MIS 506 Final Project Part 2

1. Load the required libraries

library(gutenbergr)  
library(topicmodels)  
library(tidyverse)  
library(tidytext)

1. Load the review dataset into your workspace

prince\_original <-read\_csv("prince\_raw\_data.csv")

##   
## -- Column specification --------------------------------------------------------  
## cols(  
## .default = col\_character(),  
## X = col\_double(),  
## year = col\_double(),  
## peak = col\_double()  
## )  
## i Use `spec()` for the full column specifications.

#Change the name of the column "text" to "lyrics".  
prince <-prince\_original %>%   
 select (lyrics = text, song, year, album, peak, US.Pop, US.R.B)  
  
#Create a column name decade. Then store the results back into prince.  
prince <- prince %>%  
 mutate(decade =   
 ifelse(prince$year %in% 1978:1979, "1970s",   
 ifelse(prince$year %in% 1980:1989, "1980s",   
 ifelse(prince$year %in% 1990:1999, "1990s",   
 ifelse(prince$year %in% 2000:2009, "2000s",   
 ifelse(prince$year %in% 2010:2015, "2010s",   
 "NA"))))))  
#Also, create a column name chart\_level.  
prince <- prince %>%  
 mutate(chart\_level =   
 ifelse(prince$peak %in% 1:10, "Top 10",   
 ifelse(prince$peak %in% 11:100, "Top 100", "Uncharted")))

1. Subset the data into “decade” so that each new dataframe contains the lyrics and other columns from a particular decade of Prince Songs. Choose a single dataset for a specific decade. For example, the code bellow create a dataset for the songs in the 70s.

# Create a new object “decade” to read the information from the column “decade”   
# from the prince dataset  
decade <- prince$decade  
# Use the split function to split the prince dataset based on “decade”.   
# prince will be a list of prince songs for each decade.  
prince<-split(prince,decade)  
# Create a new dataframe containing the songs in the decade 70s  
decade\_70<-prince[[1]]  
decade\_80<-prince[[2]]  
decade\_70<-prince[[3]]  
decade\_90<-prince[[4]]  
decade\_00<-prince[[5]]  
decade\_NA<-prince[[6]]

1. Prepare each of the datasets so that it can be analyzed using the topicmodels package.

tidy\_decade\_70 <- decade\_70 %>%  
 unnest\_tokens(word, lyrics) %>%  
 anti\_join(stop\_words)

## Joining, by = "word"

tidy\_decade\_70\_dtm<-tidy\_decade\_70 %>%  
 count(song, word) %>%  
 cast\_dtm(song, word, n)  
  
  
tidy\_decade\_80 <- decade\_80 %>%  
 unnest\_tokens(word, lyrics) %>%  
 anti\_join(stop\_words)

## Joining, by = "word"

tidy\_decade\_80\_dtm<-tidy\_decade\_80 %>%  
 count(song, word) %>%  
 cast\_dtm(song, word, n)  
  
  
tidy\_decade\_90 <- decade\_90 %>%  
 unnest\_tokens(word, lyrics) %>%  
 anti\_join(stop\_words)

## Joining, by = "word"

tidy\_decade\_90\_dtm<-tidy\_decade\_90 %>%  
 count(song, word) %>%  
 cast\_dtm(song, word, n)  
  
  
tidy\_decade\_00 <- decade\_00 %>%  
 unnest\_tokens(word, lyrics) %>%  
 anti\_join(stop\_words)

## Joining, by = "word"

tidy\_decade\_00\_dtm<-tidy\_decade\_00 %>%  
 count(song, word) %>%  
 cast\_dtm(song, word, n)  
  
  
tidy\_decade\_NA <- decade\_NA %>%  
 unnest\_tokens(word, lyrics) %>%  
 anti\_join(stop\_words)

## Joining, by = "word"

tidy\_decade\_NA\_dtm<-tidy\_decade\_NA %>%  
 count(song, word) %>%  
 cast\_dtm(song, word, n)

1. 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.

decade\_70\_topic\_model5<-LDA(tidy\_decade\_70\_dtm, k=5, control = list(seed = 1234))  
terms(decade\_70\_topic\_model5, k=15)

## Topic 1 Topic 2 Topic 3 Topic 4 Topic 5   
## [1,] "yeah" "baby" "love" "push" "baby"   
## [2,] "baby" "love" "da" "baby" "wanna"   
## [3,] "wanna" "2" "yeah" "na" "la"   
## [4,] "dark" "yeah" "2" "2" "rock"   
## [5,] "hot" "wanna" "style" "yeah" "music"   
## [6,] "freaks" "shake" "4" "ooh" "world"   
## [7,] "time" "space" "live" "jam" "wrong"   
## [8,] "wit" "time" "poompoom" "love" "yeah"   
## [9,] "dig" "gonna" "time" "joint" "funky"   
## [10,] "light" "cuz" "joy" "wanna" "love"   
## [11,] "dead" "pheromone" "prince" "time" "uh"   
## [12,] "heart" "feel" "baby" "grind" "play"   
## [13,] "blue" "ooh" "daddy" "everybody's" "hard"   
## [14,] "ooh" "mind" "pop" "gold" "time"   
## [15,] "jughead" "acknowledge" "hey" "rave" "everyday"

decade\_70\_topic\_model4<-LDA(tidy\_decade\_70\_dtm, k=4, control = list(seed = 1234))  
terms(decade\_70\_topic\_model4, k=15)

## Topic 1 Topic 2 Topic 3 Topic 4  
## [1,] "yeah" "baby" "love" "baby"   
## [2,] "baby" "wanna" "da" "push"   
## [3,] "wanna" "yeah" "yeah" "na"   
## [4,] "dark" "love" "2" "yeah"   
## [5,] "time" "2" "style" "2"   
## [6,] "hot" "music" "4" "la"   
## [7,] "freaks" "gonna" "time" "ooh"   
## [8,] "love" "shake" "live" "love"   
## [9,] "girl" "space" "baby" "wanna"  
## [10,] "break" "funky" "poompoom" "jam"   
## [11,] "world" "time" "prince" "rock"   
## [12,] "light" "uh" "joy" "joint"  
## [13,] "wit" "cindy" "hey" "wrong"  
## [14,] "groove" "race" "daddy" "time"   
## [15,] "ooh" "sweet" "cool" "grind"

decade\_70\_topic\_model3<-LDA(tidy\_decade\_70\_dtm, k=3, control = list(seed = 1234))  
terms(decade\_70\_topic\_model3, k=15)

## Topic 1 Topic 2 Topic 3   
## [1,] "baby" "baby" "love"   
## [2,] "yeah" "wanna" "2"   
## [3,] "push" "yeah" "da"   
## [4,] "wanna" "2" "yeah"   
## [5,] "time" "la" "na"   
## [6,] "dark" "rock" "4"   
## [7,] "ooh" "love" "style"   
## [8,] "love" "funky" "baby"   
## [9,] "hot" "music" "time"   
## [10,] "break" "gonna" "live"   
## [11,] "jam" "uh" "joy"   
## [12,] "freaks" "shake" "poompoom"  
## [13,] "body" "grind" "prince"   
## [14,] "girl" "time" "joint"   
## [15,] "stop" "space" "soul"

Based on the lyrics of Prince and his choice of song themes, I feel like his subject matter tends to go deep into a few topics rather than a wide range of topics. I chose 3, 4, and 5 for that reason. After looking at the output, I feel like 3 is appropriate. Topic 1 appears to be romance, Topic 2 looks to be about dancing and music, and Topic 3 appears to be a catch-all for all things Prince

1. Using the same value of k, run a model on at least two other decades’ datasets.

decade\_80\_topic\_model3<-LDA(tidy\_decade\_80\_dtm, k=3, control = list(seed = 1234))  
terms(decade\_80\_topic\_model3, k=15)

## Topic 1 Topic 2 Topic 3   
## [1,] "baby" "love" "yeah"   
## [2,] "love" "yeah" "da"   
## [3,] "time" "party" "dance"   
## [4,] "yeah" "wanna" "na"   
## [5,] "gotta" "time" "life"   
## [6,] "crush" "baby" "baby"   
## [7,] "housequake" "trust" "love"   
## [8,] "free" "god" "gonna"   
## [9,] "die" "gonna" "ooh"   
## [10,] "york" "mind" "2"   
## [11,] "purple" "ya" "music"   
## [12,] "rain" "head" "hot"   
## [13,] "critics" "joy" "alright"  
## [14,] "lemon" "night" "world"   
## [15,] "matic" "til" "la"

decade\_90\_topic\_model3<-LDA(tidy\_decade\_90\_dtm, k=3, control = list(seed = 1234))  
terms(decade\_90\_topic\_model3, k=15)

## Topic 1 Topic 2 Topic 3   
## [1,] "love" "eye" "boat"   
## [2,] "eye" "baby" "feel"   
## [3,] "time" "wanna" "chelsea"   
## [4,] "call" "love" "2"   
## [5,] "children" "candy" "people"   
## [6,] "night" "dance" "girl"   
## [7,] "rainbow" "mellow" "love"   
## [8,] "mama" "workin" "earth"   
## [9,] "uh" "gonna" "gonna"   
## [10,] "eye'd" "ur" "everlasting"  
## [11,] "wanna" "beautiful" "run"   
## [12,] "3121" "yeah" "call"   
## [13,] "gonna" "girl" "god"   
## [14,] "rise" "future" "cinnamon"   
## [15,] "wall" "party" "bring"

1. Examine and visualize the beta and gamma matrices First the Beta for the 70s

library(reshape2)

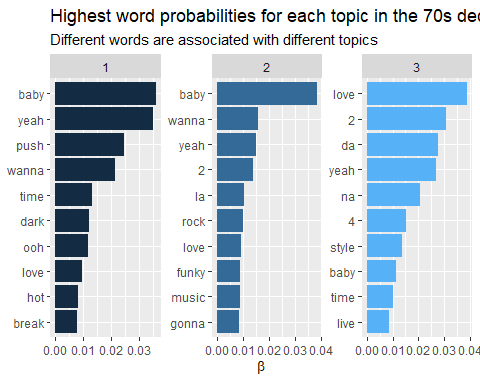
## Warning: package 'reshape2' was built under R version 4.0.5

##   
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':  
##   
## smiths

decade\_70\_topics <- tidy(decade\_70\_topic\_model3, matrix = "beta")  
  
decade\_70\_topics%>%  
 #let's group\_by each topic  
 group\_by(topic)%>%  
 #take the top 10 words in each topic  
 top\_n(10)%>%  
 ungroup%>%  
 mutate(term = reorder\_within(term, beta, topic)) %>%  
 ggplot(aes(term, beta, fill = topic)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~ topic, scales = "free") +  
 coord\_flip() +  
 labs(x = NULL, y = expression(beta),  
 title = "Highest word probabilities for each topic in the 70s decade",  
 subtitle = "Different words are associated with different topics")+  
 scale\_x\_reordered()

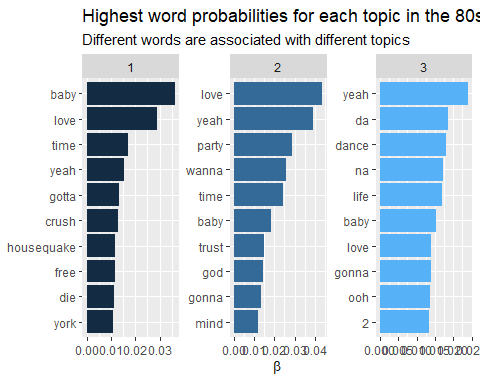
## Selecting by beta



Beta for the 80s

decade\_80\_topics <- tidy(decade\_80\_topic\_model3, matrix = "beta")  
  
decade\_80\_topics%>%  
 #let's group\_by each topic  
 group\_by(topic)%>%  
 #take the top 10 words in each topic  
 top\_n(10)%>%  
 ungroup%>%  
 mutate(term = reorder\_within(term, beta, topic)) %>%  
 ggplot(aes(term, beta, fill = topic)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~ topic, scales = "free") +  
 coord\_flip() +  
 labs(x = NULL, y = expression(beta),  
 title = "Highest word probabilities for each topic in the 80s decade",  
 subtitle = "Different words are associated with different topics")+  
 scale\_x\_reordered()

