Lab 5: Topic Modeling

Srishti Saha (ss1078)

06 March, 2020

Exercise 1: Using the link above and the downloaded file, load the lyrics dataset into your workspace.

Table 1: Top 5 rows from billboard data

Rank	Song	Artist	Year	Lyrics	Source
1	wooly bully	sam the sham and the pharaohs	1965	sam the sham miscellaneous wooly bully wooly bully sam the sham the pharaohs domingo samudio uno dos one two tres quatro matty told hatty about a thing she saw had two big horns and a wooly jaw wooly bully wooly bully wooly bully wooly bully wooly bully hatty told matty lets dont take no chance lets not belseven come and learn to dance wooly bully wooly bully wooly bully wooly bully wooly bully matty told hatty thats the thing to do get you someone really to pull the wool with you wooly bully wooly bully wooly bully wooly bully wooly bully lseven the letter l and the number 7 when typed they form a rough square l7 so the lyrics mean lets not be square	3
2	i cant help myself sugar pie honey bunch	four tops	1965	sugar pie honey bunch you know that i love you i cant help myself i love you and nobody elsein and out my life you come and you go leaving just your picture behind and i kissed it a thousand timeswhen you snap your finger or wink your eye i come arunning to you im tied to your apron strings and theres nothing that i can docant help myself no i cant help myselfsugar pie honey bunch im weaker than a man should be i cant help myself im a fool in love you seewanna tell you i dont love you tell you that were through and ive tried but every time i see your face i get all choked up insidewhen i call your name girl it starts the flame burning in my heart tearing it all apart no matter how i try my love i cannot hidecause sugar pie honey bunch you know that im weak for you cant help myself i love you and nobody elsesugar pie honey bunch do anything you ask me to cant help myself i want you and nobody elsesugar pie honey bunch you know that i love you i cant help myself i cant help myself	1
3	i cant get no satisfaction	the rolling stones	1965		1
4	you were on my mind	we five	1965	when i woke up this morning you were on my mind and you were on my mind i got troubles whoaoh i got worries whoaoh i got wounds to bind so i went to the corner just to ease my pains yeah just to ease my pains i got troubles whoaoh i got worries whoaoh i came home again when i woke up this morning you were on my miiiind and you were on my mind i got troubles whoaoh i got worries whoaoh i got wounds to bind and i got a feelin down in my shooooooes said way down in my shooooes yeah i got to ramble whoaoh i got to move on whoaoh i got to walk away my blues when i woke up this morning you were on my mind you were on my mind i got troubles whoaoh i got worries whoaoh i got wounds to bind	1
5	youve lost that lovin feelin	the righteous brothers	1965	you never close your eyes anymore when i kiss your lips and theres no tenderness like before in your fingertips youre trying hard not to show it but baby baby i know ityou lost that lovin feelin whoa that lovin feelin you lost that lovin feelin now its gone gone gone wohnow theres no welcome look in your eyes when i reach for you and now youre starting to criticize little things i do it makes me just feel like crying cause baby something beautifuls dyinyou lost that lovin feelin whoa that lovin feelin you lost that lovin feelin now its gone gone gone wohbaby baby id get down on my knees for you if you would only love me like you used to do yeah we had a love a love a love you dont find everyday so dont dont dont dont let it slip awaybaby baby baby baby i beg you please please please please i need your love need your love i need your love i need your love so bring it on back so bring it on back bring it on back bring it on backbring back that lovin feelin whoa that lovin feelin bring back that lovin feelin cause its gone gone gone and i cant go on wohbring back that lovin feelin whoa that lovin feelin bring back that lovin feelin cause its gone gone gone	1

Exercise 2: Subset the data into "decades of lyrics" so that each new dataframe contains the lyrics and other columns from a particular decade of music. Use the following decades so that each has a dataset of song lyrics: 1965-1974, 1975-1984, 1985-1994, 1995-2004, 2005-2014.

```
data_1965_1974 <- subset(billboard_data, Year<=1974 & Year>=1965)
data_1975_1984 <- subset(billboard_data, Year<=1984 & Year>=1975)
data_1985_1994 <- subset(billboard_data, Year<=1994 & Year>=1985)
```

```r
data_1995_2004 <- subset(billboard_data, Year<=2004 & Year>=1995)
data_2005_2014 <- subset(billboard_data, Year<=2014 & Year>=2005)
```

Exercise 3: Prepare each of the datasets so that it can be analyzed using the topicmodels package.

