Words for Music

Yixin Sun

Clean Dataset

- H5 File

Feature From Analysis

- 13 features at first

```
 [1] "bars_confidence"            "bars_start"                 "beats_confidence"
 [4] "beats_start"                "segments_confidence"        "segments_loudness_max"
 [7] "segments_loudness_max_time" "segments_loudness_start"    "segments_pitches"
[10] "segments_start"             "segments_timbre"            "tatums_confidence"
[13] "tatums_start"
```

- Final feature I choose

```
> names(features_df)
 [1] "duration"                "end_of_fade_in"              "key"
 [4] "key_confidence"          "loudness"                    "mode"
 [7] "mode_confidence"         "start_of_fade_out"           "tempo"
[10] "time_signature"          "time_signature_confidence"   "cluster"
```

Bag of Words

- Some words not suitable

```
 [1] "\u0096" "–"      "&"      "000"    "1"      "10"     "100"    "12"     "13"     "15"
[11] "16"     "2"       "20"     "24"     "2x"     "3"      "30"     "3x"     "4"      "40"
[21] "4x"     "5"       "50"     "6"      "7"      "8"      "9"
```

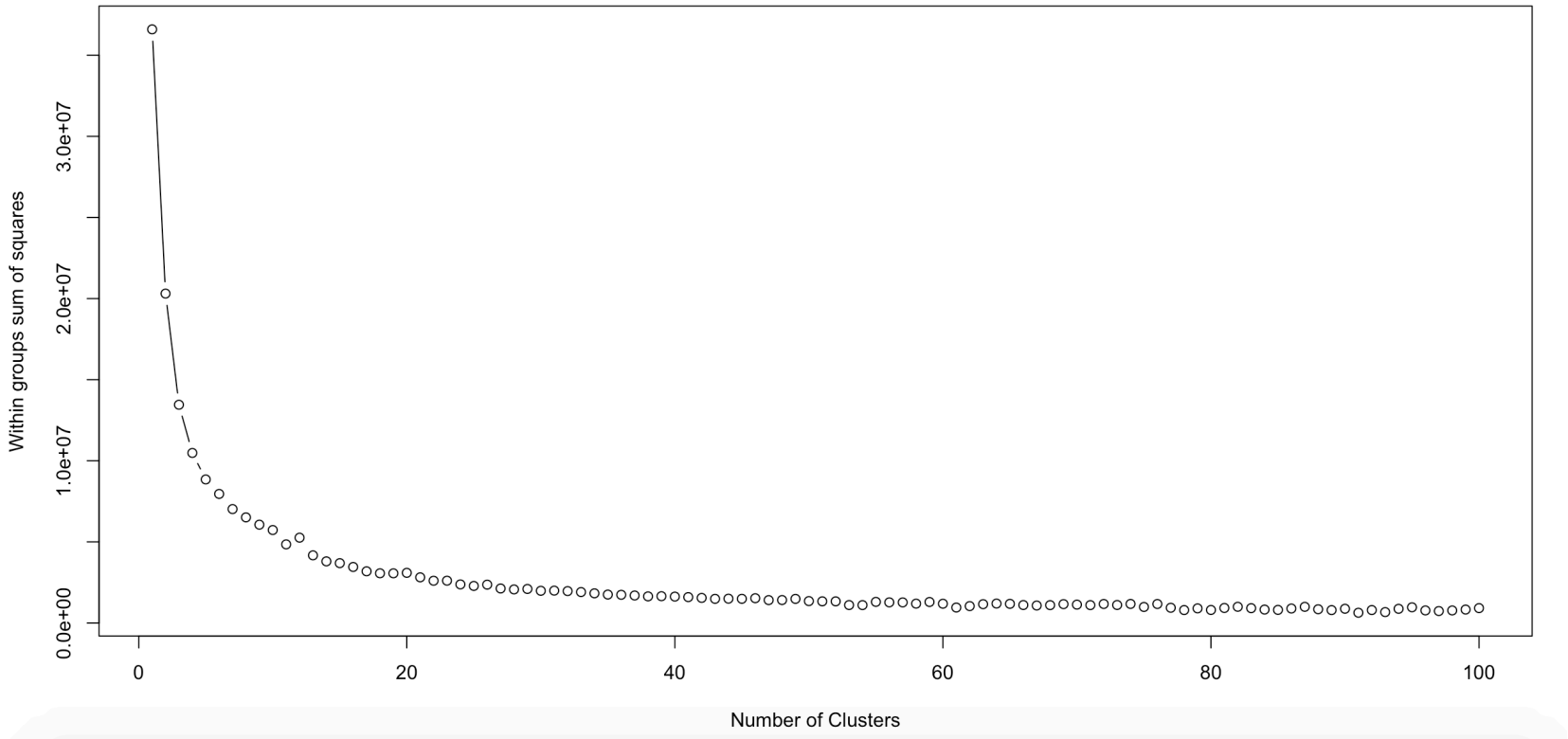
- Eliminate the effects from those words.