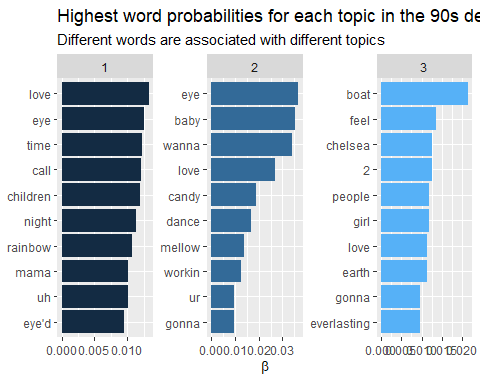
## Selecting by beta



Beta for the 90s

library(reshape2)  
decade\_90\_topics <- tidy(decade\_90\_topic\_model3, matrix = "beta")  
  
decade\_90\_topics%>%  
 #let's group\_by each topic  
 group\_by(topic)%>%  
 #take the top 10 words in each topic  
 top\_n(10)%>%  
 ungroup%>%  
 mutate(term = reorder\_within(term, beta, topic)) %>%  
 ggplot(aes(term, beta, fill = topic)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~ topic, scales = "free") +  
 coord\_flip() +  
 labs(x = NULL, y = expression(beta),  
 title = "Highest word probabilities for each topic in the 90s decade",  
 subtitle = "Different words are associated with different topics")+  
 scale\_x\_reordered()

## Selecting by beta

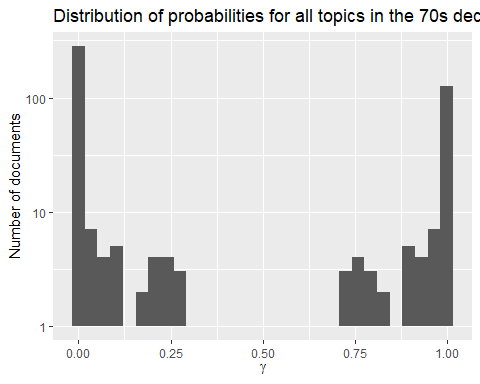


Now the Gamma The 70s Decade

decade\_70\_td\_gamma <- tidy(decade\_70\_topic\_model3, matrix = "gamma", document\_names = rownames(tidy\_decade\_70\_dtm))

Distribution of probabilities for all topics: The 70s Decade

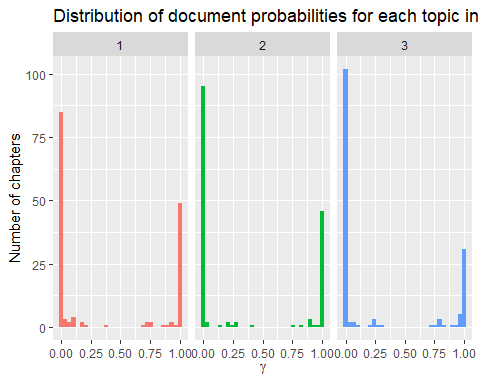
ggplot(decade\_70\_td\_gamma , aes(gamma)) +  
geom\_histogram() +  
scale\_y\_log10() +  
labs(title = "Distribution of probabilities for all topics in the 70s decade",  
y = "Number of documents", x = expression(gamma))