The first step was to select the relevant columns. We will now create a tidy text.

We will conduct the following steps to preprocess this text data:

- Removal of white spaces
- Removing numbers
- Removing intra-word punctiation
- Lemmatization
- Removing words of length less than 3 (short words)
- Removal of Stop words
- Converting to lower case words (done automatically for tidytext objects)

White Spaces

```r
# removing white spaces
tidy_reviews1$word <- gsub("\\s+","",tidy_reviews1$word)
tidy_reviews2$word <- gsub("\\s+","",tidy_reviews2$word)
tidy_reviews3$word <- gsub("\\s+","",tidy_reviews3$word)
tidy_reviews4$word <- gsub("\\s+","",tidy_reviews4$word)
tidy_reviews5$word <- gsub("\\s+","",tidy_reviews5$word)
```

Removing Numbers

```r
# removing numbers from the words list
tidy_reviews1<-tidy_reviews1[-grep("\\b\\d+\\b", tidy_reviews1$word),]
tidy_reviews2<-tidy_reviews2[-grep("\\b\\d+\\b", tidy_reviews2$word),]
tidy_reviews3<-tidy_reviews3[-grep("\\b\\d+\\b", tidy_reviews3$word),]
tidy_reviews4<-tidy_reviews4[-grep("\\b\\d+\\b", tidy_reviews4$word),]
tidy_reviews5<-tidy_reviews5[-grep("\\b\\d+\\b", tidy_reviews5$word),]
```

Punctuation

Although interword punctuation is removed in tidytext automatically, we will remove intraword punctuations separately.

```r
# removing punctuation (intraword)
tidy_reviews1$word<-removePunctuation(tidy_reviews1$word,preserve_intra_word_contractions = FALSE,
                  preserve_intra_word_dashes = FALSE)

tidy_reviews2$word<-removePunctuation(tidy_reviews2$word,preserve_intra_word_contractions = FALSE,
                  preserve_intra_word_dashes = FALSE)

tidy_reviews3$word<-removePunctuation(tidy_reviews3$word,preserve_intra_word_contractions = FALSE,
                  preserve_intra_word_dashes = FALSE)

tidy_reviews4$word<-removePunctuation(tidy_reviews4$word,preserve_intra_word_contractions = FALSE,
                  preserve_intra_word_dashes = FALSE)

tidy_reviews5$word<-removePunctuation(tidy_reviews5$word,preserve_intra_word_contractions = FALSE,
                  preserve_intra_word_dashes = FALSE)
```

Lemmatization

```
#Lemmatization
tidy_reviews1<-tidy_reviews1  %>%
  mutate(word = textstem::lemmatize_words(word))

tidy_reviews2<-tidy_reviews2  %>%
  mutate(word = textstem::lemmatize_words(word))

tidy_reviews3<-tidy_reviews3  %>%
  mutate(word = textstem::lemmatize_words(word))

tidy_reviews4<-tidy_reviews4  %>%
  mutate(word = textstem::lemmatize_words(word))

tidy_reviews5<-tidy_reviews5  %>%
  mutate(word = textstem::lemmatize_words(word))
```

Removing short words

```
## # A tibble: 9,056 x 2
##    word      n
##    <chr> <int>
##  1 ""    60143
##  2 you    9333
##  3 the    8576
##  4 and    4770
##  5 love   3374
##  6 get    2583
##  7 your   2160
##  8 that   2119
##  9 know   1768
## 10 all    1760
## # ... with 9,046 more rows
```

Stop words

```
data("stop_words")

stop_words2= as.data.frame(c("im","yes","you","gonna","", "gotta","wanna","la","thoia","dem","dat","aint","shes",
names(stop_words2)[1]<- "word"

##### for first decade
# remove the stop words in the list above
tidy_reviews1<-tidy_reviews1 %>%
      anti_join(stop_words)

# from custom list
tidy_reviews1<-tidy_reviews1 %>%
      anti_join(stop_words2)


##### for second decade
tidy_reviews2<-tidy_reviews2 %>%
      anti_join(stop_words)

tidy_reviews2<-tidy_reviews2 %>%
      anti_join(stop_words2)
```

```
##### for third decade
tidy_reviews3<-tidy_reviews3 %>%
      anti_join(stop_words)

tidy_reviews3<-tidy_reviews3 %>%
      anti_join(stop_words2)

##### for fourth decade
tidy_reviews4<-tidy_reviews4 %>%
      anti_join(stop_words)

tidy_reviews4<-tidy_reviews4 %>%
      anti_join(stop_words2)

##### for fifth decade
tidy_reviews5<-tidy_reviews5 %>%
      anti_join(stop_words)

tidy_reviews5<-tidy_reviews5 %>%
      anti_join(stop_words2)
```

