Assignment Report - A7

Shabbir, Sharad

October 31, 2017

NOTE: Best viewd in HTML version!

The goal of this assignment is to cluster artists based on their genre.

Subtask 2

Program Design

The program is split into 3 sub-packages:

- `org.neu.pdpmr.tasks.types` : It holds datatypes responsible for reading files and parsing them to provide fields. For extensibility purposes this file will read all fields and store them internally. For use within map reduce it is recommended to extract only portion of fields from it.

- `org.neu.pdpmr.tasks.Main` : Main file that executes whole project.

- `org.neu.pdpmr.tasks.subtask2` : Holds clustering mechanism for the second sub-task.
- `KMode` : It is based on KMode clustering technique.
  1. Initially we assign 1 as weight to all terms associated with any artist.
  2. We then take intersect count of *artist terms* to *centroid terms* as a distance measure.
  3. To calculate new centroid position we take mode (most repeating values) of, ***set of artist terms***, for all artists belonging to that cluster.
  4. Finally, we calculate distance to this new centroid by using same intersection technique.
- `KMeans` : It is an experimental implementation which assumes every `term` as a dimension and does Cartesian distance between them to do clustering.
  1. Initially we assign 1 as weight to all terms associated with any artist.
  2. Then we take a mean across the cluster for each dimension. This could be fractional depending on the actual values. We call this as score of centroid terms.
  3. Finally, we calculate distance to this new centroid by summing scores of matching centroid terms between `artist terms` and the `new centroid` (distance = taking inverse of sum).

Assumptions and Specifications

- **Our centroids are not contrained to be a valid point in the graph.**

It could be a subset of any combination of *terms*. Thus, we cannot use the artist similarity to construct edges between artist and centroids. In our case we assume all artists are connected to each other iff they have at least 1 common term. And distance between them is the inverse of the intersection of matching *terms*.

- We assume artist `A` is similar to artist `B` reverse is **NOT** true. As in case of $A \subset B$.

Results

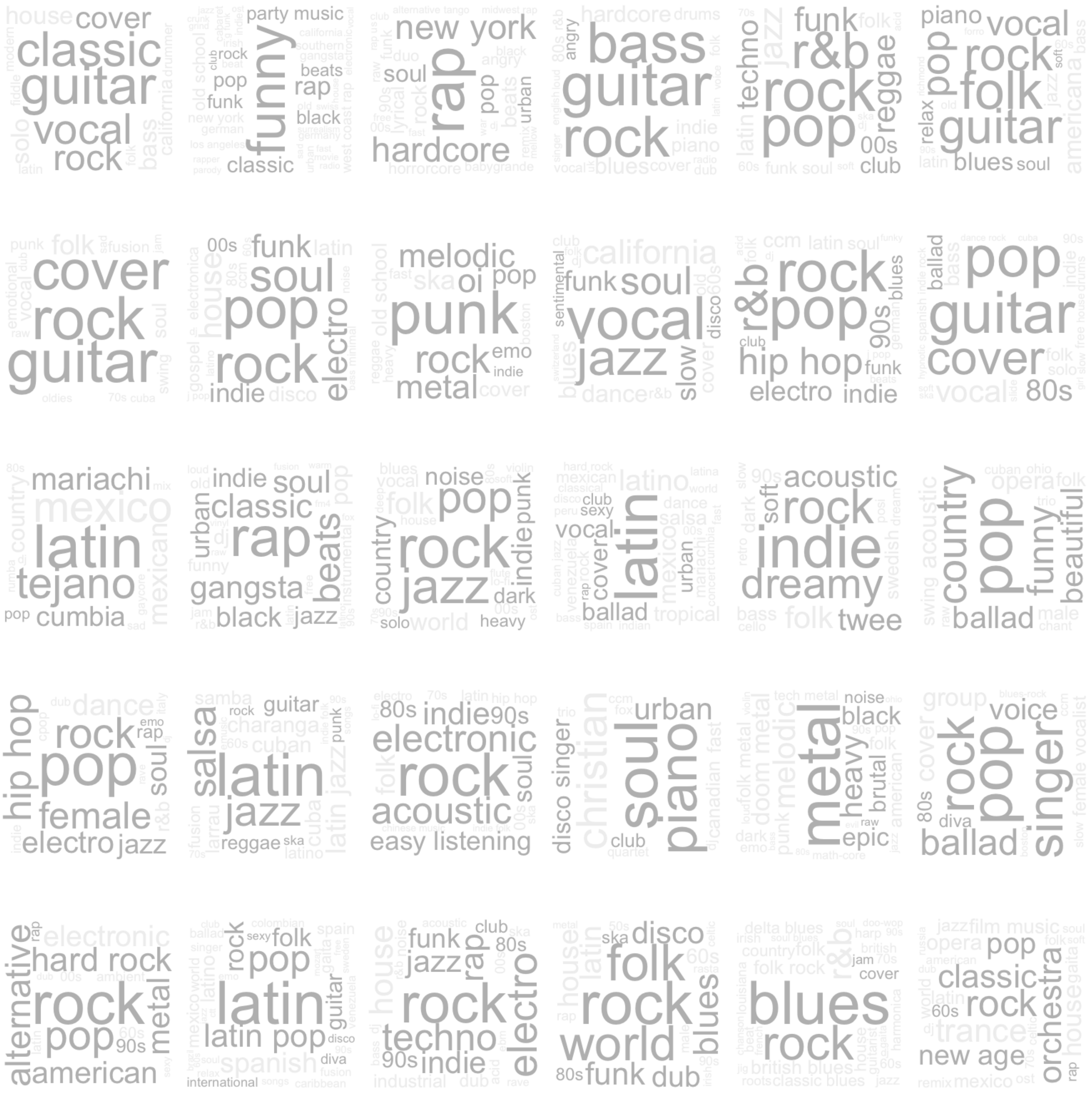
Below tag cloud shows cluster assignment of artists. This report is run on a 10000 songs subset with approx 3800 distinct artists.

