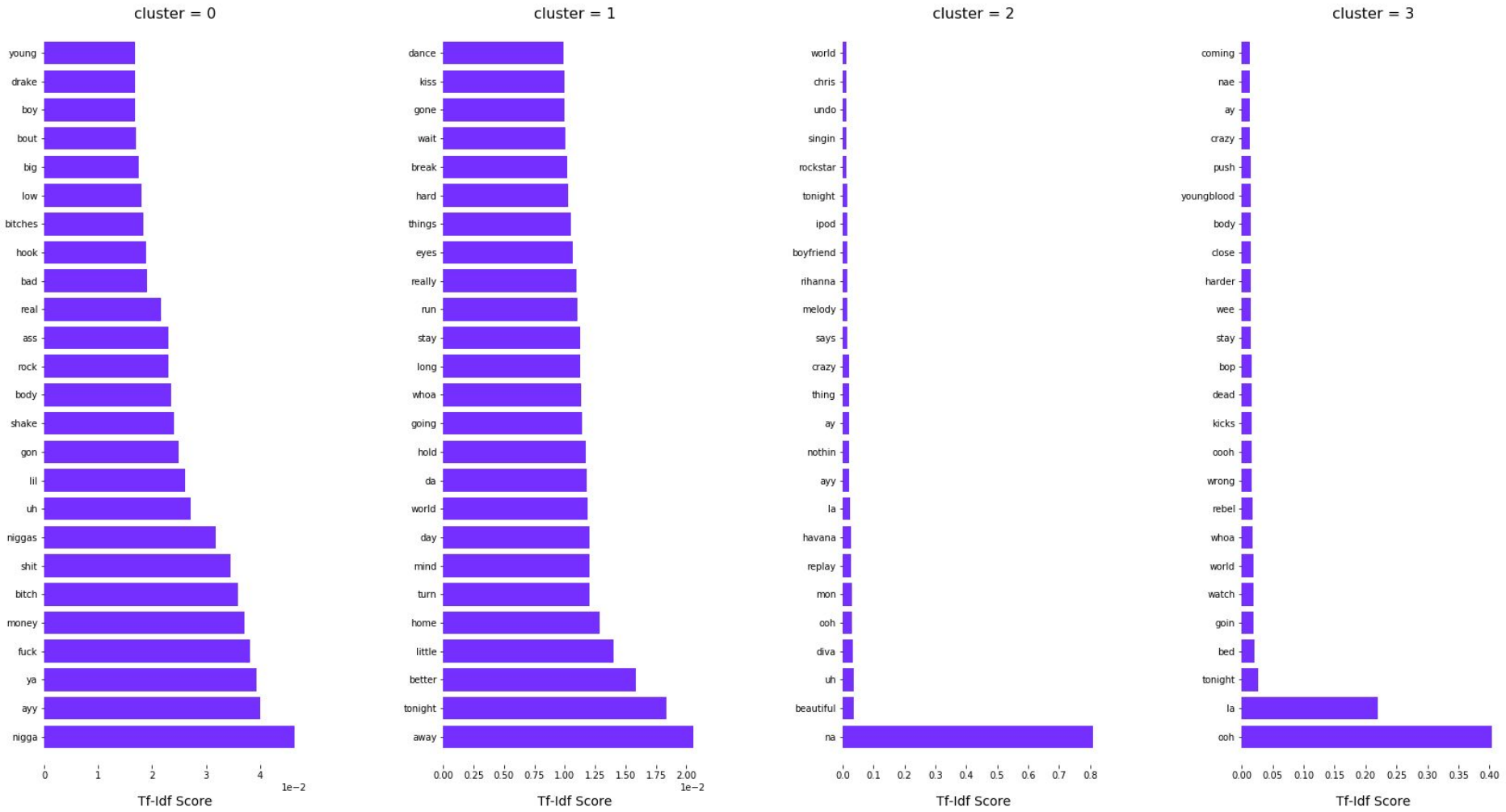

Frequency
0.0
0.1
0.2
0.3
0.4
0.5
legs walls swing sweat closed fat bend mind fo rough whisper missin nasty puttin broad askin head fed cat buck nut counter position spill baseball lil usin depend stealin plates gurl kosher slappin mouth soft fuck skeet bat ass thumb crack roc hey bitch til wait ya dick beat pussy
Word
legs
walls
swing
sweat
closed
fat
bend
mind
fo
rough
whisper
missin
nasty
puttin
broad
askin
head
fed
cat
buck
nut
counter
position
spill
baseball
lil
usin
depend
stealin
plates
gurl
kosher
slappin
mouth
soft
fuck
skeet
bat
ass
thumb
crack
roc
hey
bitch
til
wait
ya
dick
beat
pussy
1.0
0.8
0.6
0.4
0.2
0.0