Algorithm

- Cluster using the features of songs

Number of Clusters

- Non-significant change when number of cluster larger than 15



K-Means Cluster Analysis

- Using 15 clusters
- Get cluster means
- Append cluster assignment

Loading Bag of Words

- Count word frequencies songs inside each cluster
- Sort words according to probability
- Rank each word according to the sort result

Probability of Words

- Common Words
  - for example: “I”, “the”, “ you”, “to”, “and”, “me”, “is” ect.
- All cluster have similar words with high probability
- Clearly different ranking between each cluster with low probability of word

Association rules

- Within Lyric data

Clean dataset

- Transform the values in lyr data set which are not 0s into 1s, then our data set is a binary data frame

```
> head(lyr10[560:570,450:453])
      TRAGGMY12903CD9533 TRAGGVS128F4250CE8 TRAGHBP128E0793AF7 TRAGHJX128F426F76F
bruis                  0                  0                  0                  0
brush                  0                  0                  0                  0
brutal                 0                  0                  0                  0
bu                     0                  0                  0                  0
bubbl                  0                  0                  0                  0
buck                   0                  0                  0                  0
```

```
> rules = apriori(t(lyr10),
+                 parameter = list(support = 0.2, confidence = 0.4))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport support
        0.4    0.1    1 none FALSE            TRUE     0.2
 minlen maxlen target   ext
      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 470

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[4851 item(s), 2350 transaction(s)] done [0.07s].
sorting and recoding items ... [73 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [3.68s].
writing ... [1070351 rule(s)] done [0.51s].
creating S4 object  ... done [4.06s].

> summary(rules)
set of 1070351 rules

rule length distribution (lhs + rhs):sizes
     1      2      3      4      5      6      7      8      9     10
    30   1789  18289  80997 197519 289516 266623 153896  52272   9420

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   5.000   6.000   6.324   7.000  10.000

summary of quality measures:
    support         confidence          lift
 Min.   :0.2000   Min.   :0.4000   Min.   :1.000
 1st Qu.:0.2081   1st Qu.:0.7986   1st Qu.:1.245
 Median :0.2200   Median :0.8955   Median :1.306
 Mean   :0.2315   Mean   :0.8590   Mean   :1.312
 3rd Qu.:0.2421   3rd Qu.:0.9545   3rd Qu.:1.373
 Max.   :0.8081   Max.   :1.0000   Max.   :1.738

mining info:
      data ntransactions support confidence
 t(lyr10)          2350     0.2        0.4
```

- Set support to 0.2, confidence to 0.6, we have 1070351 rules, so there are too many rules, we need to set higher values, at last, we use support 0.6, confidence to 0.6, this gives only 86 rules now

```
> rules = apriori(t(lyr10),
+                 parameter = list(support = 0.6, confidence = 0.6))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport support
        0.6    0.1    1 none FALSE            TRUE     0.6
 minlen maxlen target   ext
      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 1410

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[4851 item(s), 2350 transaction(s)] done [0.06s].
sorting and recoding items ... [12 item(s)] done [0.00s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [86 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].

> summary(rules)
set of 86 rules

rule length distribution (lhs + rhs):sizes
 1  2  3
12 44 30

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   2.000   2.209   3.000   3.000

summary of quality measures:
    support         confidence          lift
 Min.   :0.6047   Min.   :0.6089   Min.   :1.000
 1st Qu.:0.6188   1st Qu.:0.8155   1st Qu.:1.066
 Median :0.6413   Median :0.8770   Median :1.176
 Mean   :0.6535   Mean   :0.8614   Mean   :1.140
 3rd Qu.:0.6749   3rd Qu.:0.9195   3rd Qu.:1.187
 Max.   :0.8081   Max.   :0.9649   Max.   :1.201

mining info:
      data ntransactions support confidence
 t(lyr10)          2350     0.6        0.6
```

Inspect the top 30 rules

- inspect for the top 30 rules with the highest confidence
- rules with high lift have typically a relatively low support