Key information to read below figure:

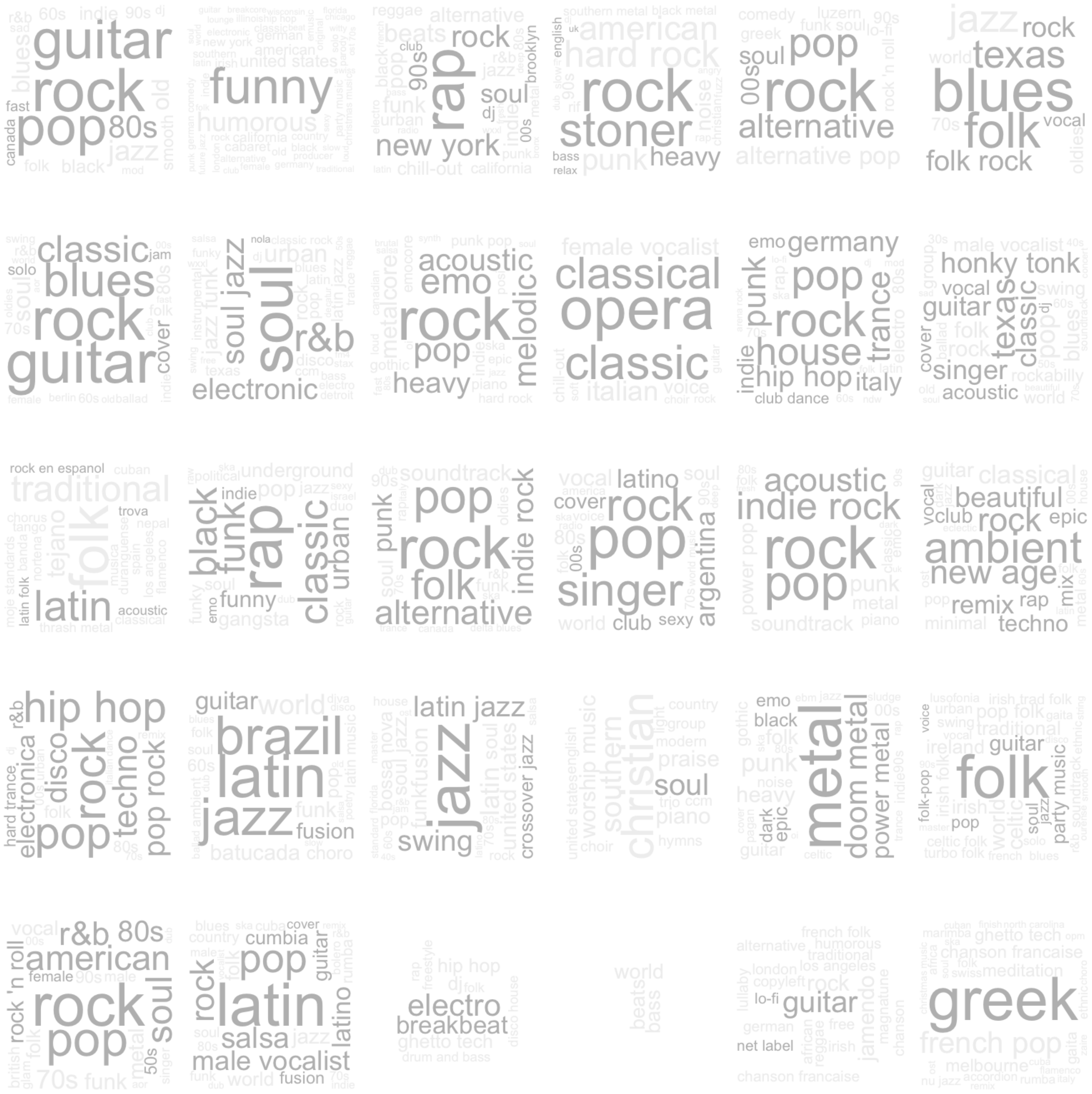
1. Every iteration result is shown as bunch of tag clouds.
2. Every tag cloud is a cluster shown in row major format. And are sorted by `CLUSTER_ID` .
3. For iteration i we highlight *terms* using iteration $i - 1$ centroids. So for de-referencing use previous iteration centroids.
4. Darker color represent centroid terms for that cluster.
5. The size of the word indicates the number of artists assigned to that cluster.
6. Note: Some centroids are not plotted due to shortage of screen space as decided by wordcloud library!