2005 - 2019

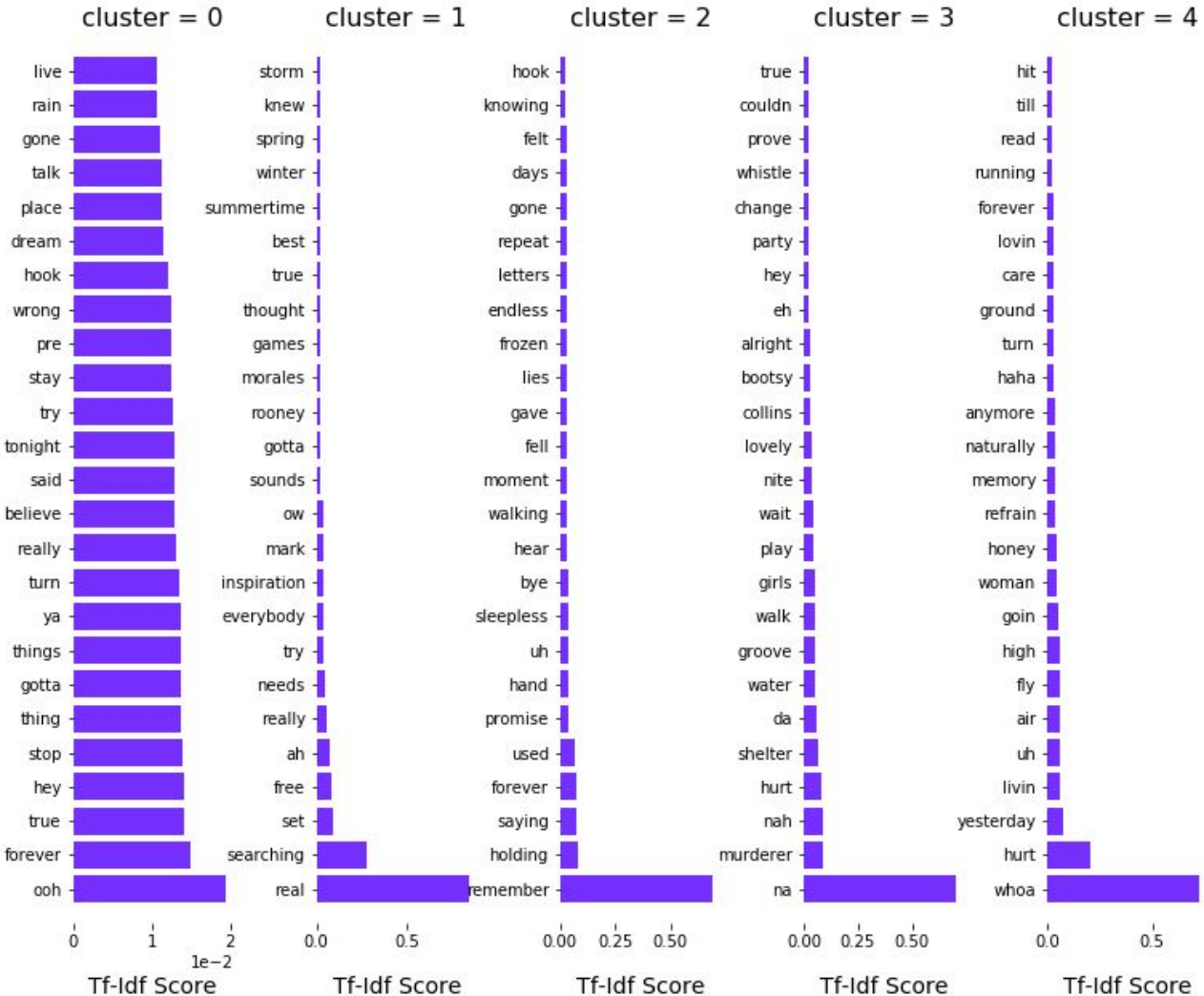
(clustering algorithm grouped bad words with each other)




cluster = 0
qué
tú
daddy
pero
dream
lo
wait
mi
feel
es
loved
deep
como
pa
cardi
te
tu
tengo
yo
woo
gang
en
el
que
la
0.0 0.1 0.2 0.3 0.4 0.5
Tf-Idf Score
cluster = 1
woo
ride
talk
beef
lotta
megan
hot
dick
prolly
em
ball
bitches
rich
lit
smoke
real
straight
boy
knew
lil
pussy
huh
type
money
ass
0.00 0.02 0.04 0.06 0.08 0.10 0.12
Tf-Idf Score
cluster = 2
air
swear
perfect
ya
person
taste
heart
feel
gave
worth
drunk
lies
round
hey
wake
matter
damn
god
nah
room
music
dance
trippin
beautiful
girlfriend
0 1 2 3 4 5 6
1e−2
Tf-Idf Score
cluster = 3
held
pressure
fine
making
nothin
line
dress
lose
goin
little
dead
place
taking
walk
outside
lookin
stay
coming
sleep
home
tonight
believe
thing
lie
wrong
0.00 0.05 0.10 0.15 0.20
Tf-Idf Score
cluster = 4
waiting
feed
went
change
wake
worst
son
class
long
hair
care
talking
arms
slip
hand
bodies
running
hide
dark
friend
daddy
fall
safe
end
away
0.0 0.1 0.2 0.3
Tf-Idf Score


