R Notebook

Please run the Text_Processing.Rmd first (I created a different data set), Or dowload the “word_lyrics.RData” file from output. Thanks!

optimism or pessimism

I read lyrics a lot from primary school to university. In my memory, most famous chinese lyrics are about sad stories. Most Chinese poets are very pessimism.

How about american poet? I’m about to find out!

```
load('../output/word_lyrics.RData')
```

```
library(tidytext)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(stringr)
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(reshape2)
library(ggplot2)
```

```
word_lyrics %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 100))
```


```
bing_word_counts <- word_lyrics %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()
```

```
## Joining, by = "word"
```

```
head(bing_word_counts,10)
```

word	sentiment	n
<chr>	<chr>	<int>

love	positive	194052
fall	negative	32478
die	negative	28362
lie	negative	26504
cry	negative	25297
break	negative	21737
hard	negative	21244
wrong	negative	19757
lost	negative	19752
burn	negative	19383

1-10 of 10 rows

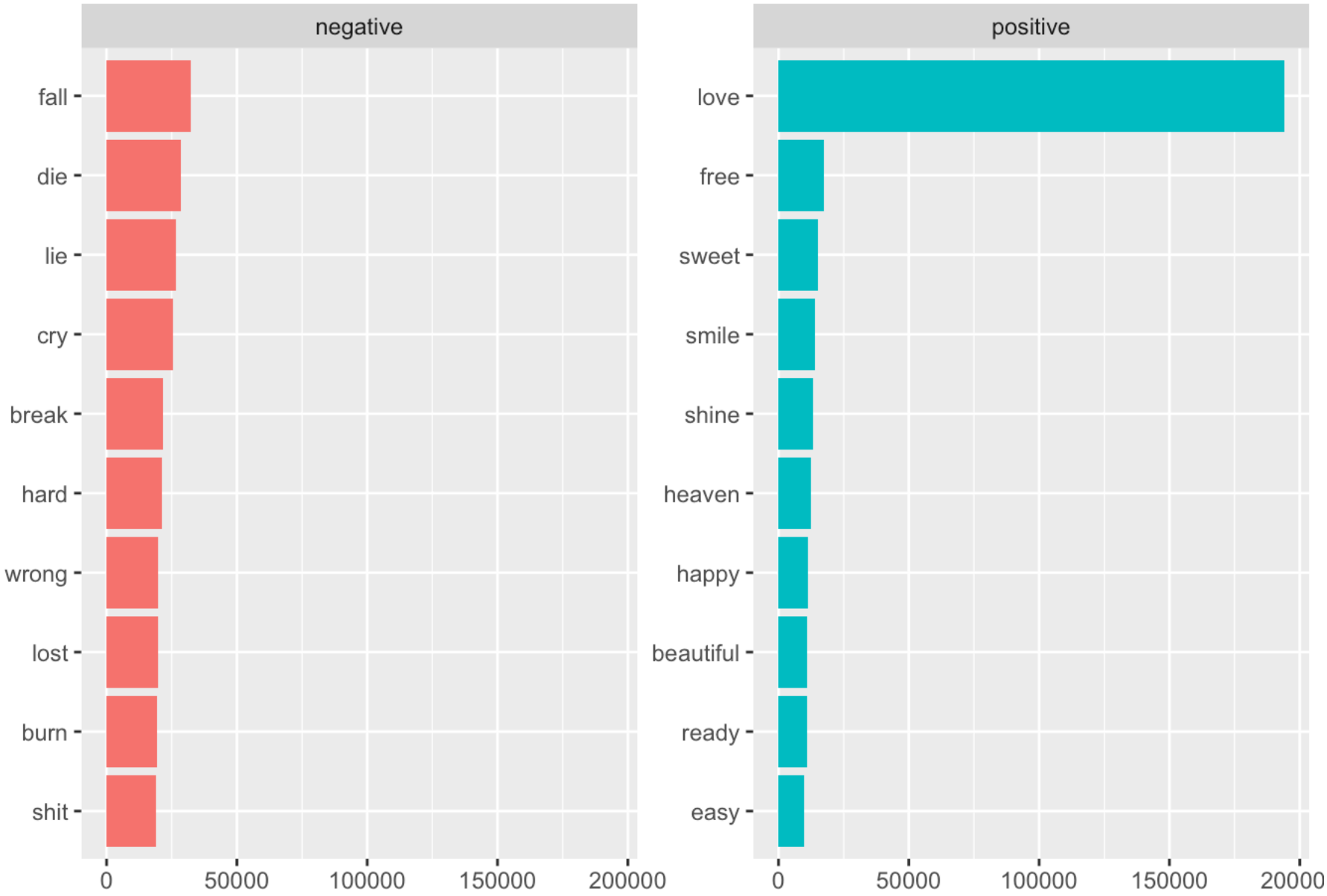
```
word_lyrics %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("blue", "red"),
                   max.words = 100)
```

```
## Joining, by = "word"
```

The number of “love” used in the lyrics is quite surprising. Lyrics are full of “love”. Does that mean american poets are truly optimism?

```
bing_word_counts %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(y = "Contribution to sentiment",
       x = NULL) +
  coord_flip()
```

```
## