Final project part 2

Jonathan Broadbridge

5/3/2021

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

fix.contractions <- function(doc) {  
 doc <- gsub("won't", "will not", doc)  
 doc <- gsub("can't", "can not", doc)  
 doc <- gsub("n't", " not", doc)  
 doc <- gsub("'ll", " will", doc)  
 doc <- gsub("'re", " are", doc)  
 doc <- gsub("'ve", " have", doc)  
 doc <- gsub("'m", " am", doc)  
 doc <- gsub("'d", " would", doc)  
 doc <- gsub("'s", "", doc)  
 return(doc)  
}  
  
  
  
  
prince <-prince\_original %>%   
 select (lyrics = text, song, year, album, peak, US.Pop, US.R.B)  
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"))))))  
  
prince <-prince %>%  
 mutate(chart\_level =   
 ifelse(prince$peak %in% 1:10, "Top 10",   
 ifelse(prince$peak %in% 11:100, "Top 100", "Uncharted")))  
prince$lyrics <- sapply(prince$lyrics, fix.contractions)  
decade <-prince$decade  
prince<-split(prince,decade)  
decade\_70<-prince[[1]]  
decade\_80<-prince[[2]]  
decade\_90<-prince[[3]]  
decade\_00<-prince[[4]]  
decade\_10<-prince[[5]]  
  
  
  
undesirable\_words <- c("prince", "chorus", "repeat", "lyrics","theres", "bridge", "fe0f", "yeah", "baby",   
"alright", "wanna", "gonna", "chorus", "verse",   
"whoa", "gotta", "make", "miscellaneous", "2",   
"4", "ooh", "uurh", "pheromone", "poompoom", "3121",   
"matic", " ai ", " ca ", " la ", "hey", " na ",   
 " da ", " uh ", " tin ", " ll", "transcription", "repeats")  
  
tidy\_prince70 <-decade\_70 %>%  
 unnest\_tokens("word",lyrics)%>%  
 anti\_join(stop\_words)%>%  
 filter (!word %in% undesirable\_words) %>%  
 filter(nchar(word) > 2)

## Warning: Outer names are only allowed for unnamed scalar atomic inputs

## Joining, by = "word"

tidy\_prince70$word <- gsub("\\s+","", tidy\_prince70$word)  
tidy\_prince70$word <- gsub("[^a-zA-Z]","", tidy\_prince70$word)  
  
prince70dtm <- tidy\_prince70 %>%  
 count(song, word) %>%  
 cast\_dtm(song, word, n)  
  
topic\_model1<-LDA(prince70dtm, k=6, control = list(seed = 1234))  
topic\_model1

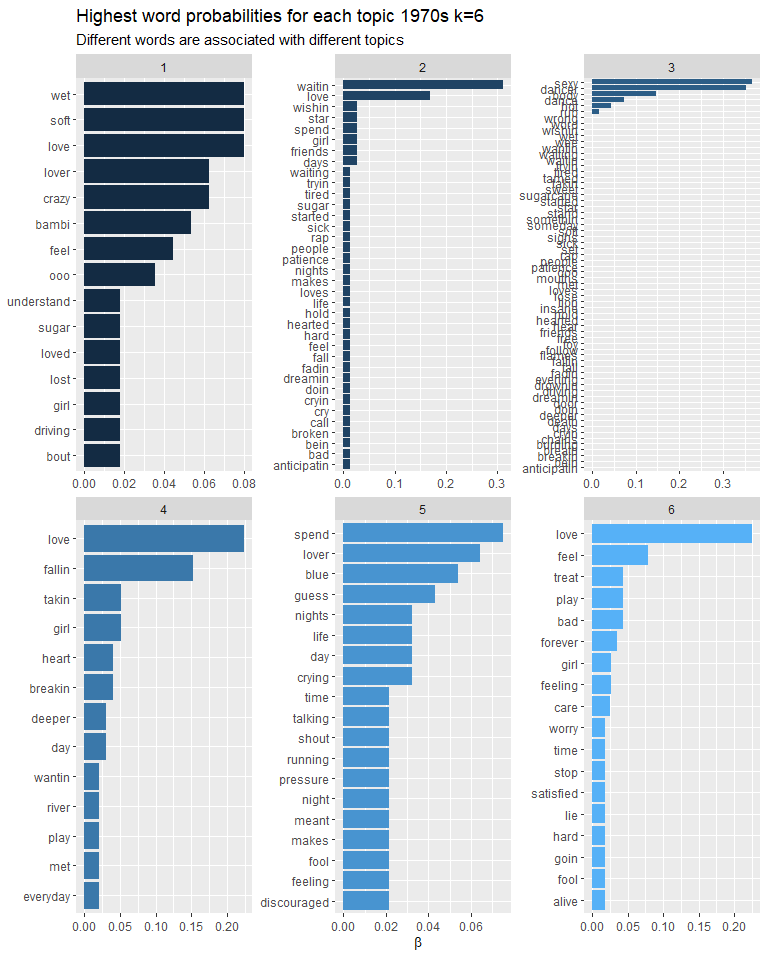
## A LDA\_VEM topic model with 6 topics.

topics1 <- tidy(topic\_model1, matrix = "beta")  
topics1

## # A tibble: 1,134 x 3  
## topic term beta  
## <int> <chr> <dbl>  
## 1 1 adore 5.94e-319  
## 2 2 adore 2.58e-315  
## 3 3 adore 1.76e-314  
## 4 4 adore 1.02e- 2  
## 5 5 adore 4.34e-300  
## 6 6 adore 1.05e-318  
## 7 1 barely 6.67e-319  
## 8 2 barely 3.51e-315  
## 9 3 barely 1.61e-314  
## 10 4 barely 1.02e- 2  
## # ... with 1,124 more rows

topics1%>%  
 group\_by(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 1970s k=6",  
 subtitle = "Different words are associated with different topics")+  
 scale\_x\_reordered()

## Selecting by beta



topic\_model2<-LDA(prince70dtm, k=8, control = list(seed = 1234))  
topic\_model2