cluster = 0
young
drake
boy
bout
big
low
bitches
hook
bad
real
ass
rock
body
shake
gon
lil
uh
niggas
shit
bitch
money
fuck
ya
ayy
nigga
0
1
2
3
4
1e−2
Tf-Idf Score
cluster = 1
dance
kiss
gone
wait
break
hard
things
eyes
really
run
stay
long
whoa
going
hold
da
world
day
mind
turn
home
little
better
tonight
away
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
1e−2
Tf-Idf Score
cluster = 2
world
chris
undo
singin
rockstar
tonight
ipod
boyfriend
rihanna
melody
says
crazy
thing
ay
nothin
ayy
la
havana
replay
mon
ooh
diva
uh
beautiful
na
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Tf-Idf Score
cluster = 3
coming
nae
ay
crazy
push
youngblood
body
close
harder
wee
stay
bop
dead
kicks
oooh
wrong
rebel
whoa
world
watch
goin
bed
tonight
la
ooh
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Tf-Idf Score

Mid 1980s


cluster = 0
live
rain
gone
talk
place
dream
hook
wrong
pre
stay
try
tonight
said
believe
really
turn
ya
things
gotta
thing
stop
hey
true
forever
ooh
0
1
2
1e−2
Tf-Idf Score
cluster = 1
storm
knew
spring
winter
summertime
best
true
thought
games
morales
rooney
gotta
sounds
ow
mark
inspiration
everybody
try
needs
really
ah
free
set
searching
real
0.0
0.5
Tf-Idf Score
cluster = 2
hook
knowing
felt
days
gone
repeat
letters
endless
frozen
lies
gave
fell
moment
walking
hear
bye
sleepless
uh
hand
promise
used
forever
saying
holding
remember
0.00
0.25
0.50
Tf-Idf Score
cluster = 3
true
couldn
prove
whistle
change
party
hey
eh
alright
bootsy
collins
lovely
nite
wait
play
girls
walk
groove
water
da
shelter
hurt
nah
murderer
na
0.00
0.25
0.50
Tf-Idf Score
cluster = 4
hit
till
read
running
forever
lovin
care
ground
turn
haha
anymore
naturally
memory
refrain
honey
woman
goin
high
fly
air
uh
livin
yesterday
hurt
whoa
0.0
0.5
Tf-Idf Score


EVERLASTING'
UNDERSTAND'
FINISHED'
HOPED'
HEART'
WAIT'
FILLED'
FAR'
CHARMING'
COLLISION'
FILES'
KISS'
ROMANTIC'
MAN'
JULIET'
FIGURES'
PRINCE'
FINNA'
EARTH'
MEAN'
SWEAR'
FIRED'
HUG'
FOREVER'
PROMISE'
TENDERNESS'
CINDERELLA'
PRECIOUS'
FIRES'
KISSED'
DREAM'
FELT'
OPEN'
FIRESIDE'
FINALLY'
EYES'
CHANCE'
WORLD'
'HOLDING'
ROMEO'
READ'
COMES'
MOVIES'
TRUE'
RESCUE'
HEAVEN'
LOOK'
WAITING'
DAY'
FORGET'

1990s - 2000s


TOMORROW'
PLANNIN'
YO'
BEATS'
GOING'
SLEEPIN'
FLAME'
FEET'
WARM'
LEAVE'
DANCING'
SKIES'
SUN'
WANNA'
TIGHT'
WAKE'
SEND'
LADY'
DREAMIN'
TURNED'
SHINE'
BUGGIN'
LOVIN'
SOLO'
ACT'
INSTEAD'
BRIGHTER'
TONIGHT'
GREY'
FRIEND'
POST'
SKY'
FIGHT'
DORIS'
MISS'
BED'
DARE'
HIT'
SPARK'
BRIGHT'
FOOL'
BRAIN'
STARTS'
MINUTE'
BOOM'
TIL'
BANG'HANGING'
CRUEL'
HIGH'

1980s - 1990s

What you could do with this project

- ★ Collect data about the frequency of…
  - Parts of speech (nouns, verbs, etc.)
  - Different kinds of nouns (People, nationalities, products, etc.)
  - Bad words (I'm not typing them)
- ★ Divide the data by…
  - Time (specific years, decades, etc.)
  - Genre
  - Artist
  - Lyrical content (length, unique words, etc.)
- ★ Create cool graphs with the data collected
  - Stuff we wanted to do but couldn't do in time includes…
    - Collect information on the usage of product and company names over time
    - Display information about use of profanity over the decades
    - Display most popular verb by decade
    - How closely words follow Zipf's Law