SI 618 Final Project Report – Part A

Music Lyrics Analysis

Huanchen Lu

1. Motivation

As an important part of our life and culture, music is a way for artists to convey their feelings and values, and popular music can tell us what people want to hear. For this project, I mainly focused on analyzing the change of music lyrics over past 55 years. Specifically, I extracted top 100 songs of each decade from The World's Music Charts and fetched lyrics of each song from SongLyrics.

My motivation for this project is very simple – I love music and I also love data. Also, I think this project is useful because by studying popular music lyrics, I could gain insights about the society's values and feelings.

2. Data Sources

a. The World's Music Charts (http://tsort.info/music/ds1960.htm)

The World's Music Charts is a collection of world music chart information. From the website's decade-end music charts, I fetched the top 100 songs of each decade to form my first dataset. The dataset contained several important variables, including song rank, song title, artist and year. The dataset covered 5 decades: 1960s, 1970s, 1980s, 1990s and 2000s. It also contained top 100 songs from 2011 to 2014.

b. SongLyrics (http://www.songlyrics.com/)

After I obtained all the songs from The World's Music Charts, I used SongLyrics, a website for searching lyrics of songs, to get lyrics for each song and combined all the lyrics with my first dataset.

3. Data Manipulation

a. Data Collection

Since The World's Music Charts and SongLyrics do not provide APIs for me to fetch their data, I used urllib2 and BeautifulSoup to parse the web pages and grabbed the data to create my dataset. The fetch_top_list() function in data.py was used for grabbing

songs from The World's Music Charts. Each relevant variable was stored in a list. And after successfully grabbing one song, get_lyrics() function was called in fetch_top_list() to extract the lyrics by song title and artist name from SongLyrics. Then, the record will be write to csv file. Here is a screenshot of my dataset.

	1	2	3	4	5
1	Position	Artist	Song Title	Year	Lyrics
2	1	Elvis Presl	Are You Lonesome Tonight?	1960	Are you lonesome tonight? Do you
3	2	Elvis Presl	It's Now Or Never	1960	It's now or never
4	3	Chubby Ch	The Twist	1960	Come on, baby, let's do the twist
5	4	Percy Faith	Theme From 'A Summer Place'	1960	There's a summer place
6	5	Rocco Gra	Marina	1960	you're hard to hug
7	6	The Everly	Cathy's Clown	1960	Don't want your love anymore
8	7	The Drifter	Save the Last Dance For Me	1960	You can dance
9	8	Roy Orbiso	Only The Lonely (Know The Way I Feel)	1960	Dum dum dummy doo wah
10	9	Sam Cook	Wonderful World	1960	Don't know much about history
11	10	Brenda Le	I'm Sorry	1960	I'M SORRY SO SORRY
12	11	Elvis Presl	Stuck On You	1960	You can shake an apple off an

b. Data Pre-processing

Although The World's Music Charts is an extremely clean data source with no missing data, SongLyrics sometimes would return empty lyrics. For these missing lyrics, I ignored all of them. And I also removed all the stop words from the lyrics.

4. Analysis and Visualization

a. Term Frequency Calculation

I calculated the term frequencies for lyrics of each decades and stored the result in dictionary. After the calculation was done, I wrote the frequency dictionary to txt files. Here are some results of the term frequencies of 1960s:

words	frequency
love	193
yeah	155
baby	132
day	71
back	70
hey	65
good	63
time	53
night	48
girl	48
jude	46
make	45

b. Tf-idf Weight Calculation

The tf-idf weight of each term was calculated as:

$$\log(term\ frequency + 1) \times (\log\left(\frac{number\ of\ total\ documents}{document\ frequency}\right) + 1)$$

I calculated the tf-idf weight for each term of lyrics of each decades and stored the result in dictionary. After the calculation was done, I wrote the frequency dictionary to txt files. Here are some results of the tf-idf weights of 1960s:

words	tf-idf
jude	10.748686025
submarine	9.2011837797
paperback	8.22016540165
wooly	7.90965018161
excitations	7.90965018161
bully	7.56022479207
monday	7.52041688092
aquarius	7.16072165678
writer	6.83748143236
eloise	6.69434683405
grapevine	6.69434683405
speedy	6.42826373707
gonzales	6.42826373707
satisfaction	6.42826373707

c. Visualization

I used Tagul (https://tagul.com/) to generate word clouds for my results.

i. Term Frequency Visualization

1960s:


feeling
loves
heard
girl
day
vibrations
tonight
twist
money
life
easy
dance
downtown
words
make
paperback
goodbye
ride
time
heart
honey
mine
sugar
back
lovin
days
hey
reach
woman
wan
night
rolling
good
walk
mary
writer
dum
moon
sunshine
live
leave
yeah
hold
fire
pretty
head
huh
bad
washed
buy
wooly
arms
feel
jude
people
eyes
sweet
baby
blue
hard
yellow
monday
hand
love
bit
gim
long
bully
submarine
find
give
run
happy
lonely
things
man
care
bop
ticket
home
believer
stand
mind
darling
friend