Cluster assignment visualization

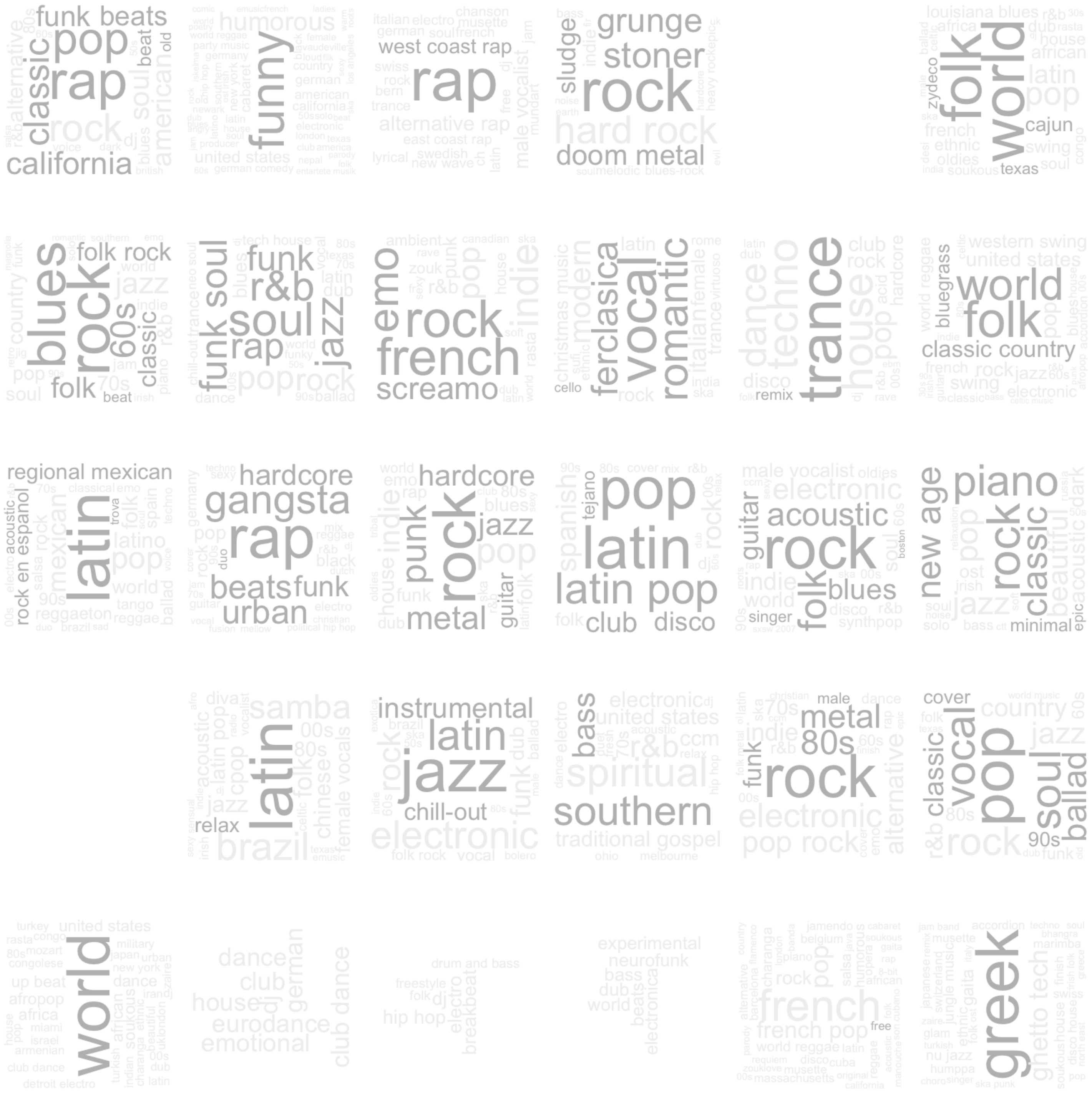
Iter 1


house cover classic guitar vocal rock solo fiddle modern bass california drummer latin folk
party music funny rap beats black pop funk rock classic old school new york german los angeles california southern gangsta west coast rap electronic
new york rap hardcore soul duo angry black pop beats urban lyrical 90s rock horrorcore babygrande alternative tango midwest rap
hardcore drums bass guitar rock angry indie piano blues cover vocal singer radio dub 80s r&b
funk folk r&b rock pop reggae techno jazz latin 00s club 60s funk soul soft 70s acid
piano vocal rock pop folk guitar blues soul relax latin jazz americana bass old
punk folk fusion cover rock guitar emotional vocal raw soul swing oldies 70s cuba jam
00s funk latin soul house pop rock electro indie disco gospel electronica 80s ccm noise
melodic pop oi punk ska fast old school rock emo indie metal cover reggae heavy boston
club california funk soul vocal jazz sentimental blues slow cover disco 60s old dance r&b
ccm latin soul r&b rock pop 90s blues hip hop funk electro indie club german beats acid folk
ballad bass pop guitar cover indie vocal 80s folk solo drums 90s
80s mariachi mexico latin country tejano mexicano cumbia pop rumba mix gaycore
loud old indie soul pop classic rap urban beats gangsta funny black jazz instrumental r&b
blues noise pop vocal folk rock jazz country indie punk dark solo world heavy violin house
hard rock mexican latino latina classical club sexy vocal latin dance salsa mexico cover ballad mariachi tropical venezuela urban cumbia
90s acoustic soft rock indie dark dreamy swedish bass folk twee cello retro
cuban ohio opera folk country pop funny beautiful swing acoustic ballad male chant trio
hip hop dub dance rock emo rap pop soul female electro jazz r&b italy indie
samba rock guitar salsa charanga cuban latin jazz punk reggae ska 60s fusion cuba latino 70s
electro 70s latin hip hop 80s indie 90s electronic rock soul acoustic easy listening folk chinese music ska
trio ccm urban fox christian soul piano disco singer club quartet canadian fast
tech metal noise black doom metal melodic metal heavy brutal folk punk epic raw dark emo evil american math-core
group blues-rock voice rock pop singer cover 80s diva ballad female vocalist slow ccm
electronic rap alternative hard rock rock pop metal 00s ambient 60s 90s american dub latin
club colombian sexy folk spain ballad singer rock pop latin guitar latin pop spanish mexico world diva disco international songs caribbean relax soul
acoustic club funk ska jazz rap 80s house rock electro techno 90s indie industrial dub rave bass dj noise
metal 50s disco ska folk house latin rock 60s world blues 80s funk dub rap celtic rasta male
delta blues soul doo-wop irish harp country folk british folk rock r&b jam 70s cover blues rock louisiana harmonica british blues classic blues roots jazz 60s
jazz film music soul russia opera pop folk american classic rock trance 60s latin world new age orchestra house remix mexico ost rap

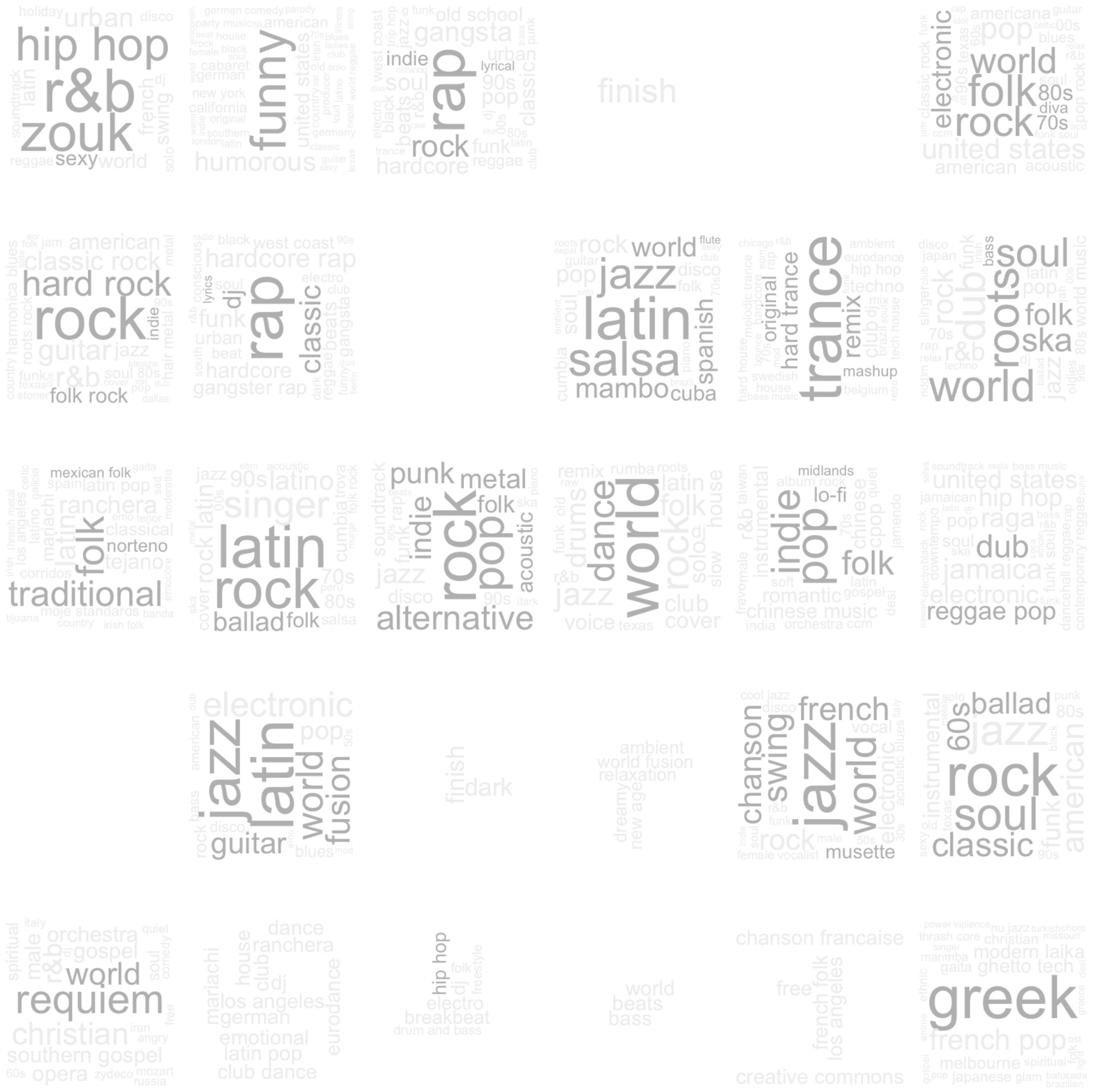
Iter 2


r&b 60s indie 90s dj sad blues guitar rock fast canada pop 80s jazz smooth old folk black mod
guitar breakcore wisconsin florida chicago lounge illinoiship hop electronic classicbeat german new york american southern irish united states funny humorous rock california country cabaret old black producer alternative club female germany traditional
reggae alternative beats rock club 90s rap r&b jazz soul dj brooklyn funk urban radio indie 80s new york chill-out california punk latin
southern metal black metal uk english american hard rock rock noise stoner bass relax punk heavy rap
comedy luzern 90s funk soul lo-fi greek pop soul 00s rock rock 'n roll alternative alternative pop
jazz rock world texas blues 70s folk vocal folk rock oldies
swing r&b classic jam solo blues 80s world soul rock folk 70s cover guitar indie female berlin 60s oldballad
salsa nola classic rock funky urban jazz soul jazz funk blues latin soul r&b instrumental latin jazz disco texas ccm bass electro electronic detroit
brutal synth punk pop soul acoustic emo metalcore rock melodic pop ska canadian gothic heavy epic indie piano hard rock
female vocalist classical opera chill-out classic italian voice choir rock guitar
emo germany punk pop rap trance rock 70s house indie hip hop italy club dance 60s
30s male vocalist 40s honky tonk vocal texas swing guitar classic folk rock pop singer cover blues rockabilly old acoustic world
rock en espanol cuban traditional trova chorus nepal tango tejano folk spain los angeles flamenco latin acoustic classical thrash metal
ska political underground indie pop jazz black funk rap classic urban soul funky funny gangsta emo
