LDA

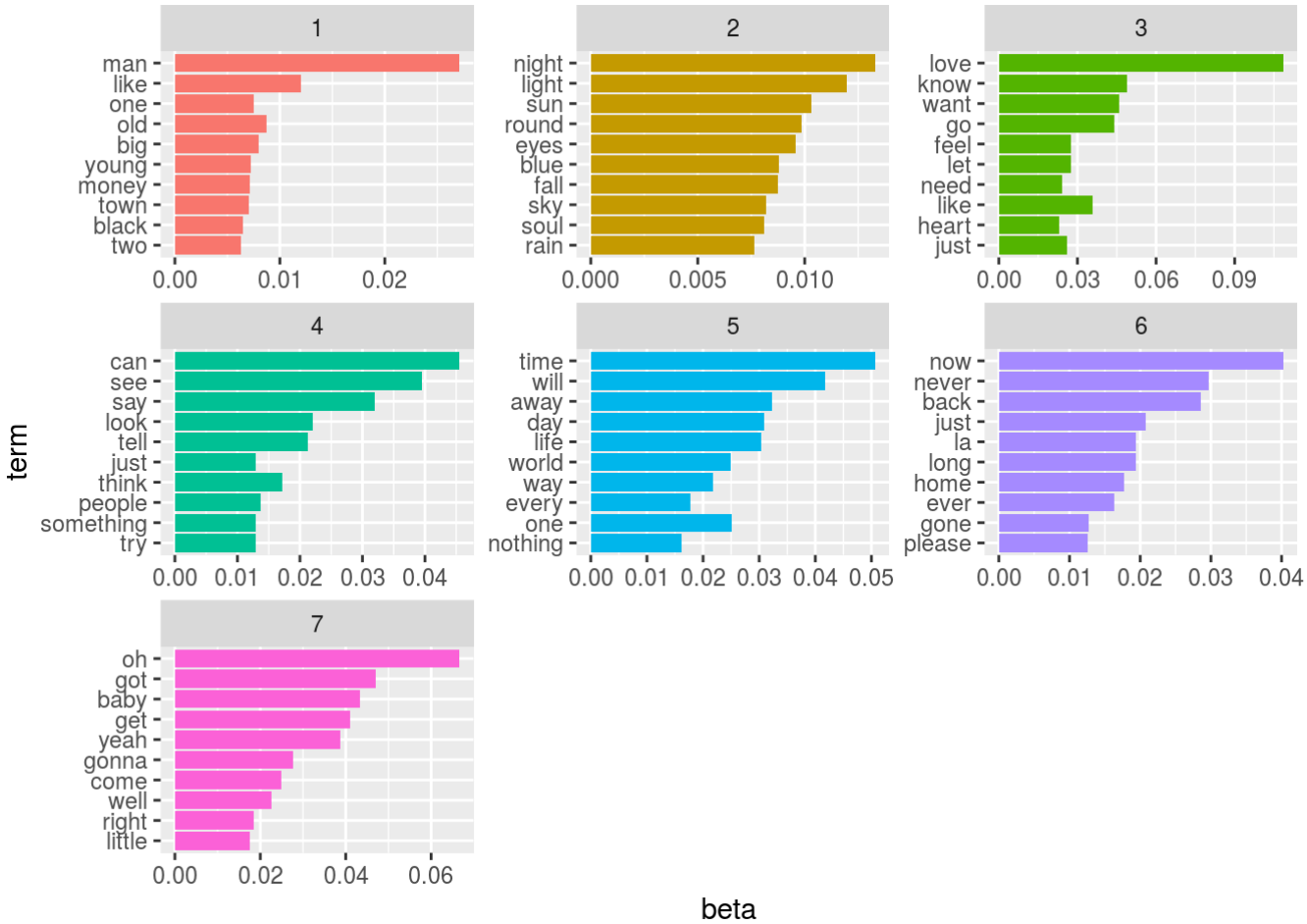
LDA

1. stopwords for Lyrics 2. threshold (# of min Songs) to delete part of artists 3. release date for each Song. Then For each Artist, show topic changes over time

```
# Setting seed
set.seed(826)

# Read Data
BI = read.csv('~/Documents/text_project/British_Invasion/Preprocessing/BI.csv',
              stringsAsFactors = FALSE)
ly <-dfm(BI$Lyrics, stem=F, removePunct = T, tolower=T, remove_Numbers = T,
          remove = c(stopwords(kind="english")))
```

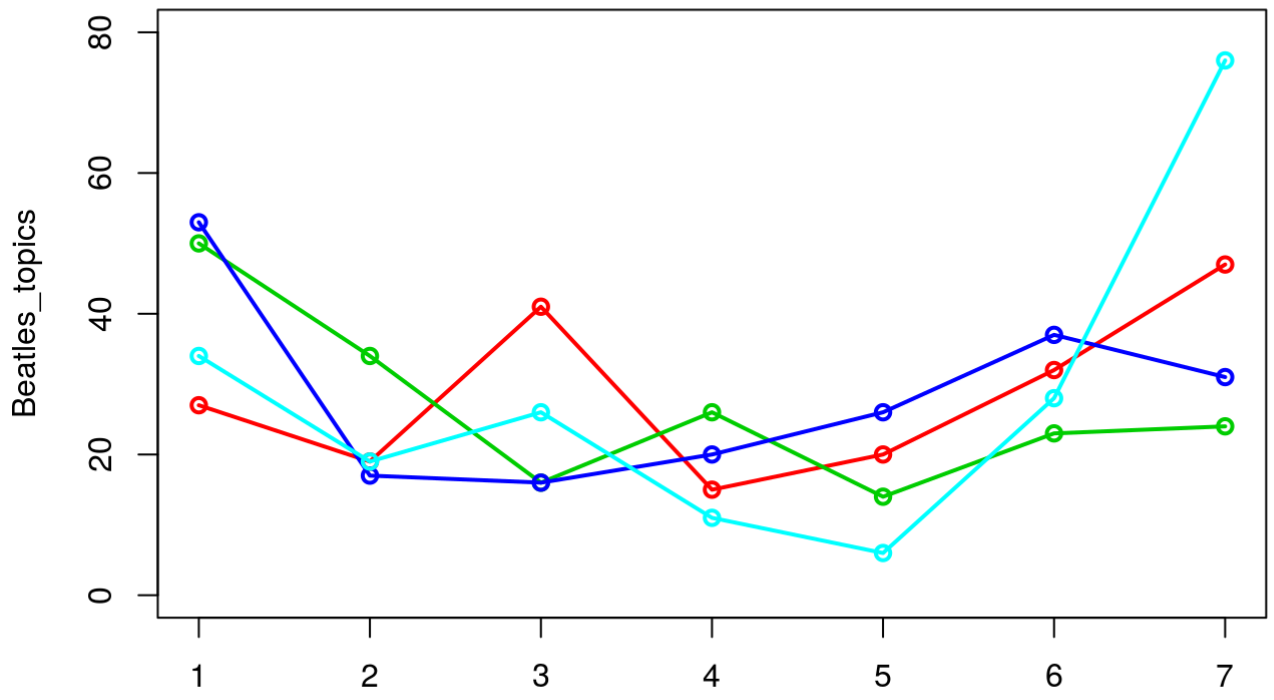
```