Converting to lower case

This step happens automatically in the tidytext format

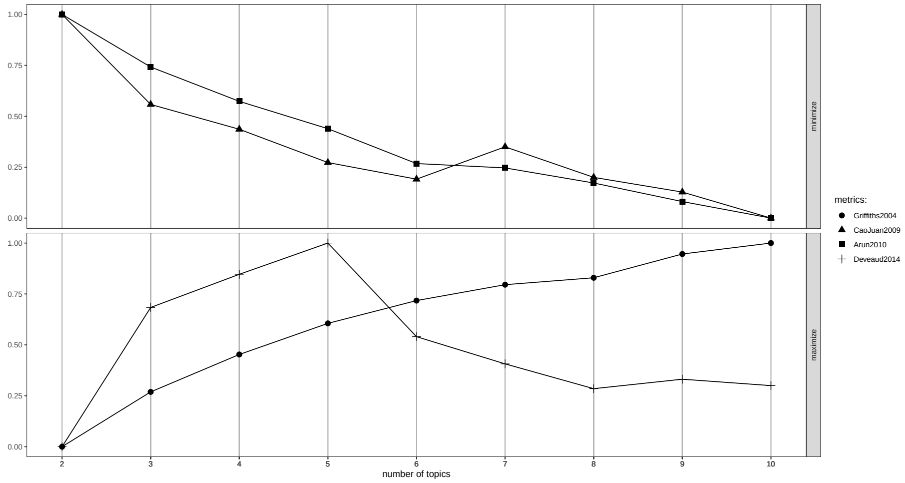
```
## # A tibble: 8,739 x 2
##    word      n
##    <chr> <int>
##  1 love   3374
##  2 baby   1196
##  3 time   1157
##  4 night  1058
##  5 feel    844
##  6 dance   653
##  7 girl    610
##  8 heart   567
##  9 ive     556
## 10 ill     529
## # ... with 8,729 more rows
```

Document Term matrix

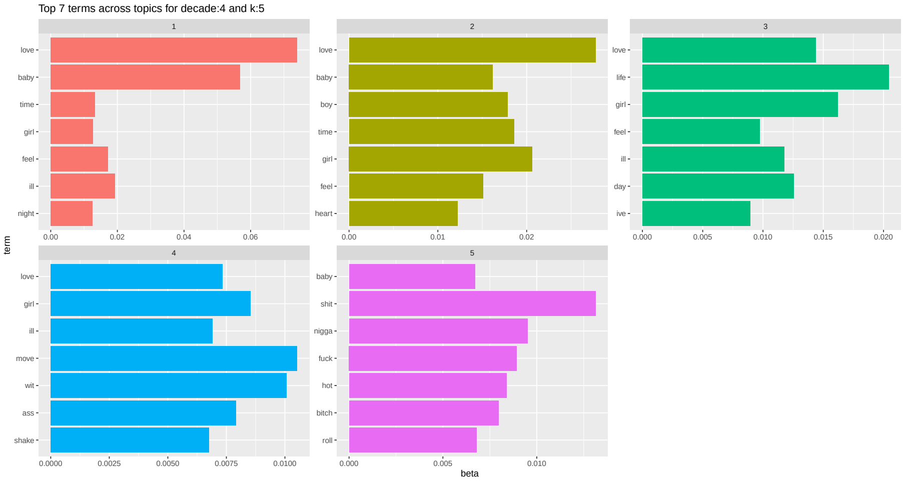
```
## DTM from tidytext
DTM1<-  tidy_reviews1 %>%
  dplyr::count(Song, word) %>%
  cast_dtm(Song, word, n)

DTM2<-  tidy_reviews2 %>%
  dplyr::count(Song, word) %>%
  cast_dtm(Song, word, n)

DTM3<-  tidy_reviews3 %>%
  dplyr::count(Song, word) %>%
  cast_dtm(Song, word, n)

DTM4<-  tidy_reviews4 %>%
  dplyr::count(Song, word) %>%
  cast_dtm(Song, word, n)

DTM5<-  tidy_reviews5 %>%
  dplyr::count(Song, word) %>%
  cast_dtm(Song, word, n)
```