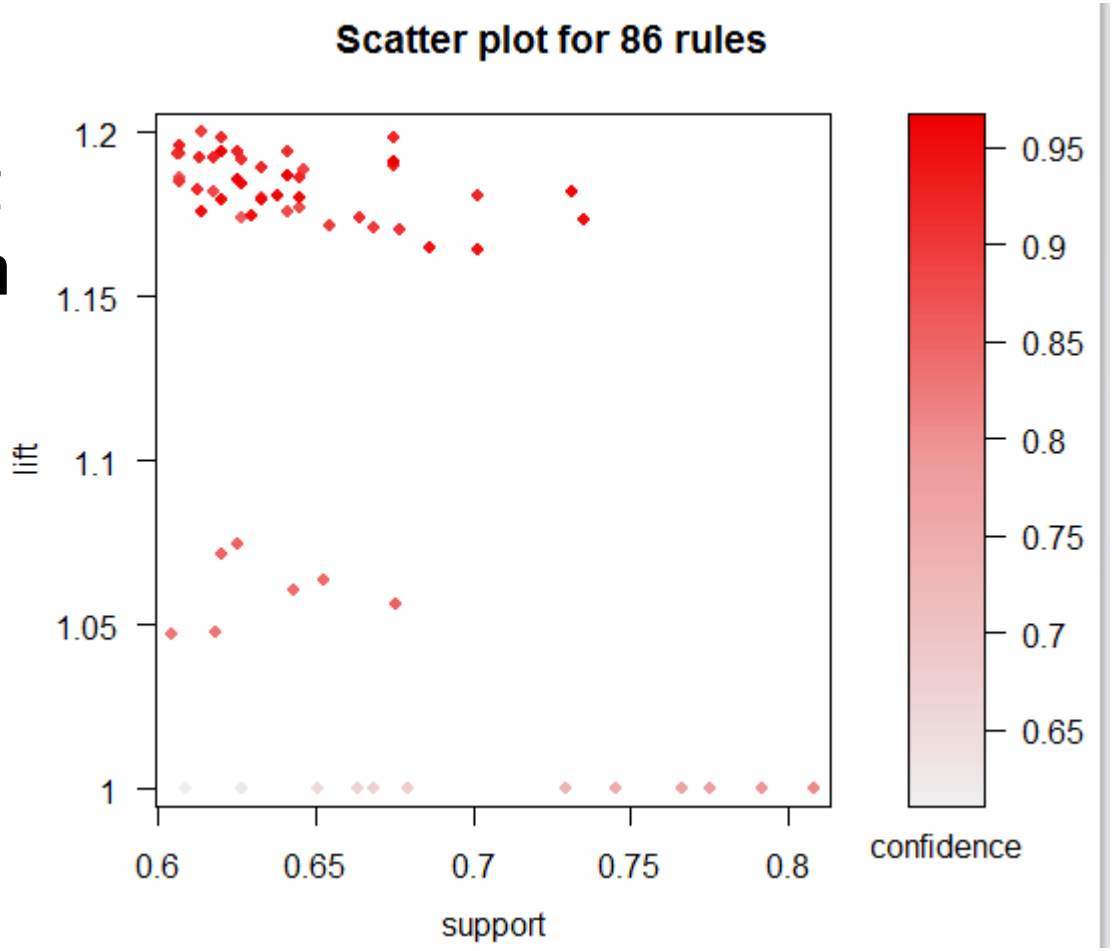
```
> inspect(head(rules, n = 30, by = "confidence"))
     lhs           rhs   support   confidence lift
[1]  {a,and}    => {the} 0.6204255 0.9649239  1.194087
[2]  {and,to}   => {the} 0.6748936 0.9623786  1.190937
[3]  {and,i}    => {the} 0.6412766 0.9592616  1.187080
[4]  {a,to}     => {the} 0.6255319 0.9582790  1.185864
[5]  {and,you}  => {the} 0.6263830 0.9570871  1.184389
[6]  {and}      => {the} 0.7314894 0.9550000  1.181806
[7]  {is}       => {the} 0.6378723 0.9541693  1.180778
[8]  {i,to}     => {the} 0.6451064 0.9534591  1.179899
[9]  {to,you}   => {the} 0.6327660 0.9532051  1.179585
[10] {in}       => {the} 0.6204255 0.9529412  1.179258
[11] {i,you}    => {the} 0.6140426 0.9499671  1.175578
[12] {it}       => {the} 0.6297872 0.9493265  1.174785
[13] {to}       => {the} 0.7353191 0.9484083  1.173649
[14] {you}      => {the} 0.6863830 0.9410735  1.164572
[15] {i}        => {the} 0.7012766 0.9406393  1.164035
[16] {and,you}  => {to}  0.6068085 0.9271782  1.195866
[17] {a,the}    => {to}  0.6255319 0.9256927  1.193951
[18] {it}       => {to}  0.6131915 0.9243105  1.192168
[19] {and,i}    => {to}  0.6178723 0.9242521  1.192092
[20] {and,the}  => {to}  0.6748936 0.9226294  1.190000
[21] {the,you}  => {to}  0.6327660 0.9218847  1.189039
[22] {i,the}    => {to}  0.6451064 0.9199029  1.186483
[23] {a,the}    => {and} 0.6204255 0.9181360  1.198678
[24] {the,to}   => {and} 0.6748936 0.9178241  1.198270
[25] {is}       => {to}  0.6127660 0.9166136  1.182240
[26] {and}      => {to}  0.7012766 0.9155556  1.180876
[27] {i,the}    => {and} 0.6412766 0.9144417  1.193855
[28] {to,you}   => {and} 0.6068085 0.9141026  1.193412
[29] {it}       => {and} 0.6063830 0.9140475  1.193340
[30] {i,to}     => {and} 0.6178723 0.9132075  1.192243
```