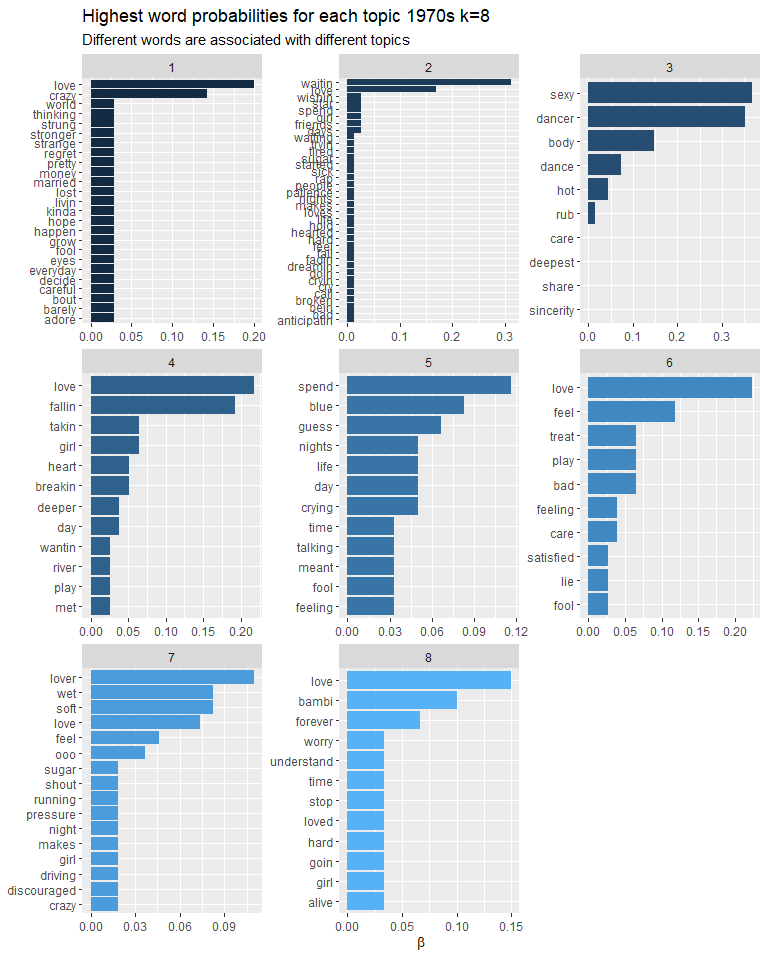
## A LDA\_VEM topic model with 8 topics.

topics2 <- tidy(topic\_model2, matrix = "beta")  
topics2

## # A tibble: 1,512 x 3  
## topic term beta  
## <int> <chr> <dbl>  
## 1 1 adore 2.86e- 2  
## 2 2 adore 5.09e-181  
## 3 3 adore 8.36e-181  
## 4 4 adore 4.83e-181  
## 5 5 adore 1.38e-180  
## 6 6 adore 5.36e-181  
## 7 7 adore 1.27e-181  
## 8 8 adore 1.38e-180  
## 9 1 barely 2.86e- 2  
## 10 2 barely 5.09e-181  
## # ... with 1,502 more rows

topics2%>%  
 group\_by(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 1970s k=8",  
 subtitle = "Different words are associated with different topics")+  
 scale\_x\_reordered()

## Selecting by beta



topic\_model3<-LDA(prince70dtm, k=10, control = list(seed = 1234))  
topic\_model3

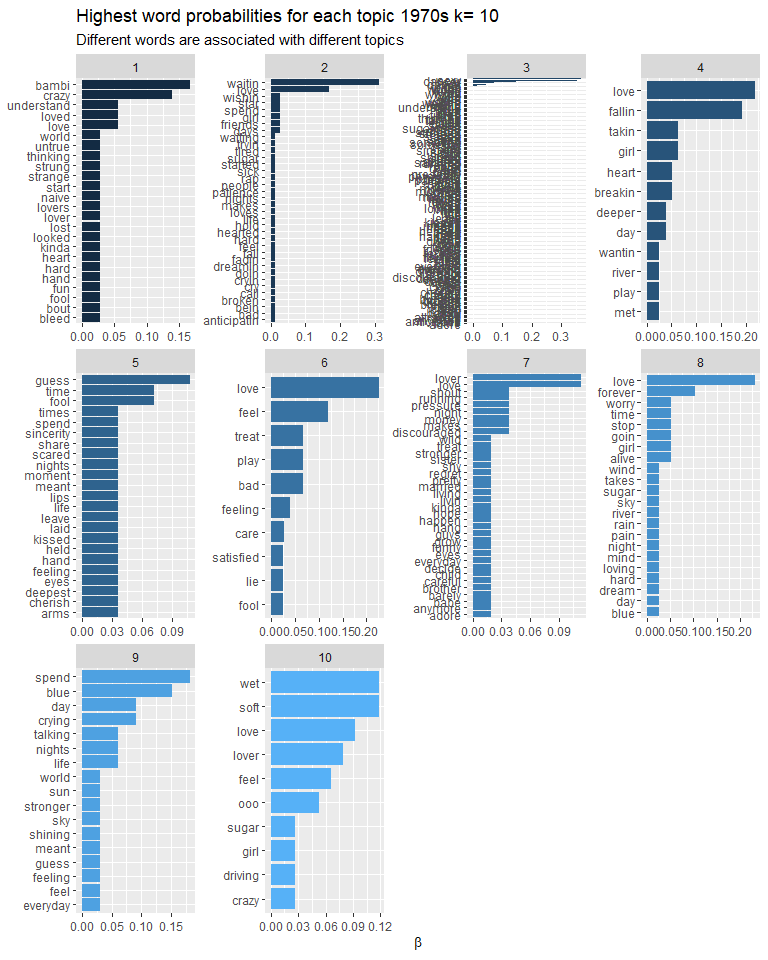
## A LDA\_VEM topic model with 10 topics.