# Find K
# start.time <- Sys.time()
# ############################
# result <- FindTopicsNumber(
#   ly,
#   topics = seq(from = 2, to = 20, by = 1),
#   metrics = c("Griffiths2004", "CaoJuan2009", "Arun2010", "Deveaud2014"),
#   method = "Gibbs",
#   control = list(seed = 826),
#   mc.cores = 3L,
#   verbose = TRUE)
# ############################
# end.time <- Sys.time()
# time.taken <- end.time - start.time
# time.taken
#
# FindTopicsNumber_plot(result)
K=7
```

```
BI_mod<-LDA(ly, k = 7, method = "Gibbs",  control = list(seed = 826))

# Quickly extracts the word weights and transforms them into a data frame
BI_topics <- tidy(BI_mod, matrix = "beta")

# Generates a df of top terms
BI_top_terms <- BI_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

BI_top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```

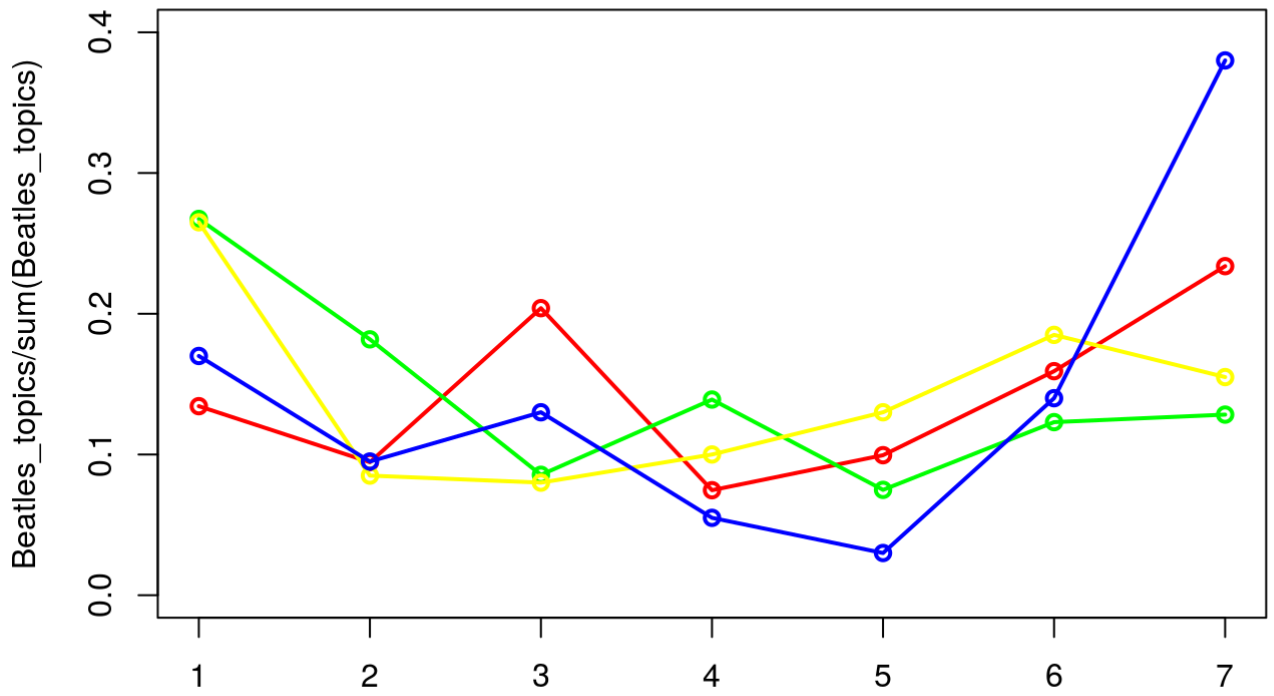


```
## 4 Visualizing topic trends over time
# Store the results of the distribution of topics over documents
doc_topics<-BI_mod@gamma
# Store the results of words over topics
words_topics<-BI_mod@beta
# Arrange topics
K=7
max<-apply(doc_topics, 1, which.max)
which.max2<-function(x){
  which(x == sort(x,partial=(K-1))[K-1])
}
max2<- apply(doc_topics, 1, which.max2)
max2<-sapply(max2, max)