Distribution of document probabilities for each topic: The 70s Decade

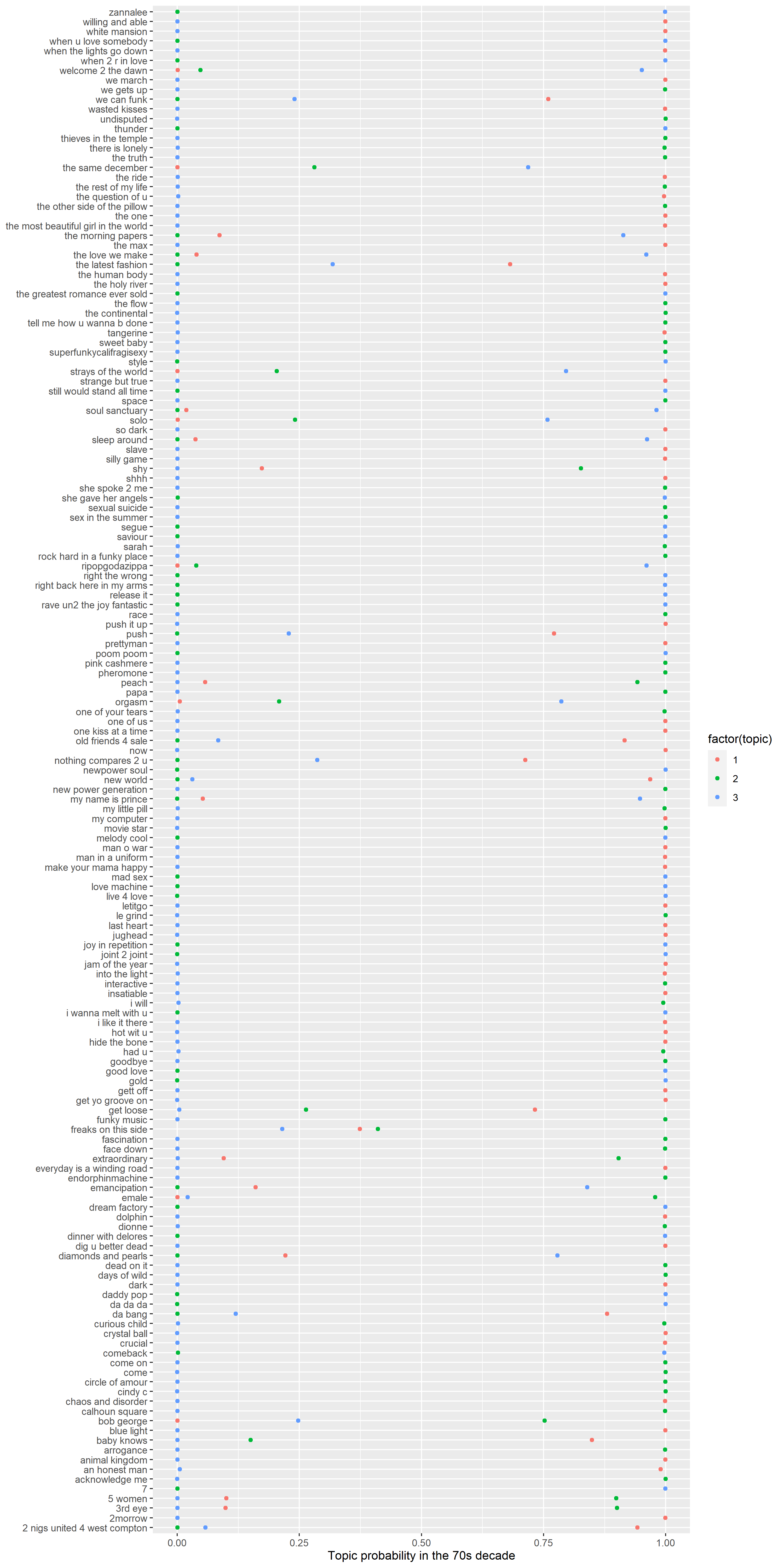
ggplot(decade\_70\_td\_gamma, aes(gamma, fill = as.factor(topic))) +  
 geom\_histogram(show.legend = FALSE) +  
 facet\_wrap(~ topic, ncol = 3) +  
 labs(title = "Distribution of document probabilities for each topic in the 70s decade",  
 y = "Number of chapters", x = expression(gamma))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Distribution of topic probability for each document: The 70s Decade

p1 <- ggplot(decade\_70\_td\_gamma, aes(x=document, y=gamma)) +   
geom\_point(aes(color=factor(topic))) +   
 labs(x=NULL, y="Topic probability in the 70s decade") +  
 coord\_flip()  
  
ggsave(p1, filename = "70s\_gamma.png", height = 20, width = 10)  
knitr::include\_graphics("70s\_gamma.png")

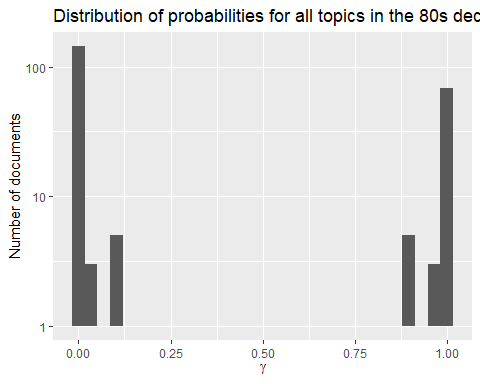


The 80s Decade

decade\_80\_td\_gamma <- tidy(decade\_80\_topic\_model3, matrix = "gamma", document\_names = rownames(tidy\_decade\_80\_dtm))