topics3 <- tidy(topic\_model3, matrix = "beta")  
topics3

## # A tibble: 1,890 x 3  
## topic term beta  
## <int> <chr> <dbl>  
## 1 1 adore 3.72e-44  
## 2 2 adore 3.72e-44  
## 3 3 adore 3.72e-44  
## 4 4 adore 3.72e-44  
## 5 5 adore 3.72e-44  
## 6 6 adore 3.72e-44  
## 7 7 adore 1.89e- 2  
## 8 8 adore 3.72e-44  
## 9 9 adore 3.72e-44  
## 10 10 adore 3.72e-44  
## # ... with 1,880 more rows

topics3%>%  
 group\_by(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 1970s k= 10",  
 subtitle = "Different words are associated with different topics")+  
 scale\_x\_reordered()

## Selecting by beta



### Topic model Summary

After running threee sample topic models I have decided to select the topic model with a k value of 8. This topic model had the most consistently high beta values represented. The 6 topic model had lower beta values across the board and the 10 topic model had 8 topics with a high beta and 2 with low beta.

tidy\_prince80 <-decade\_80 %>%  
 unnest\_tokens("word",lyrics)%>%  
 anti\_join(stop\_words)%>%  
 filter (!word %in% undesirable\_words) %>%  
 filter(nchar(word) > 2)

## Warning: Outer names are only allowed for unnamed scalar atomic inputs

## Joining, by = "word"

tidy\_prince80$word <- gsub("\\s+","", tidy\_prince80$word)  
tidy\_prince80$word <- gsub("[^a-zA-Z]","", tidy\_prince80$word)  
  
prince80dtm <- tidy\_prince80 %>%  
 count(song, word) %>%  
 cast\_dtm(song, word, n)  
  
topic\_model4<-LDA(prince80dtm, k=8, control = list(seed = 1234))  
topic\_model4

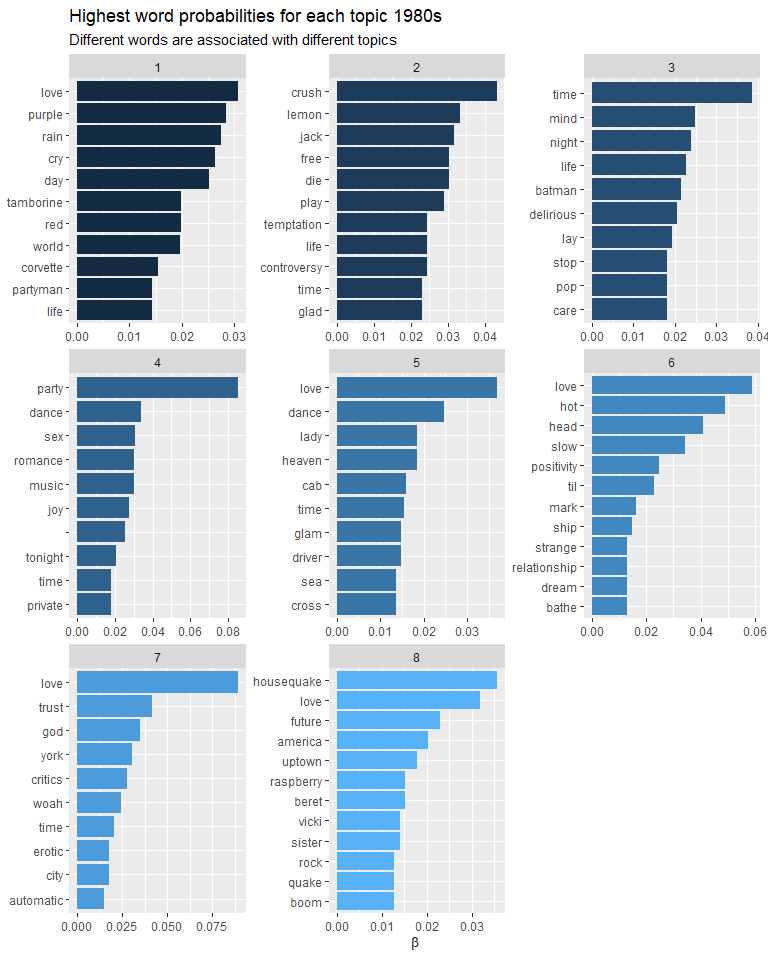
## A LDA\_VEM topic model with 8 topics.

topics4 <- tidy(topic\_model4, matrix = "beta")  
topics4

## # A tibble: 12,776 x 3  
## topic term beta  
## <int> <chr> <dbl>  
## 1 1 "" 7.43e-266  
## 2 2 "" 2.10e-267  
## 3 3 "" 1.14e- 3  
## 4 4 "" 2.52e- 2  
## 5 5 "" 3.04e-267  
## 6 6 "" 3.54e-266  
## 7 7 "" 4.36e-269  
## 8 8 "" 8.31e-268  
## 9 1 "astray" 5.42e-273  
## 10 2 "astray" 9.58e-275  
## # ... with 12,766 more rows

topics4%>%  
 group\_by(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 1980s",  
 subtitle = "Different words are associated with different topics")+  
 scale\_x\_reordered()

## Selecting by beta



tidy\_prince90 <-decade\_90 %>%  
 unnest\_tokens("word",lyrics)%>%  
 anti\_join(stop\_words)%>%  
 filter (!word %in% undesirable\_words) %>%  
 filter(nchar(word) > 2)

## Warning: Outer names are only allowed for unnamed scalar atomic inputs

## Joining, by = "word"

tidy\_prince90$word <- gsub("\\s+","", tidy\_prince90$word)  
tidy\_prince90$word <- gsub("[^a-zA-Z]","", tidy\_prince90$word)  
  
prince90dtm <- tidy\_prince90 %>%  
 count(song, word) %>%  
 cast\_dtm(song, word, n)  
  
topic\_model5<-LDA(prince90dtm, k=8, control = list(seed = 1234))  
topic\_model5