dub soundtrack 90s pop punk rock indie rock folk soul alternative trance canada delta blues
vocal latino soul cover rock 80s pop 00s singer argentina world club sexy
80s acoustic indie rock rock power pop pop punk metal soundtrack piano
guitar classical vocal beautiful club rock epic ambient new age pop remix rap mix minimal techno metal
r&b hip hop hard trance electronica disco rock techno pop rock pop 80s 70s
guitar world diva disco blues folk soul 60s brazil latin jazz funk fusion batucada choro
house latin jazz salsa bossa nova soul jazz funkfusion jazz latin soul united states crossover jazz pop swing rock
country group modern praise worship music southern christian soul trio ccm united states english piano choir hymns
emo ebm jazz sludge 00s black gothic folk punk noise heavy metal doom metal power metal dark epic guitar celtic
lusofonia irish trad folk urban pop folk voice traditional ireland guitar folk irish folk-pop pop world celtic soul jazz party music celtic folk turbo folk french blues
vocal r&b 80s 00s american female 90s male rock 'n roll rock pop soul british folk metal 50s 70s funk
blues ska cuba cover country cumbia male rock pop guitar latin latino salsa jazz male vocalist soul 80s funk world fusion 70s indie
hip hop freestyle dj folk electro breakbeat ghetto tech drum and bass disco house
world beats bass
french folk alternative humorous traditional london los angeles copyleft rock lo-fi guitar german free net label irish jamendo magnatune chanson chanson francaise
cuban finish north carolina marimba ghetto tech chanson francaise meditation greek french pop melbourne accordion nu jazz rumba italy remix