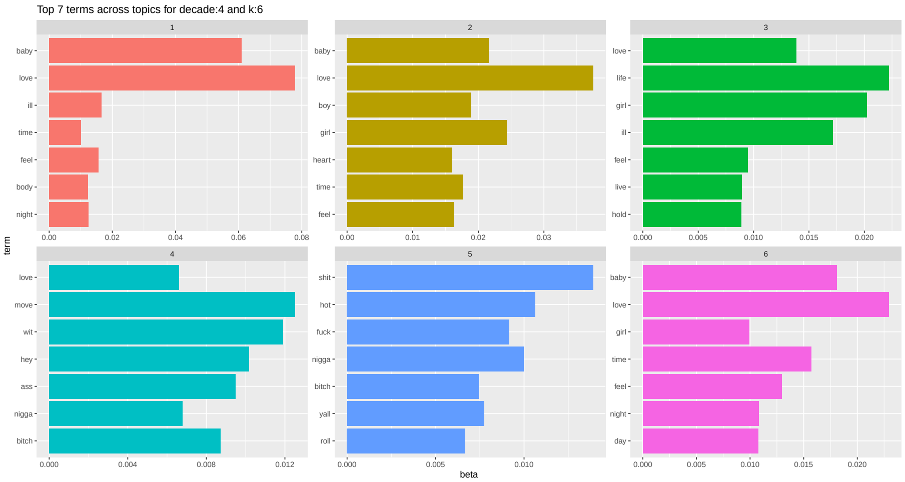
Exercise 4: Choose a single dataset and run three models to try and identify an appropriate value for k (the number of topics). State which value of k you choose after running these three models as well as why you picked those particular three values of k to run for each of your models.

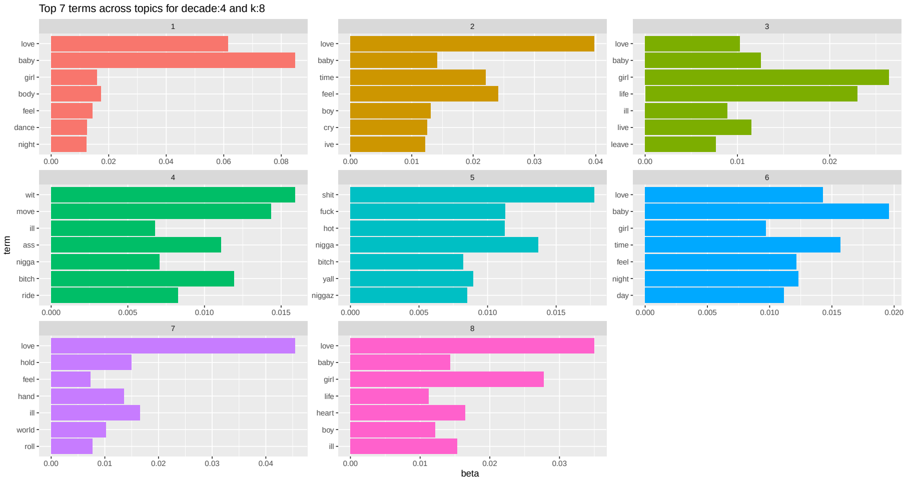
```
## fit models... done.
## calculate metrics:
##   Griffiths2004... done.
##   CaoJuan2009... done.
##   Arun2010... done.
##   Deveaud2014... done.
```



I will be trying for three values of k: (5,6,8). I selected these values based on the plot obtained above. I looked at the minima and maxima of the metrics described above and made these choices. Growth from 8 to 10 in the metric 'Griffiths2004' to be maximized is not that steep. Hence, I am choosing 8 instead of 10.