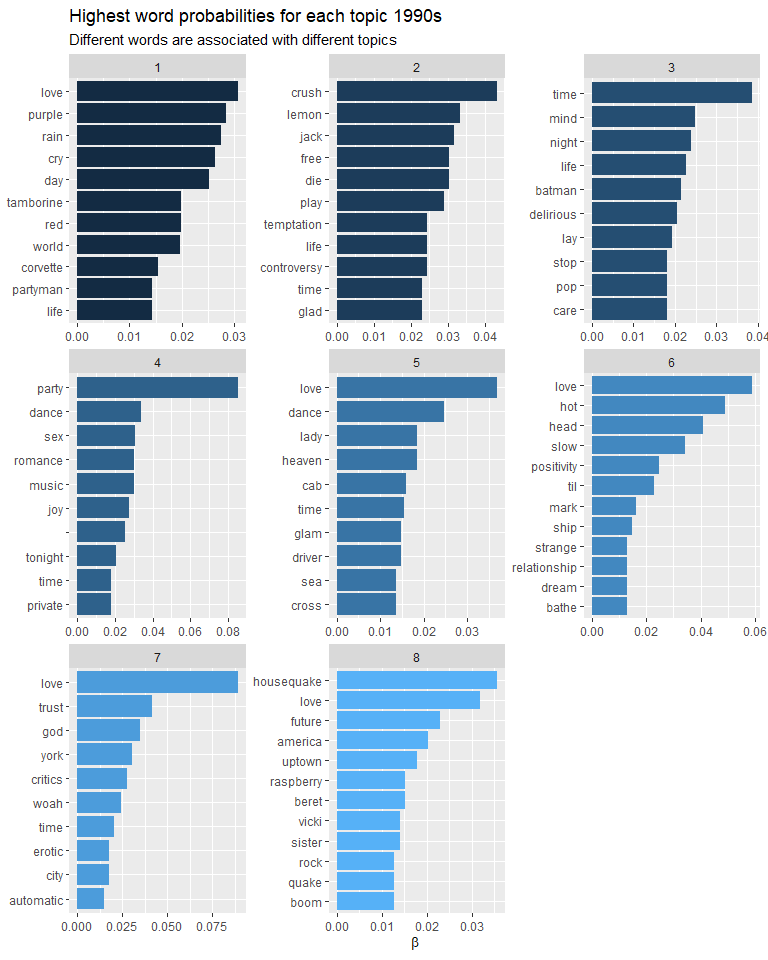
## A LDA\_VEM topic model with 8 topics.

topics5 <- tidy(topic\_model4, matrix = "beta")  
topics5

## # A tibble: 12,776 x 3  
## topic term beta  
## <int> <chr> <dbl>  
## 1 1 "" 7.43e-266  
## 2 2 "" 2.10e-267  
## 3 3 "" 1.14e- 3  
## 4 4 "" 2.52e- 2  
## 5 5 "" 3.04e-267  
## 6 6 "" 3.54e-266  
## 7 7 "" 4.36e-269  
## 8 8 "" 8.31e-268  
## 9 1 "astray" 5.42e-273  
## 10 2 "astray" 9.58e-275  
## # ... with 12,766 more rows

topics5%>%  
 group\_by(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 1990s",  
 subtitle = "Different words are associated with different topics")+  
 scale\_x\_reordered()

## Selecting by beta



td\_gamma1 <- tidy(topic\_model2, matrix = "gamma", document\_names = rownames(prince70\_dfm))  
td\_gamma1

## # A tibble: 112 x 3  
## document topic gamma  
## <chr> <int> <dbl>  
## 1 baby 1 0.996   
## 2 bambi 1 0.000577  
## 3 crazy you 1 0.994   
## 4 for you 1 0.00200   
## 5 i feel for you 1 0.000319  
## 6 i wanna be your lover 1 0.000368  
## 7 in love 1 0.000156  
## 8 my love is forever 1 0.000311  
## 9 sexy dancer 1 0.000179  
## 10 so blue 1 0.000368  
## # ... with 102 more rows

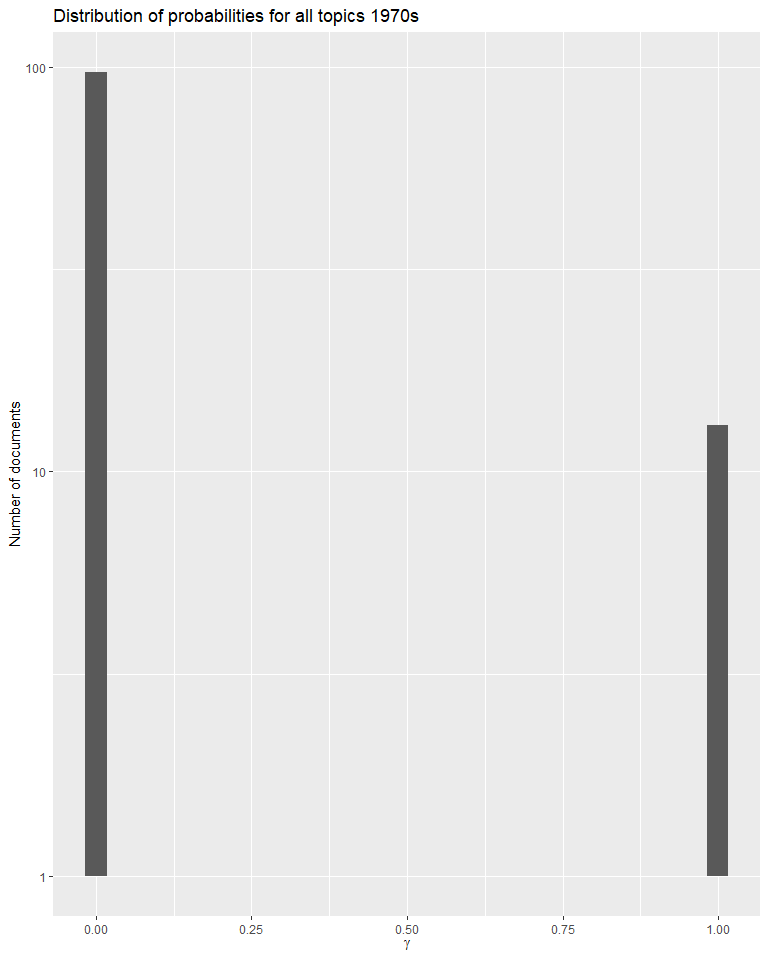
td\_gamma2 <- tidy(topic\_model4, matrix = "gamma", document\_names = rownames(prince80\_dfm))  
td\_gamma2