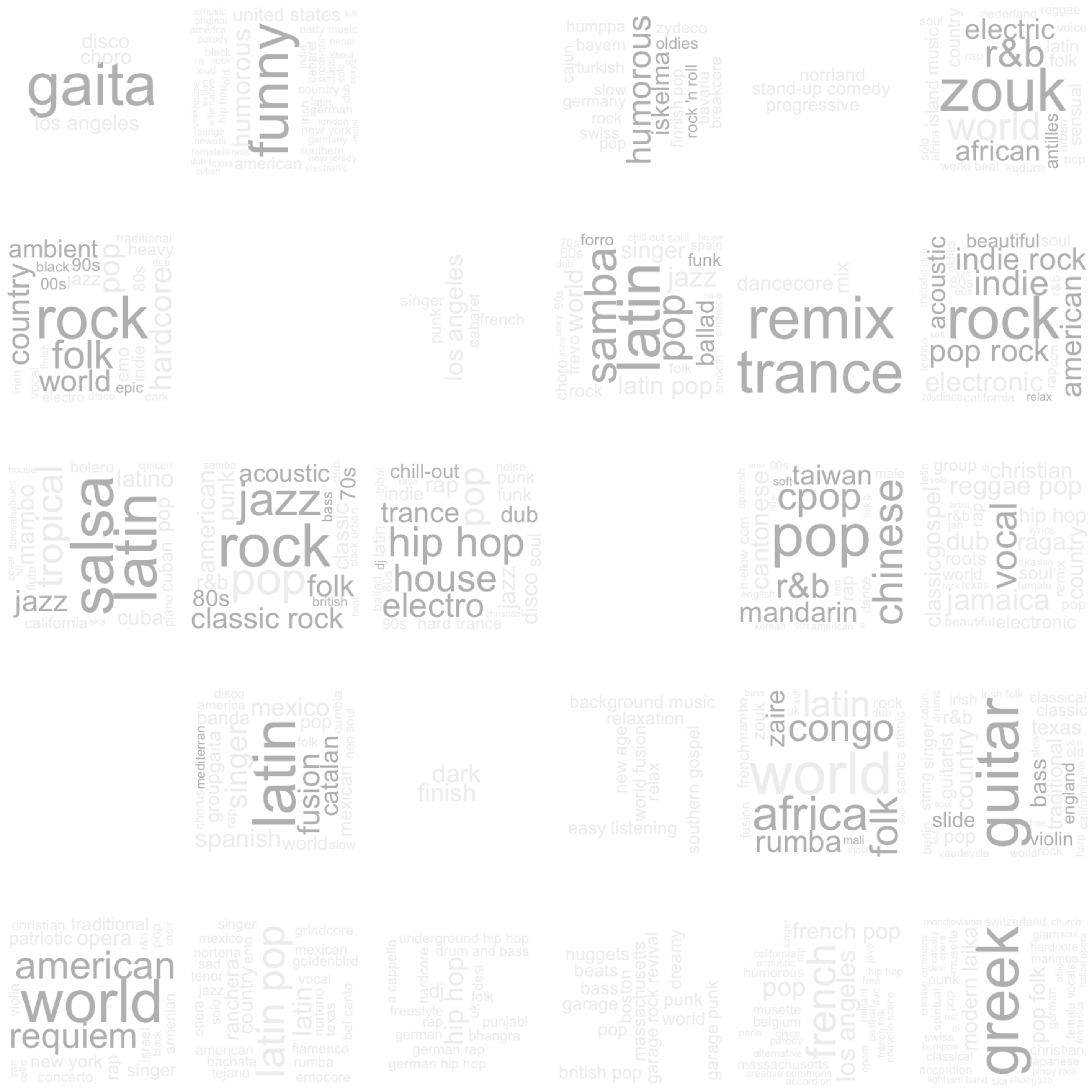
Iter 3


funk beats
pop
rap
classic
rock
california
humorous
funny
west coast rap
rap
alternative rap
grunge
stoner
rock
hard rock
doom metal
sludge
louisiana blues
folk
world
cajun
blues
rock
folk rock
60s
folk
funk
r&b
soul
rap
jazz
funk soul
emo
rock
french
screamo
indie
ferclasica
vocal
romantic
trance
techno
dance
house
remix
world
folk
bluegrass
classic country
regional mexican
latin
rock en espanol
hardcore
gangsta
rap
beats
funk
urban
hardcore
punk
rock
jazz
metal
pop
latin
latin pop
club
disco
acoustic
rock
guitar
folk
blues
singer
piano
new age
rock
classic
minimal
samba
latin
brazil
relax
instrumental
latin
jazz
chill-out
electronic
bass
spiritual
southern
traditional gospel
metal
80s
rock
funk
electronic
pop rock
cover
pop
vocal
soul
ballad
90s
world
united states
dance
club
house
eurodance
emotional
drum and bass
hip hop
experimental
neurofunk
french
french pop
greek