index<-seq(1:nrow(doc_topics))
top2<-data.frame(max = max, max2 = max2, index = index)#date = ymd(blm_tweets_sum$date
2)
top2 = cbind.data.frame(top2, BI)
```

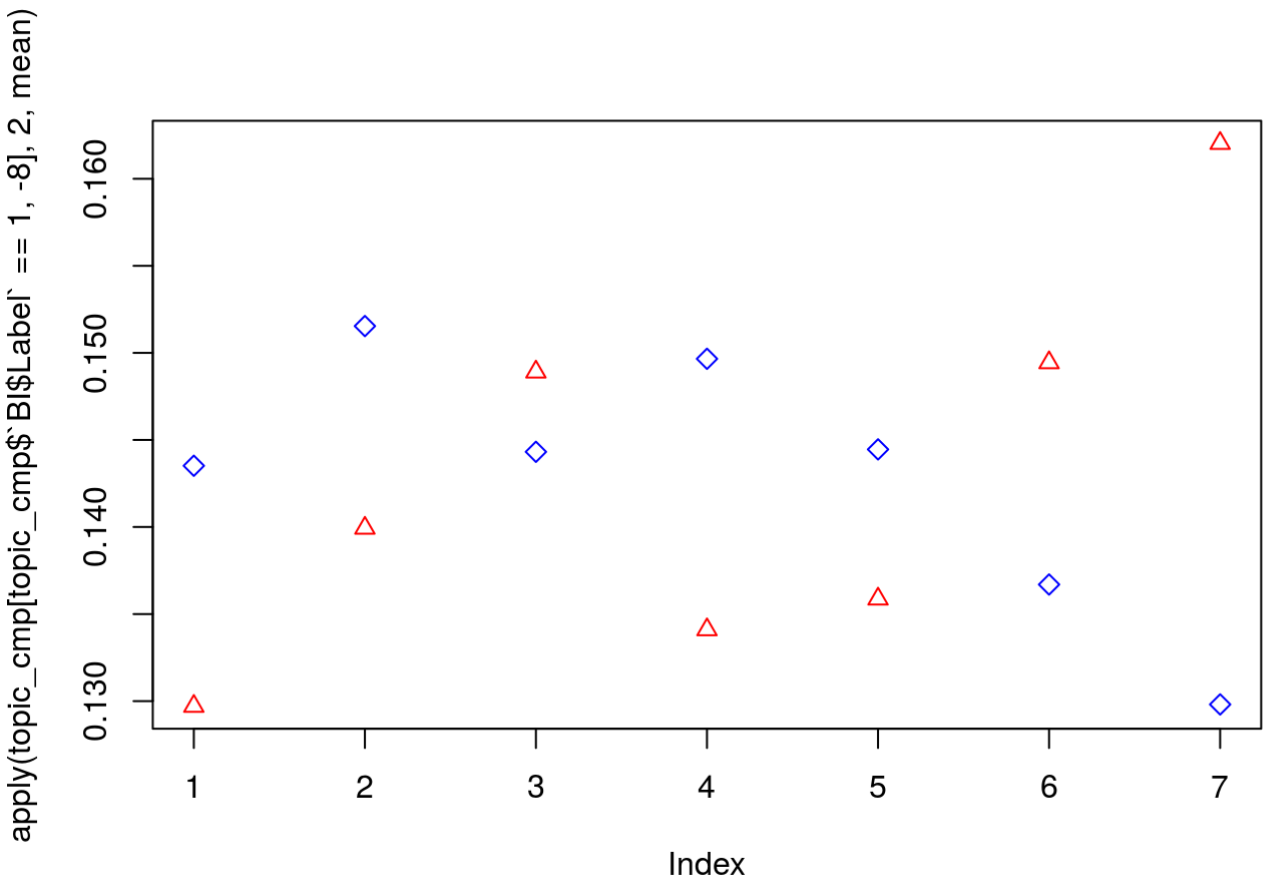
```
## Topic Compare: Big Four (British Invasion 1)
Beatles_topics = table(top2[top2$Artist == 'The Beatles',]$max)
Who_topics = table(top2[top2$Artist == 'The Who',]$max)
Kinks_topics = table(top2[top2$Artist == 'The Kinks',]$max)
RollingStones_topics = table(top2[top2$Artist == 'The Rolling Stones',]$max)

plot(Beatles_topics, type='o', col=2, ylim = c(0,80))
points(Who_topics, col=3, type = 'o')
points(Kinks_topics, col=4, type = 'o')