## # A tibble: 608 x 3  
## document topic gamma  
## <chr> <int> <dbl>  
## 1 1999 1 0.000104   
## 2 adore 1 0.000146   
## 3 all the critics love u in new york 1 0.000115   
## 4 alphabet st 1 0.000146   
## 5 america 1 0.000144   
## 6 anna stesia 1 0.0000920  
## 7 anotherloverholenyohead 1 0.000237   
## 8 around the world in a day 1 0.999   
## 9 automatic 1 0.000146   
## 10 batdance 1 0.000129   
## # ... with 598 more rows

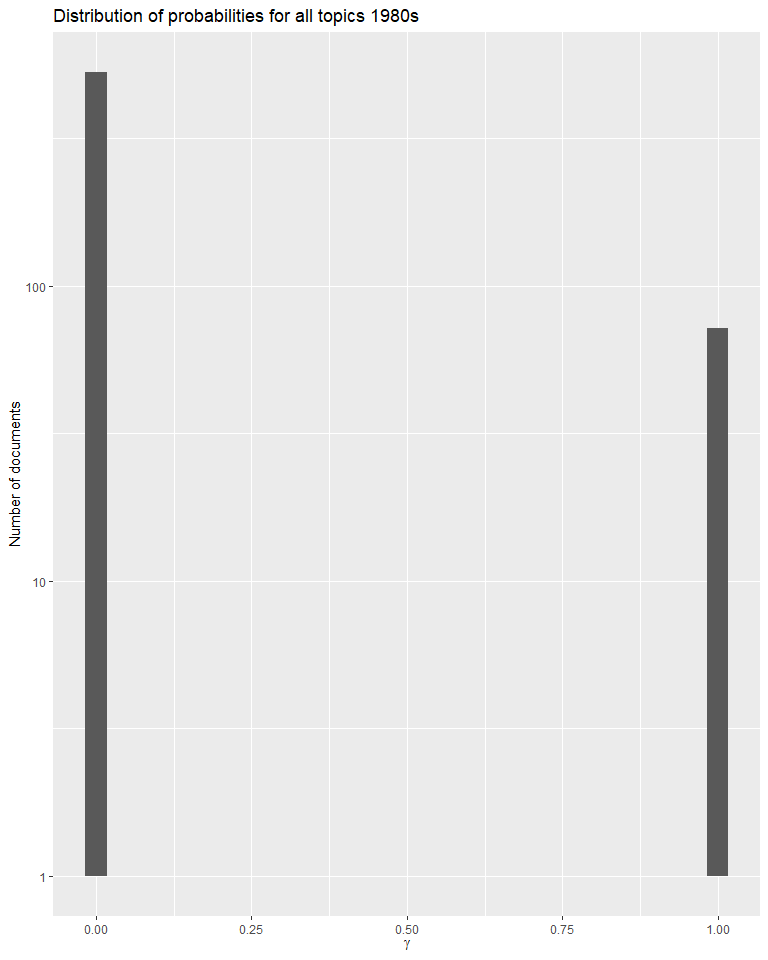
td\_gamma3 <- tidy(topic\_model5, matrix = "gamma", document\_names = rownames(prince90\_dfm))  
td\_gamma3

## # A tibble: 1,256 x 3  
## document topic gamma  
## <chr> <int> <dbl>  
## 1 2 nigs united 4 west compton 1 0.000234   
## 2 2morrow 1 0.999   
## 3 3rd eye 1 0.998   
## 4 5 women 1 0.000203   
## 5 7 1 0.999   
## 6 acknowledge me 1 0.0000876  
## 7 an honest man 1 0.00259   
## 8 animal kingdom 1 0.998   
## 9 arrogance 1 0.000348   
## 10 baby knows 1 0.000210   
## # ... with 1,246 more rows

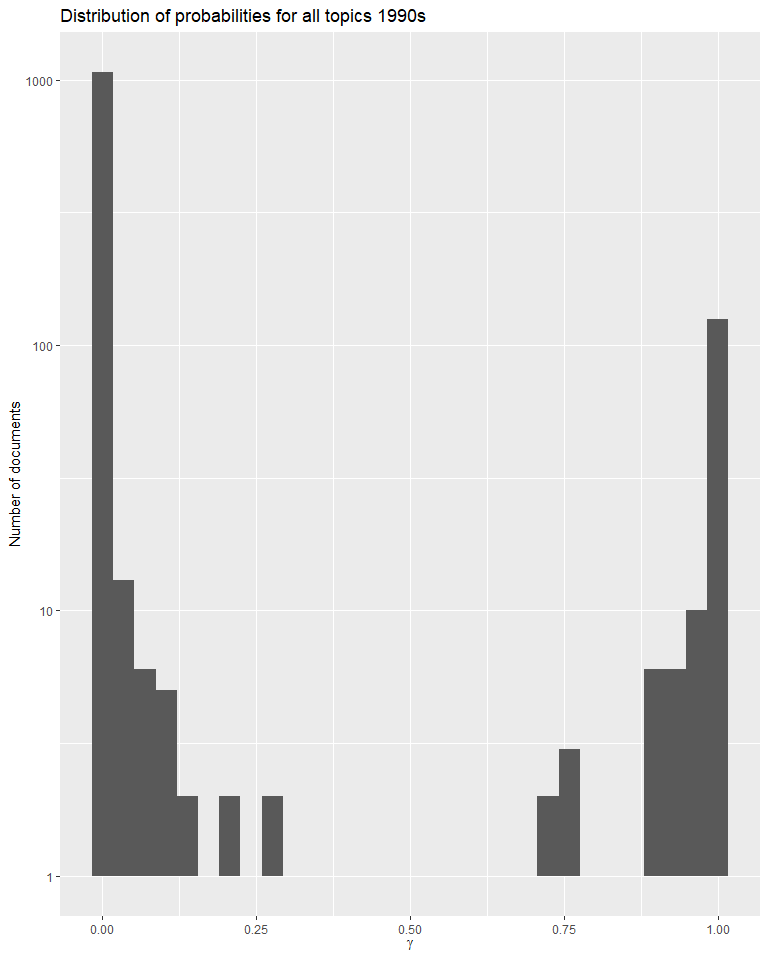
ggplot(td\_gamma1, aes(gamma)) +geom\_histogram() +  
 scale\_y\_log10() +  
 labs(title = "Distribution of probabilities for all topics 1970s",  
 y = "Number of documents", x = expression(gamma))