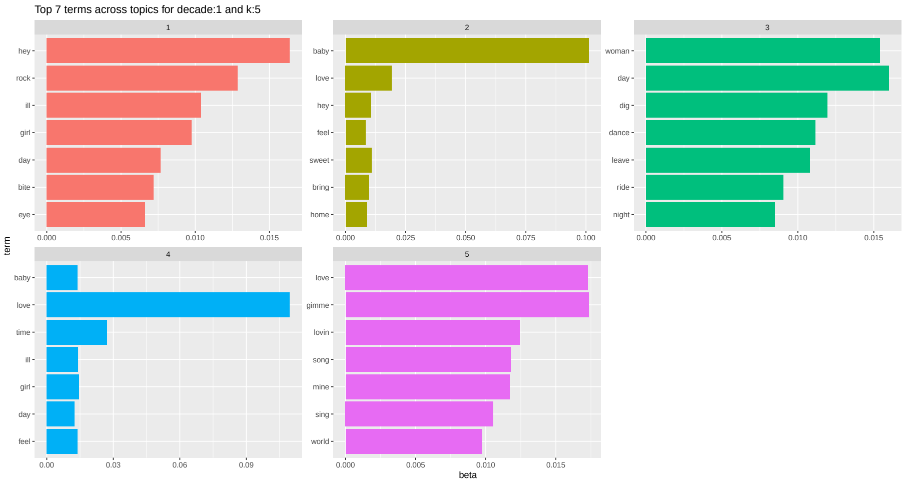
I am selecting k=5 as the 5th topic shows some diversity in the sense it has words like 'bitch', 'nigga', 'shit' etc. which may indicate songs not suitable for all people (or age groups). These lyrics could be offensive. All other topics have similar phrases. As the value of k, (for instance 6 and 8), we get repetitive topics of the same kind (as the first 4). Thus, I will extract 5 topics.

Exercise 5: Using the same value of k, run a model on each of the other decades lyrics datasets

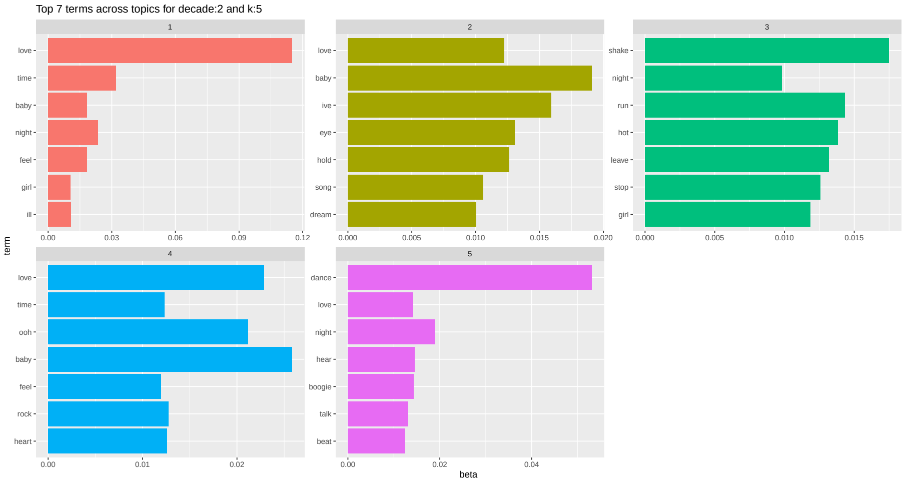
Please note that for the ease of coding, I had assigned the following number to each decade:

- 1: 1965-1974
- 2: 1975-1984
- 3: 1985-1994
- 4: 1995-2004
- 5: 2005-2014

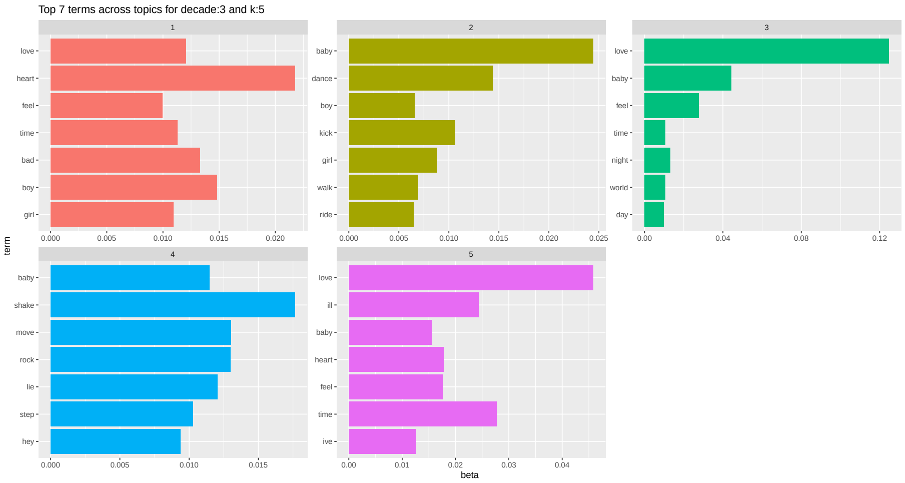
```
## [1] "For decade (1) 1965-1974..."
```