points(RollingStones_topics, col=5, type = 'o')
```



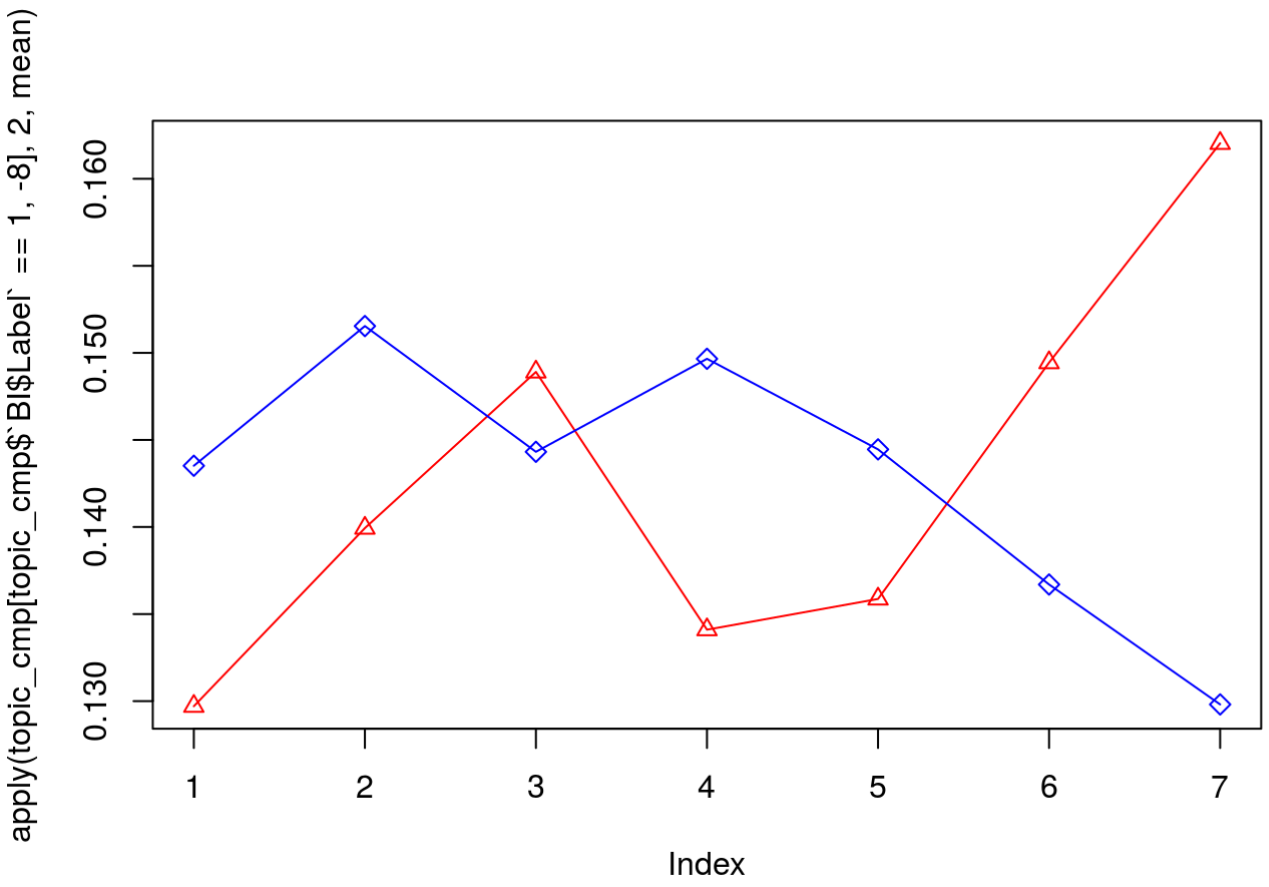

```
plot(Beatles_topics/sum(Beatles_topics), type='o', col='red', ylim = c(0,0.4))
points(Who_topics/sum(Who_topics), col='green', type = 'o')
points(Kinks_topics/sum(Kinks_topics), col='yellow', type = 'o')
points(RollingStones_topics/sum(RollingStones_topics), col='blue', type = 'o')
```



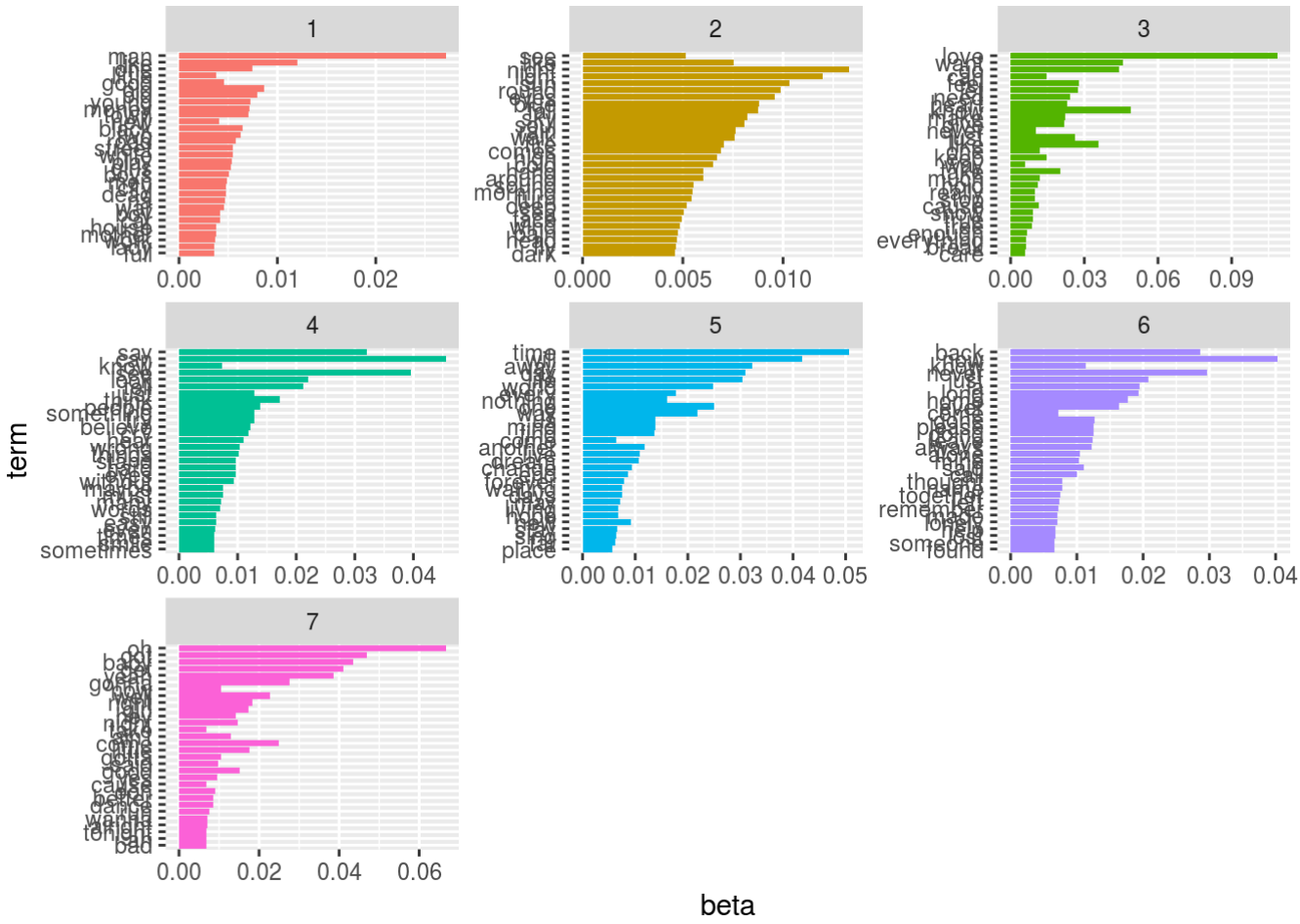

```
##BI1 V.S. BI2 over 7 Topics
topic_cmp = cbind.data.frame(doc_topics, BI$Label)
plot(apply(topic_cmp[topic_cmp$`BI$Label`==1,-8], 2, mean), col=2, pch=2)
points(apply(topic_cmp[topic_cmp$`BI$Label`==2,-8], 2, mean), col=4, pch=5)
```




```
plot(apply(topic_cmp[topic_cmp$`BI$Label`==1,-8], 2, mean), col=2, pch=2, type = 'o')
lines(apply(topic_cmp[topic_cmp$`BI$Label`==2,-8], 2, mean), col=4, pch=5)
points(apply(topic_cmp[topic_cmp$`BI$Label`==2,-8], 2, mean), col=4, pch=5)
```




```
BI_top_terms <- BI_topics %>%
  group_by(topic) %>%
  top_n(30, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
BI_top_terms %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()
```


1
2
3
4
5
6
7
term
beta
0.00
0.01
0.02
0.000
0.005
0.010
0.03
0.06
0.09
0.04
0.05