ggplot(td\_gamma2, aes(gamma)) +geom\_histogram() +  
 scale\_y\_log10() +  
 labs(title = "Distribution of probabilities for all topics 1980s",  
 y = "Number of documents", x = expression(gamma))

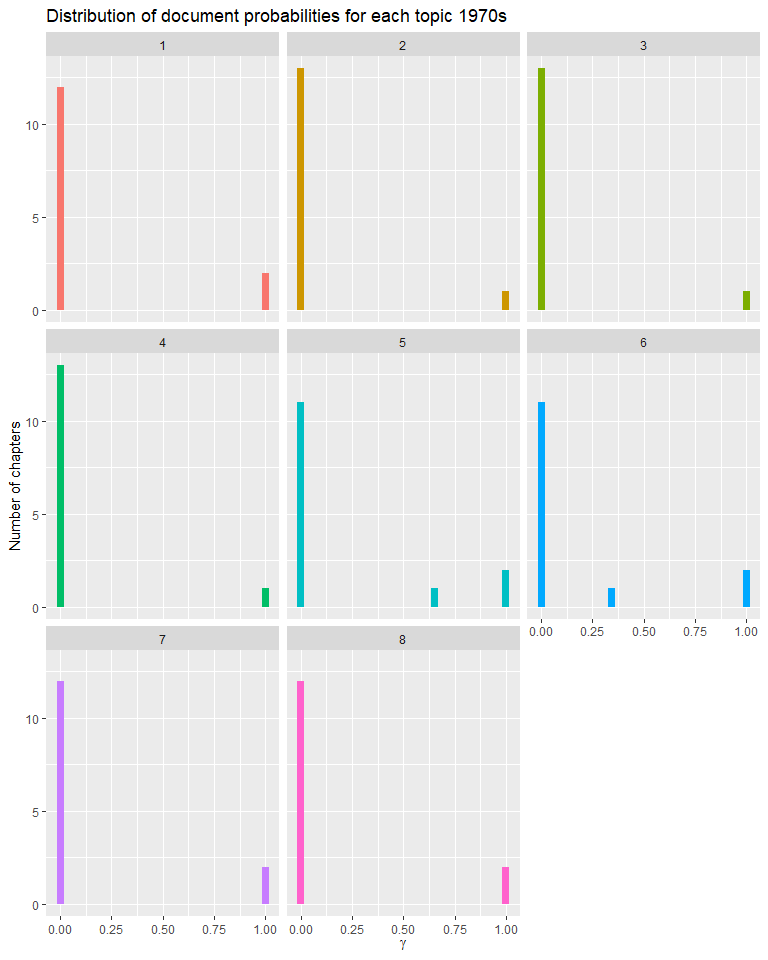


ggplot(td\_gamma3, aes(gamma)) +geom\_histogram() +  
 scale\_y\_log10() +  
 labs(title = "Distribution of probabilities for all topics 1990s",  
 y = "Number of documents", x = expression(gamma))



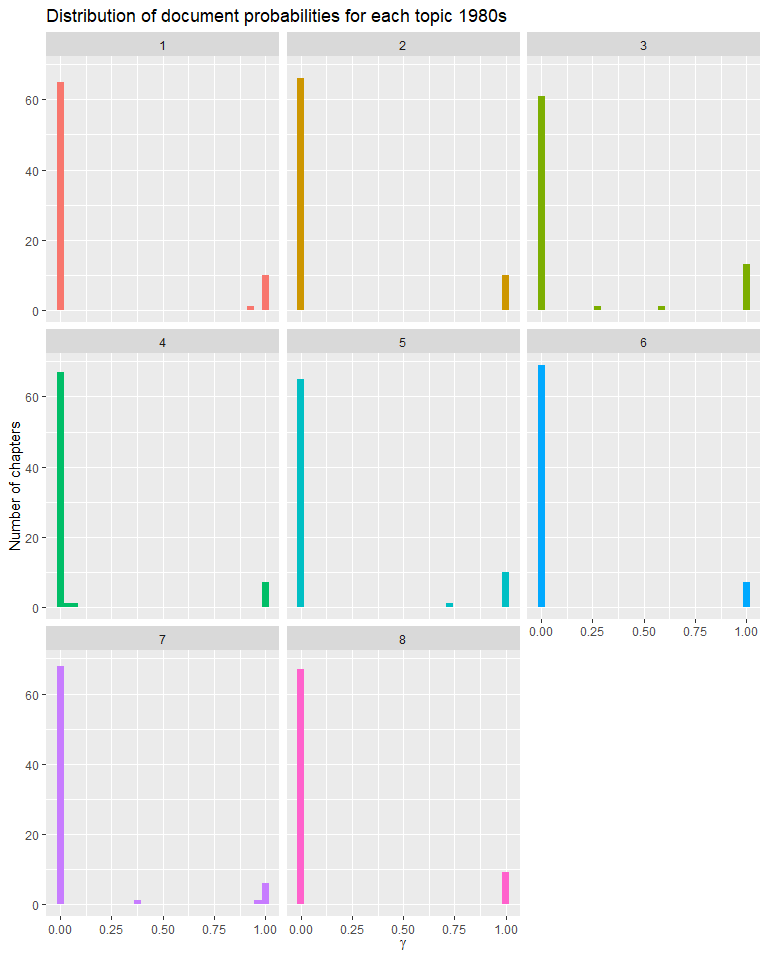
ggplot(td\_gamma1, aes(gamma, fill = as.factor(topic))) +  
 geom\_histogram(show.legend = FALSE) +  
 facet\_wrap(~ topic, ncol = 3) +  
 labs(title = "Distribution of document probabilities for each topic 1970s",  
 y = "Number of chapters", x = expression(gamma))