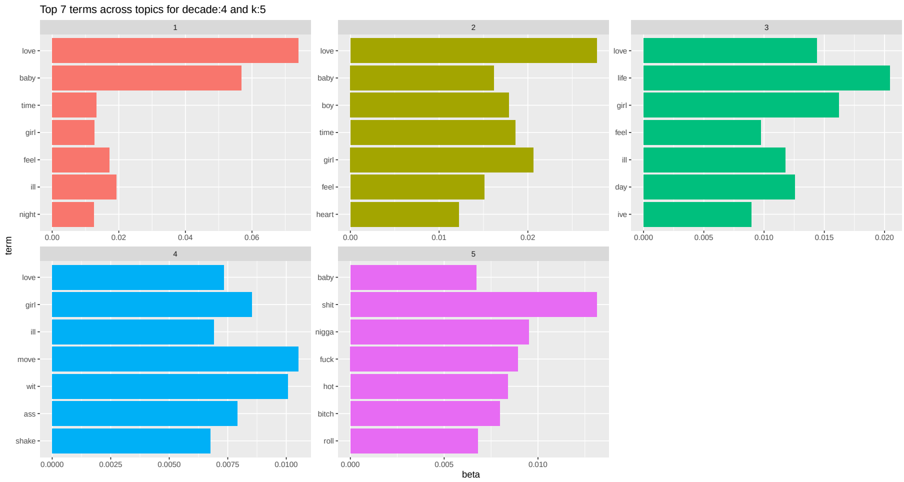
```
## [1] "For decade (2) 1975-1984..."
```



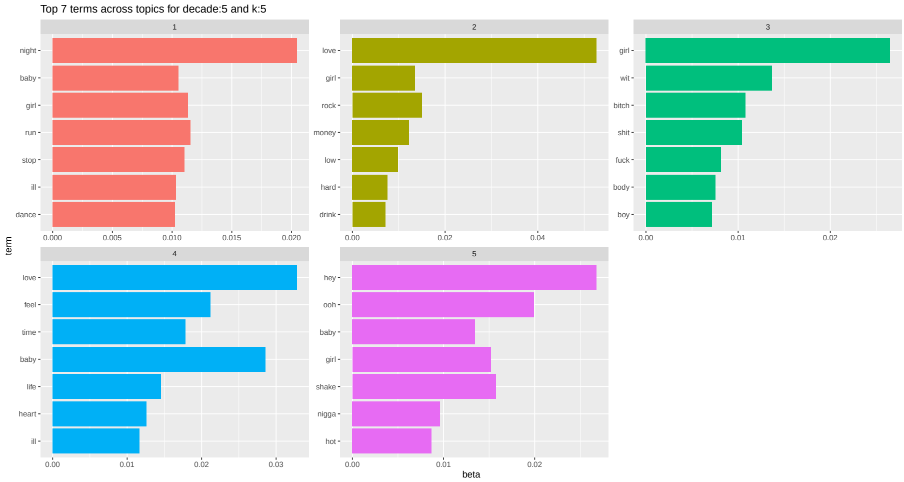
```
## [1] "For decade (3) 1985-1994..."
```



```
## [1] "For decade (4) 1995-2004..."
```



```
## [1] "For decade (5) 2005-2014..."
```



Exercise 6: Based on your output, does it seem like your value of k was a good choice for all decades of lyrics?

The value k=5 has created topics of varying degrees of uniqueness across topics. In some decades, it was not as successful. However, it seems to have quite well worked for a few other decades. For example, for decade 3 i.e.1985-1994, the topics seems to be very similar (except topic 4 which might contain words implying dance/party songs).

However in decade 2 i.e. 1975-1984, the topics seems to be quite disparate with different top words.

I believe it is also the kind of lyrics most songs have. It is difficult to create unique topics out of song lyrics as the genres will be similar. I believe, k=5 did a fair job of creating different topics out of song lyrics across decades. Although, it can be improved using more refined text preprocessing techniques, it was a fair choice for this example.