Interpretation

- So we find that lots of rules are in the following format:
- E.g. if we both have ‘a’ and ‘and’ then we probability also have ‘the’ with support 0.6204255, confidence 0.9649239 and lift 1.194087.
- There are lots of words like ‘I’, ’a’, ’and’, ’the’,
- ‘to’, ‘is’, ‘it’, it seems we need to remove to find a better pattern.

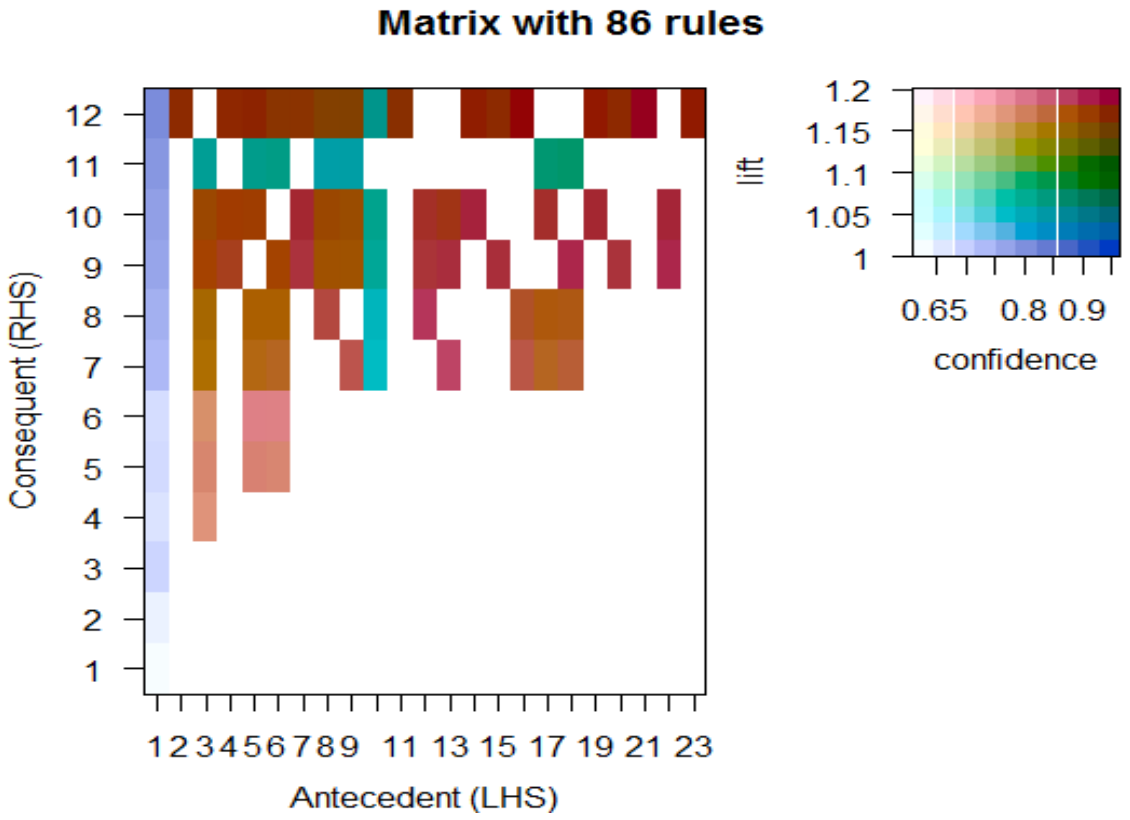
Rules scatter plot

- We can see that rules with high lift have support from
- 0.6 to 0.7 which is not a wide range