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



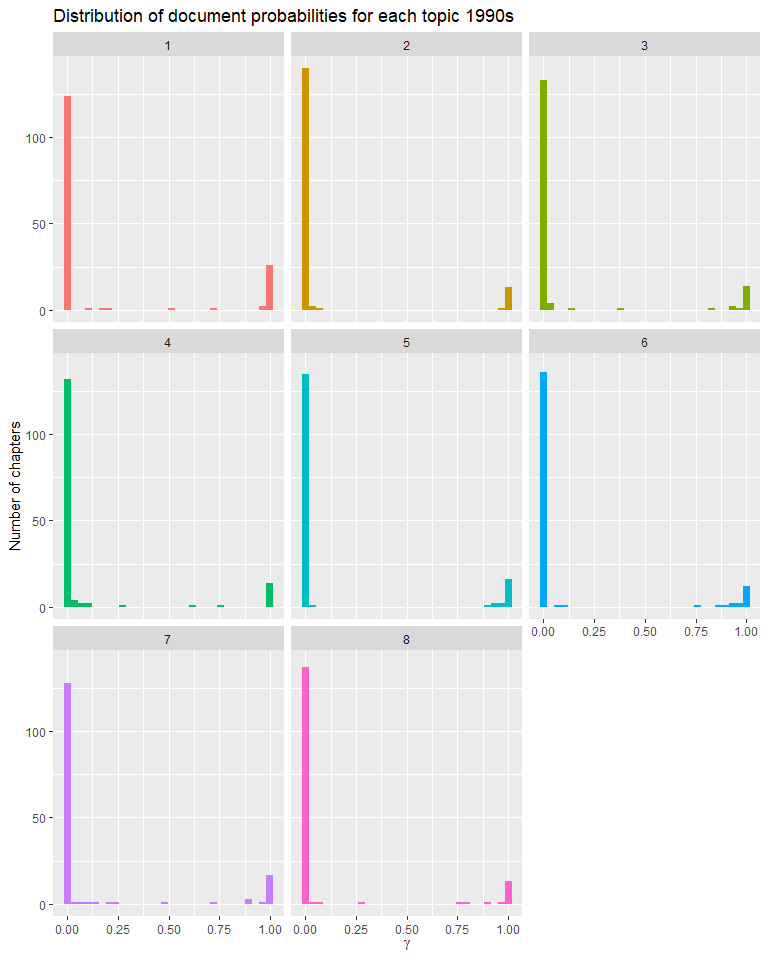
ggplot(td\_gamma2, aes(gamma, fill = as.factor(topic))) +  
 geom\_histogram(show.legend = FALSE) +  
 facet\_wrap(~ topic, ncol = 3) +  
 labs(title = "Distribution of document probabilities for each topic 1980s",  
 y = "Number of chapters", x = expression(gamma))

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

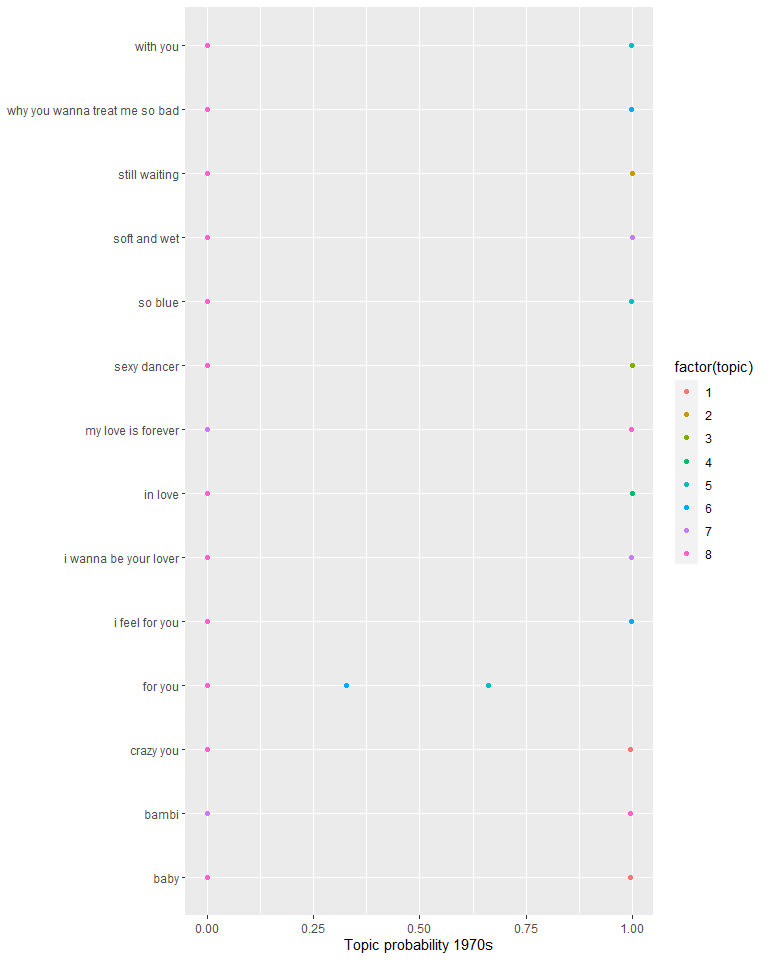


ggplot(td\_gamma3, aes(gamma, fill = as.factor(topic))) +  
 geom\_histogram(show.legend = FALSE) +  
 facet\_wrap(~ topic, ncol = 3) +  
 labs(title = "Distribution of document probabilities for each topic 1990s",  
 y = "Number of chapters", x = expression(gamma))