Distribution of probabilities for all topics: The 80s Decade

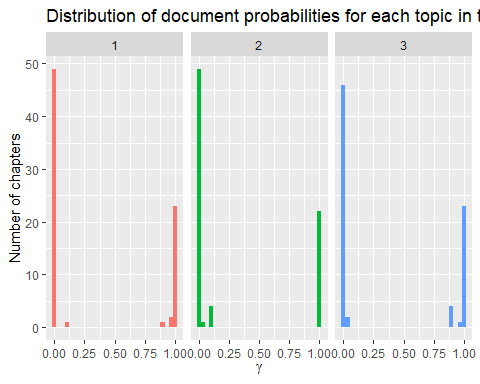
ggplot(decade\_80\_td\_gamma , aes(gamma)) +  
geom\_histogram() +  
scale\_y\_log10() +  
labs(title = "Distribution of probabilities for all topics in the 80s decade",  
y = "Number of documents", x = expression(gamma))



Distribution of document probabilities for each topic: The 80s Decade

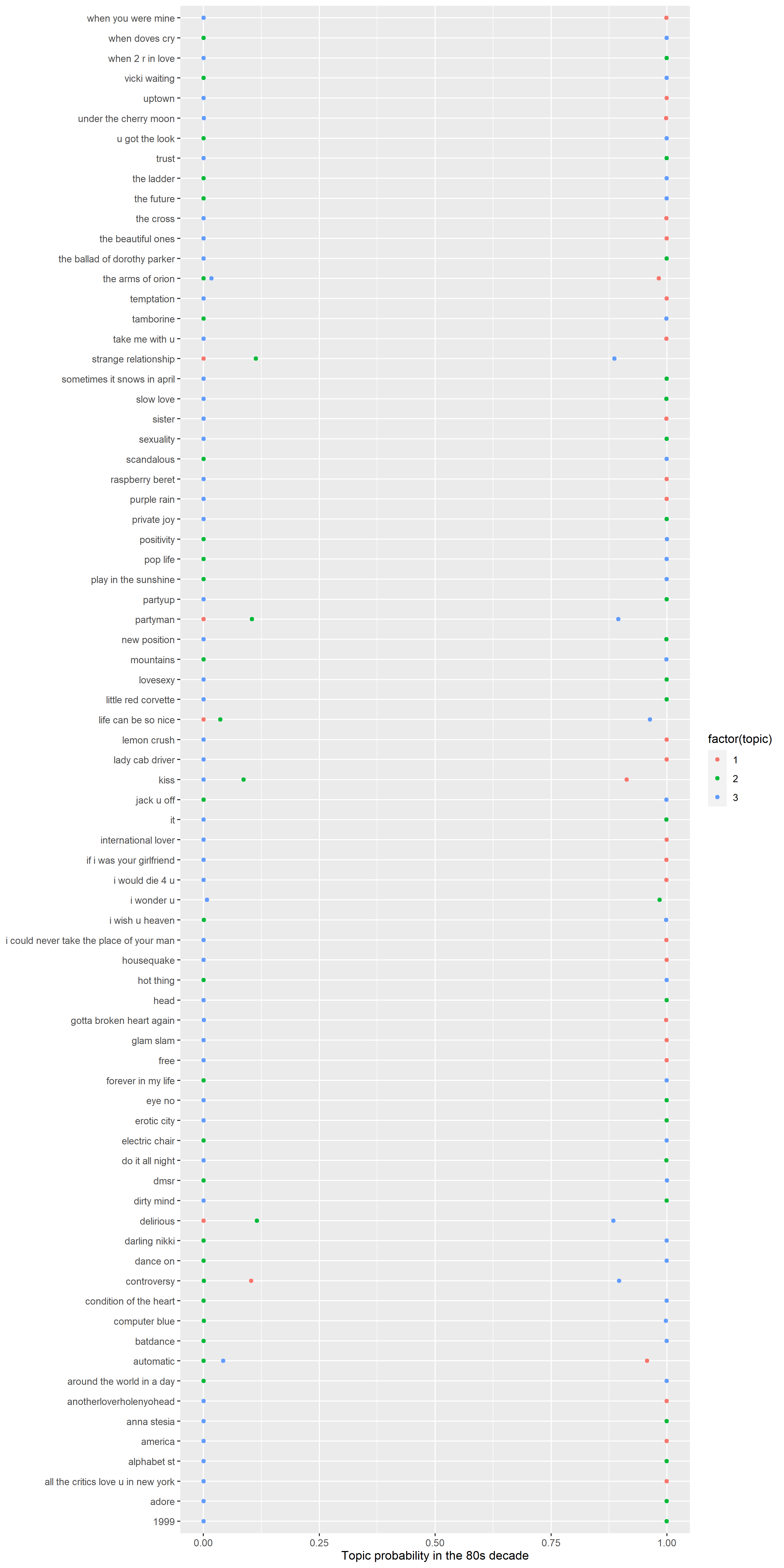
ggplot(decade\_80\_td\_gamma, aes(gamma, fill = as.factor(topic))) +  
 geom\_histogram(show.legend = FALSE) +  
 facet\_wrap(~ topic, ncol = 3) +  
 labs(title = "Distribution of document probabilities for each topic in the 80s decade",  
 y = "Number of chapters", x = expression(gamma))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Distribution of topic probability for each document: The 80s Decade