1970s:


stuff
hot
face
softly
feeling
ring
eyes
place
rock
dancing
night
day
leave
mind
arms
alive
gave
light
thing
stay
stop
survive
song
man
freak
goodbye
woman
people
good
hey
born
heart
music
sweet
feel
sun
force
give
back
lady
long
made
hard
waterloo
baker
happy
singing
life
home
word
till
baby
lay
bell
living
deep
times
cool
whoa
live
lord
time
die
wrong
ai
yeah
boys
words
girl
beautiful
sky
knock
world
breaking
stayin
love
bye
run
joy
young
fever
show
daddy
ooo
lola
make
find

1980s:

love
yeah
baby
time
feel
call
make
red
heart
night
life
give
day
hand
dance
beat
girl
wo
rock
dust
turn
world
amadeus
eyes
stop
back
good
crazy
jump
ooh
roll
end
remember
wine
ah
la
lord
round
cry
shack
tonight
boy
bites
left
find
fun
hear
west
man
girls
side
karma
run
hey
thing
faith
live
prayer
forever
long
relax
celebrate
bang
head
eileen
play
mind
war
times
worry
hold
inside
easy
close
leave
called
move
show
wake

1990s:

love
baby
yeah
life
back
feel
make
heart
time
give
night
good
miss
jump
dance
daddy
hurt
day
uh
nah
stop
snow
find
shake
ice
fly
man
boy
black
sign
eyes
world
ai
du
la
shout
call
girl
body
mind
thinking
livin
free
end
hold
touch
live
dream
feeling
shoe
barbie
rain
sweet
true
kiss
light
lonely
long
bop
mine
licky
inside
song
scrub
ground
close
move
hear
yay
joe
forever
pain
show
belong
men
face
god
blue
break
mac
hard
murderer
heaven
wrong
words
sade
made

2000s:

2010 – 2014:

As we can see from the word clouds above, baby, love, yeah are the most frequent words in all decades.

ii. Tf-idf Weights Visualization

1960s:

sunshine
heba
back
dummy
eloise
unknown
hela
clover
wah
bully
loretta
gladly
love
downtown
penny
sender
sock
day
boots
dock
dum
aquarius
believer
washed
twist
time
lonesome
haer
vibrations
crimson
marina
sleepy
disguise
writer
suspicious
ticket
banker
baby
excitations
speedy
trap
ahhhhh
san
good
wastin
honky
submarine
pickin
potato
helloa
gonzales
tripper
manage
mary
painted
paperback
delilah
twistin
whiter
francisco
yeah
crucify
twisty
puppet
judy
wooly
doo
wanting
runaway
ruin
monday
daydream
tonk
na
massachusetts
hey
log
cha
jude
heals
dizzy
la
girl
ca
strawberry
night
yay
hatty
tavern
whirlpool
yellow
satisfaction
grapevine
suburban
mashed

1970s:

rosie
paloma
wop
meanest
na
kintyre
postman
remembered
chevy
love
angie
galileo
knock
crumble
suzie
heathcliff
da
mistrust
ribbon
babylon
freak
hare
born
cathy
stayin
gon
tragedy
expert
smoker
mist
lovey
mucho
boogie
ballroom
levee
ma
alive
rye
evenin
yeah
kung
blanca
dah
krishna
sharona
hallelujah
baker
breaking
cracklin
wept
toker
fernando
fever
dovey
gurur
baby
survive
ca
blitz
grease
fishes
shaft
venus
waterloo
heart
hot
champions
wuthering
mull
force
died
stop
bye
facing
picker
bell
timing
hoo
sat
jive
fu
lotta
zion
closes
joy
chic
lola
night
dee
time
grinner
singing
troubled
lord
crocodile
wood

1980s:


time
susie
na
thee
poca
life
make
preach
yahoo
bang
um
baby
bites
chorando
aonde
rattle
dia
drives
karma
give
chameleon
doves
red
babaluma
zing
shack
odds
wan
toora
respectfully
loo
night
nothings
sledgehammer
feel
prayer
amadeus
dance
egyptian
ebony
bette
dutchie
esa
99
bamba
heap
beat
instinctively
ghostbusters
bam
amor
davis
celebrate
eileen
rizla
plunder
yeah
west
dep
loora
turnaround
balloons
funkytown
flair
papa
gon
vale
gracia
tease
batman
love
thriller
defeated
lembrar
jitterbug
recordacao
call
biddly
claims
fina
prance
pon
ao
tiger
jukebox
rye
tainted
arriba
heart
celebration
rival

1990s:


jump
undress
mr
sayes
janet
injustice
barbie
belongs
grey
miggida
rhyme
ca
vida
lakes
gon
monica
love
mocha
blaze
dop
life
da
mmm
make
dee
dio
uh
zigazig
joyride
duba
yeah
macarena
shake
deserts
hosanna
murderer
mac
wan
unbreak
ba
unnatural
greens
give
join
sade
joe
source
time
defenders
duh
night
baby
pressurin
snow
compares
bop
vogue
torn
gangsta
moth
vanilla
licky
ice
rpt
lift
shoe
heart
nay
rub
lyrical
di
creation
lifted
hotstepper
loca
informer
conscience
passenger
knocked
pissing
daydreams
raider
kross
gloom
denial
dance
kris
stabbed
back
nah
body
shout
undo
feel
scrub
genie
spending

2000s:

2010-2014:

As we can see from the word clouds above, some bad words, like fuck and shit, started to show up after 1990s.