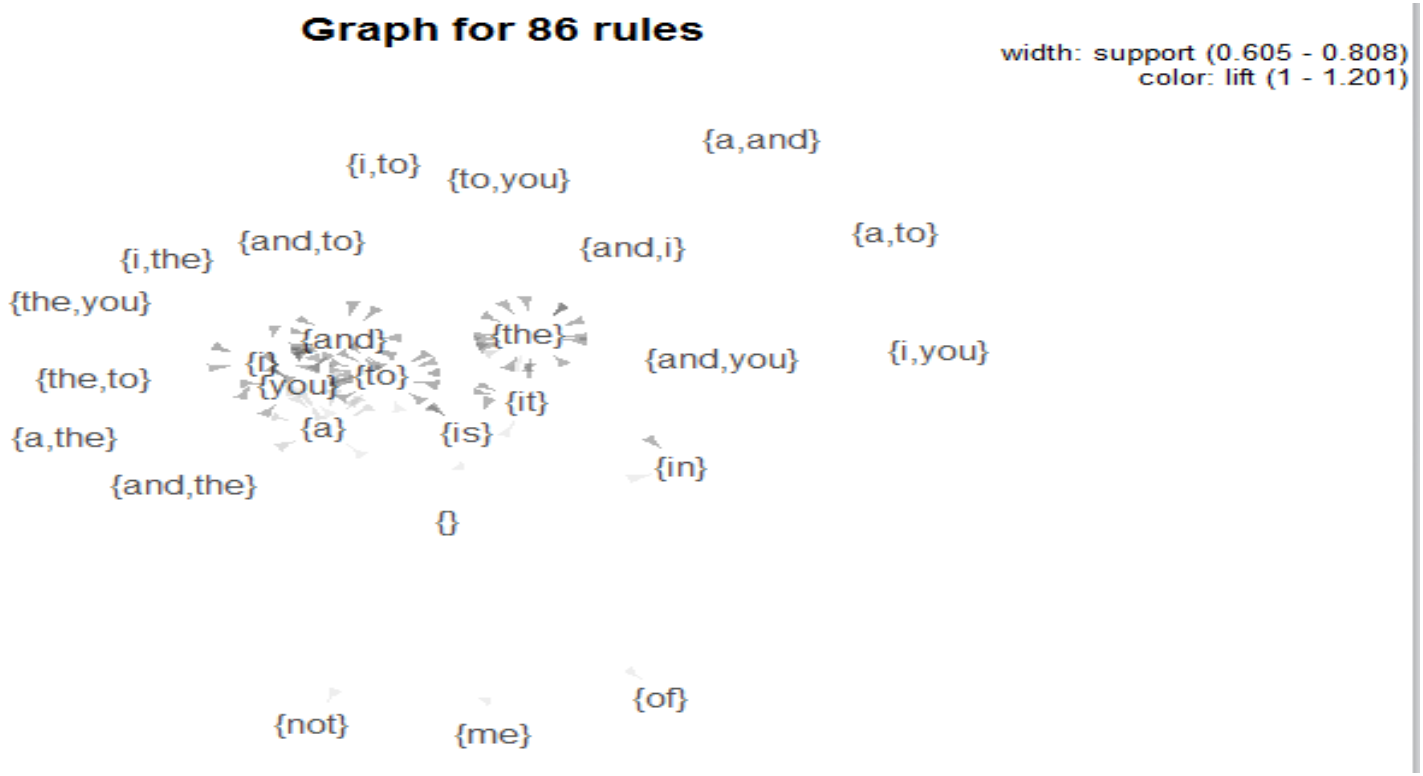


Rules plot

- High confidence/high support rules can be identified in the plot as hot/red (high confidence) and dark/intense (high support). on a black and white printout, the different colors are not distinguishable



- This plot representation focuses on how the rules are composed of individual words and shows which rules share words, and we can find the words, ‘I’, ’and’, ‘you’, ‘to’, ‘the’ are on the center of the rules.



Thank You!