p1 <- ggplot(decade\_80\_td\_gamma, aes(x=document, y=gamma)) +   
geom\_point(aes(color=factor(topic))) +   
 labs(x=NULL, y="Topic probability in the 80s decade") +  
 coord\_flip()  
  
ggsave(p1, filename = "80s\_gamma.png", height = 20, width = 10)  
knitr::include\_graphics("80s\_gamma.png")

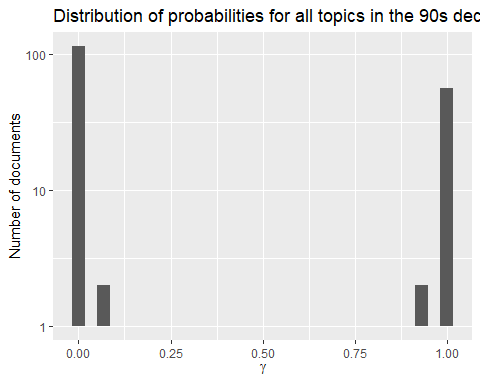


The 90s Decade

decade\_90\_td\_gamma <- tidy(decade\_90\_topic\_model3, matrix = "gamma", document\_names = rownames(tidy\_decade\_90\_dtm))

Distribution of probabilities for all topics: The 90s Decade

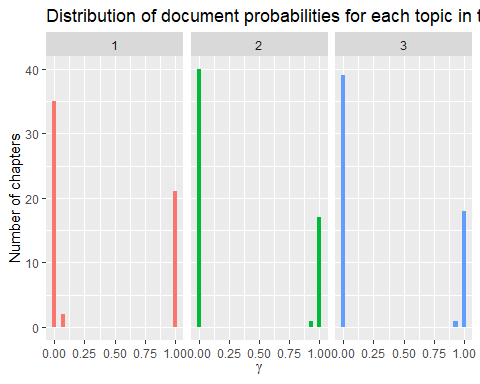
ggplot(decade\_90\_td\_gamma , aes(gamma)) +  
geom\_histogram() +  
scale\_y\_log10() +  
labs(title = "Distribution of probabilities for all topics in the 90s decade",  
y = "Number of documents", x = expression(gamma))