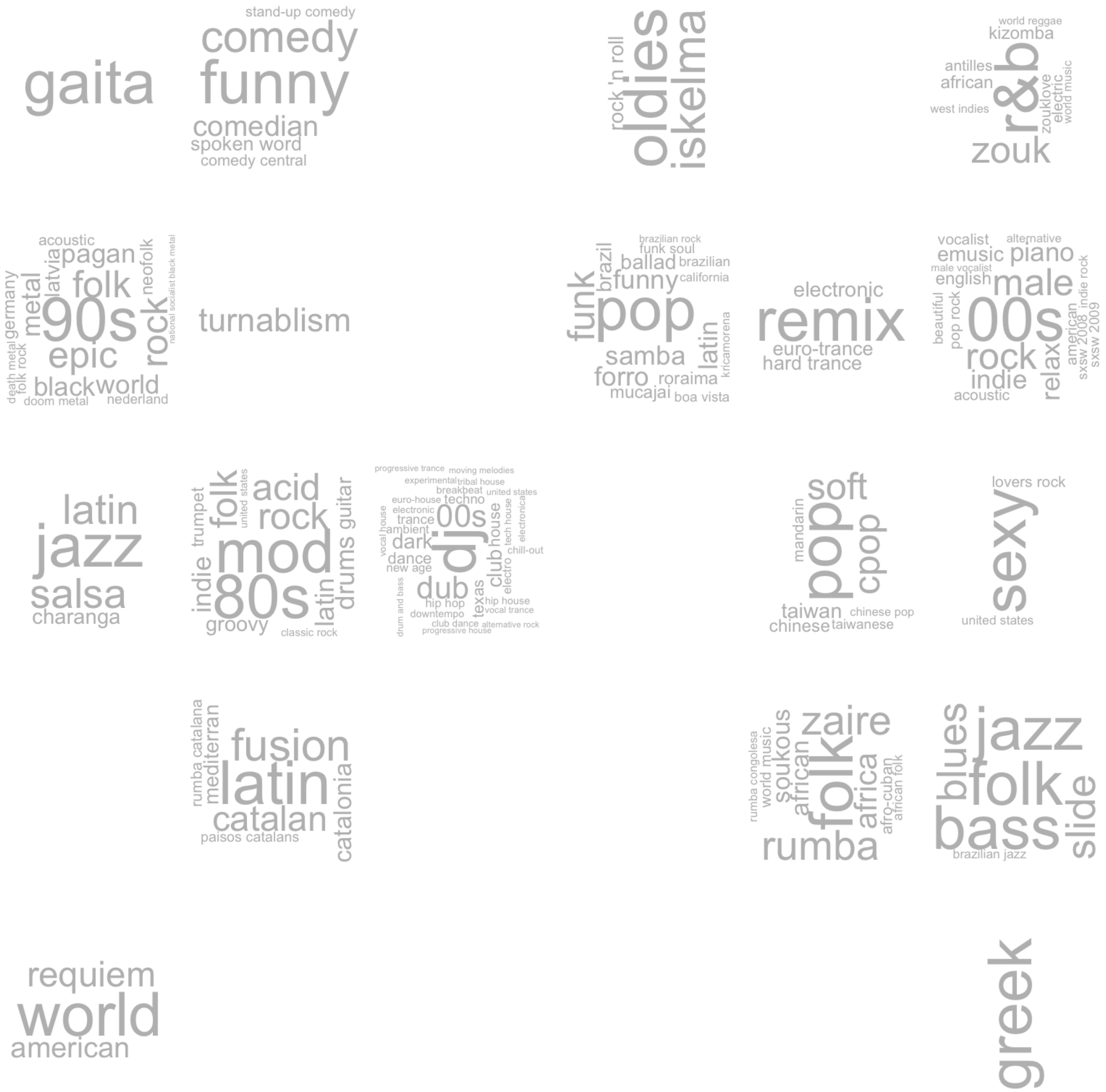
Iter 6


hip hop r&b zouk urban sexy holiday disco reggae world french swing soundtrack latin dj solo
funny humorous american german comedy parody cabaret german new york california united states house southern london latin germany
rap gangsta indie lyrical old school soul 90s pop rock hardcore funk urban classic beats black west coast reggae trance
finish
folk rock world pop electronic 80s 70s diva americana united states american acoustic blues
rock hard rock classic rock american guitar r&b folk rock indie jazz soul 80s country harmonica blues hair metal
rap dj classic hardcore rap hardcore gangster rap funk urban beat lyrics soul black west coast electro beats gangsta
latin jazz salsa mambo cuba world spanish rock pop disco folk soul cumbia flute
trance hard trance original remix techno hip hop ambient eurodance mashup belgium swedish house
world soul roots dub ska folk r&b jazz 70s pop latin funk bass
folk traditional mexican folk ranchera latin pop norteno tejano classical mariachi corridos
latin rock singer latino 90s ballad folk 70s 80s salsa cumbia trova
rock pop punk metal indie folk acoustic alternative jazz disco soundtrack
world dance drums jazz rock folk latin remix club cover solo slow house voice
pop indie folk lo-fi midlands album rock instrumental romantic chinese music chinese cpop soft
dub hip hop raga jamaica united states electronic reggae pop pop soul downtempo
electronic jazz latin world fusion pop guitar disco blues american
finish dark
ambient world fusion relaxation dreamy new age
jazz chanson french swing world electronic musette rock vocal
rock soul classic jazz ballad 60s american instrumental funk
requiem world gospel orchestra christian southern gospel opera male r&b
dance ranchera house club dj los angeles german emotional latin pop club dance mariachi eurodance
hip hop electro breakbeat drum and bass dj folk freestyle
world beats bass
chanson francaise free french folk los angeles creative commons
greek french pop modern laika ghetto tech christian thrash core melbourne japanese

Iter 10