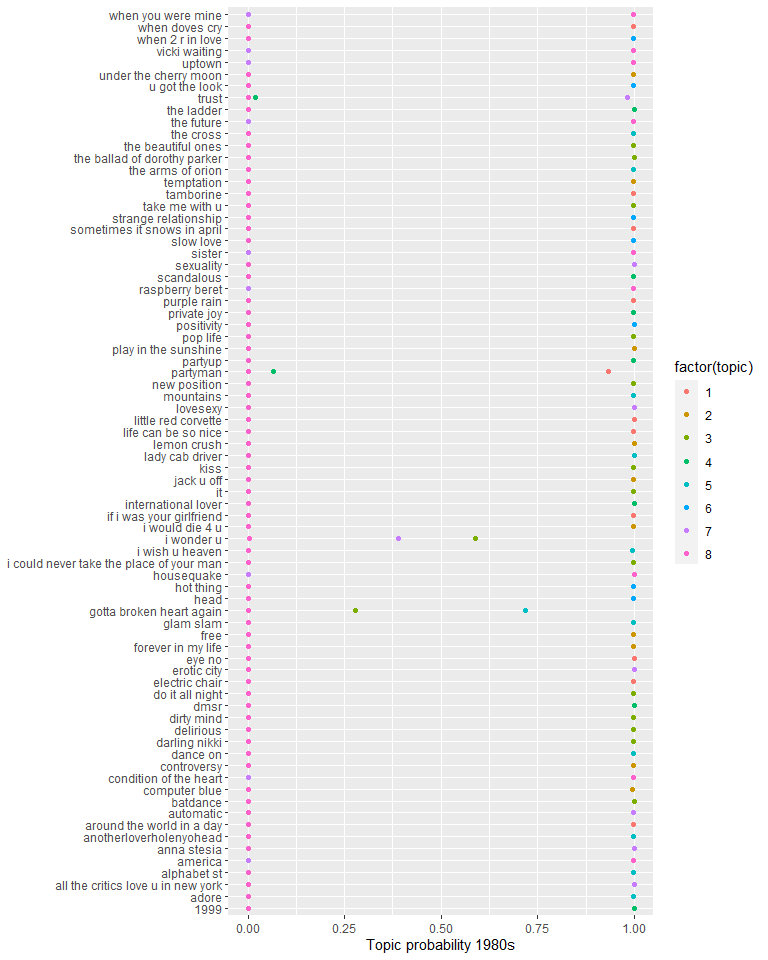
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



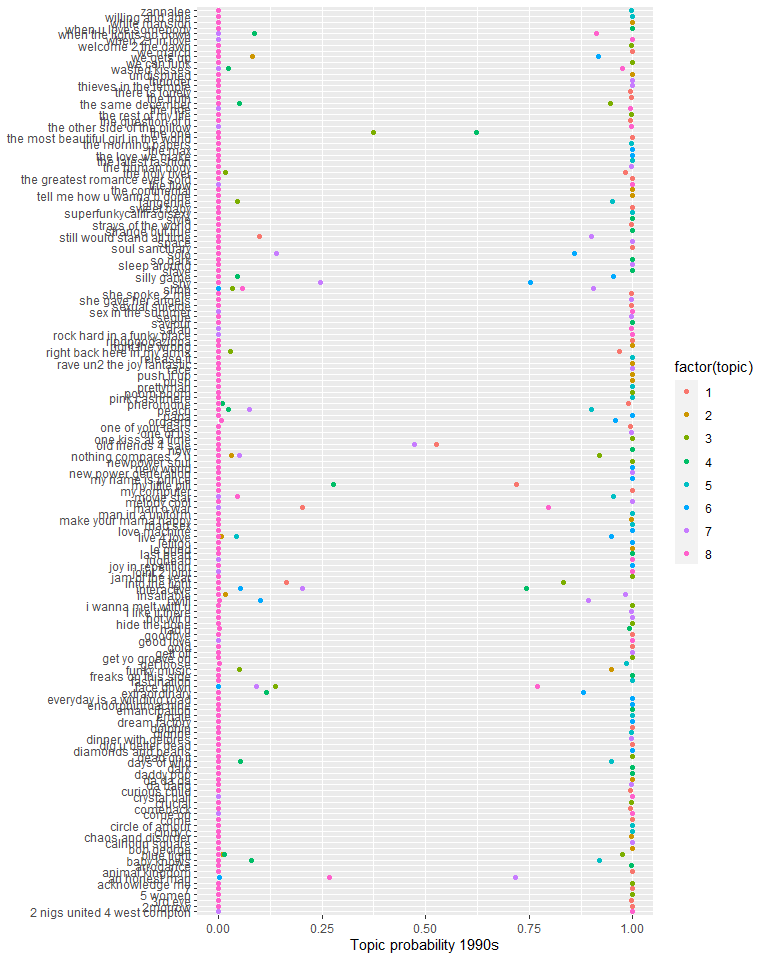
ggplot(td\_gamma1, aes(x=document, y=gamma)) +   
 geom\_point(aes(color=factor(topic))) +   
 labs(x=NULL, y="Topic probability 1970s") +  
 coord\_flip()



ggplot(td\_gamma2, aes(x=document, y=gamma)) +   
 geom\_point(aes(color=factor(topic))) +   
 labs(x=NULL, y="Topic probability 1980s") +  
 coord\_flip()



ggplot(td\_gamma3, aes(x=document, y=gamma)) +   
 geom\_point(aes(color=factor(topic))) +   
 labs(x=NULL, y="Topic probability 1990s") +  
 coord\_flip()



### Summary of outputs

Based on the different outputs we have created using beta and gamma matrices shows that our choice for a k value did not work equally well for all decades tested. In the 1970’s we can reasonably use the K value of 8. This selected value gave us high beta values as well as a gamma test that showed that topics were largely in a topic or not in a topic. In the 1980’s and 1990’s this breaks down. The beta values for both time periods were significantly lower thatn for the 1970’s and the gamma test showed a large variety of words that were not correlated with either being in a topic or not. That is to say that these words were present in many topics during this time period.