Distribution of document probabilities for each topic: The 90s Decade

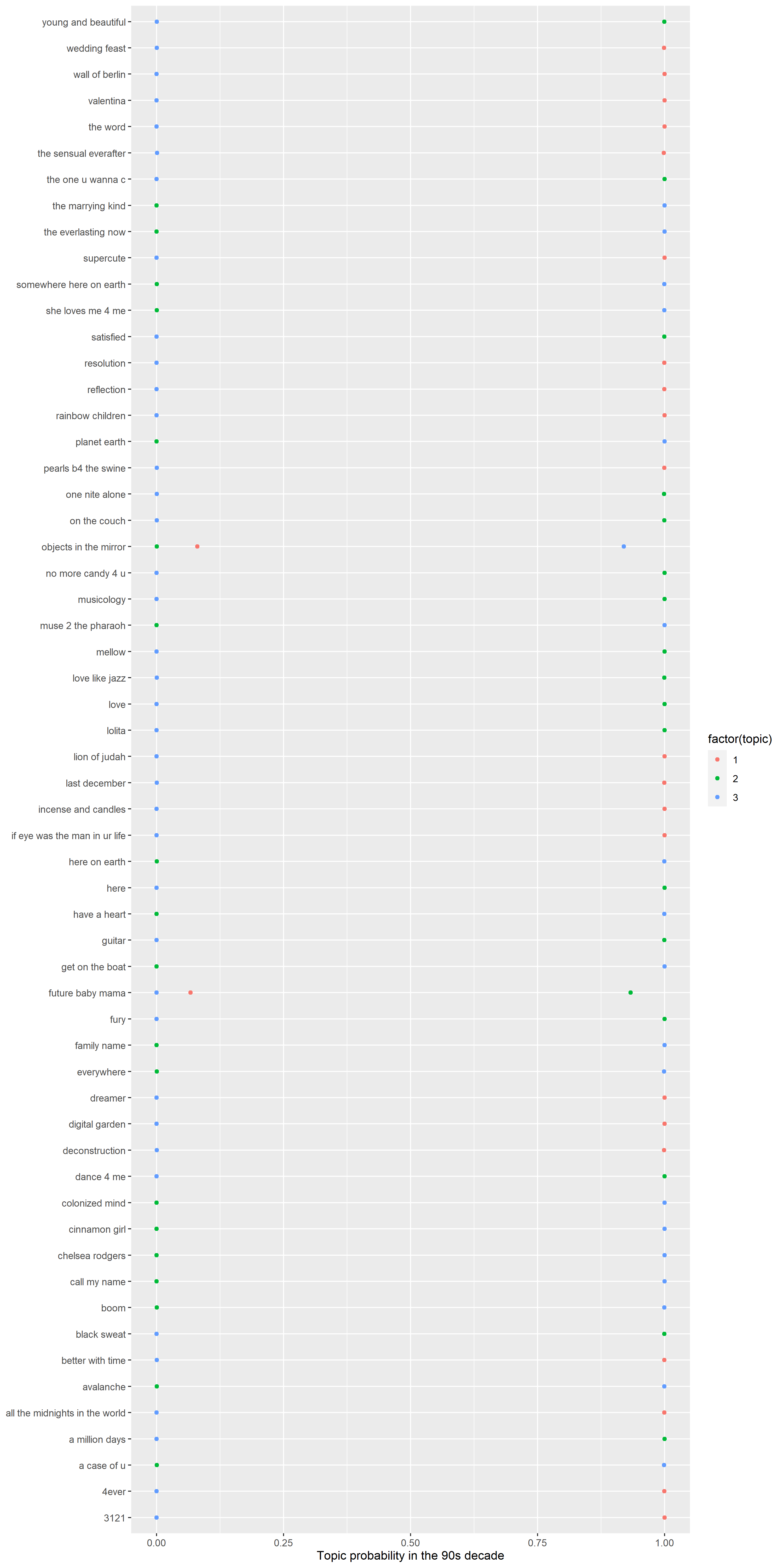
ggplot(decade\_90\_td\_gamma, aes(gamma, fill = as.factor(topic))) +  
 geom\_histogram(show.legend = FALSE) +  
 facet\_wrap(~ topic, ncol = 3) +  
 labs(title = "Distribution of document probabilities for each topic in the 90s decade",  
 y = "Number of chapters", x = expression(gamma))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Distribution of topic probability for each document: The 90s Decade

p1 <- ggplot(decade\_90\_td\_gamma, aes(x=document, y=gamma)) +   
geom\_point(aes(color=factor(topic))) +   
 labs(x=NULL, y="Topic probability in the 90s decade") +  
 coord\_flip()  
  
ggsave(p1, filename = "90s\_gamma.png", height = 20, width = 10)  
knitr::include\_graphics("90s\_gamma.png")



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

I think so. Most of the songs seem to fit nicely into one of the topics, with some expected overlap. Choosing a high value for k is acceptable when you have complex subject matter, but the lyrics of Prince songs aren’t exactly thought provoking and insightful, nor were they meant to be. I think 3 topics is enough for dealing with Prince lyrics throughout the decades.