Final centroids after all iterations


gaita
stand-up comedy
comedy
funny
comedian
spoken word
comedy central
rock 'n roll
oldies
iskelma
world reggae
kizomba
antilles
african
west indies
r&b
zouklove
electric
world music
zouk
acoustic
pagan
latvia
folk
neofolk
national socialist black metal
germany
metal
90s
rock
epic
death metal
folk rock
black
world
doom metal
nederland
turnablism
brazilian rock
funk soul
ballad
brazilian
brazil
funny
california
funk
pop
latin
kricamorena
samba
forro
roraima
mucajai
boa vista
electronic
remix
euro-trance
hard trance
vocalist
alternative
emusic
piano
male vocalist
english
male
beautiful
pop rock
00s
indie rock
american
sxsw 2008
sxsw 2009
rock
indie
relax
acoustic
latin
jazz
salsa
charanga
trumpet
folk
united states
acid
rock
drums
guitar
indie
mod
80s
latin
groovy
classic rock
progressive trance
moving melodies
experimental
tribal house
breakbeat
united states
euro-house
techno
electronic
trance
00s
tech house
electronica
ambient
vocal house
dark
dj
house
club
chill-out
dance
new age
electro
dub
drum and bass
hip hop
texas
hip house
downtempo
vocal trance
club dance
alternative rock
progressive house
soft
mandarin
pop
cpop
taiwan
chinese pop
chinese
taiwanese
lovers rock
sexy
united states
rumba catalana
mediterran
fusion
latin
catalan
paisos catalans
catalonia
rumba congolesa
world music
soukous
african
zaire
folk
africa
afro-cuban
african folk
rumba
blues
jazz
folk
bass
slide
brazilian jazz
requiem
world
american
greek

Conclusion

KModes is an interesting clustering results. We observed that centroids are mostly stabilized by 6th iteration on small corpus (with some exceptions). As algorithm progresses the clusters seems to merge together to form a bigger distinction between *terms* as seen in the case of `POP` which was prominent in every cluster earlier in iteration 1. It is later aggregated to its own *term* towards the end of the algorithm.

Task Assignment

Shabbir was responsible for sub-problem 2 including sub-problem 2 reporting and respective Makefile.

Sharad was responsible for sub-problem 1 including sub-problem 1 reporting and respective Makefile.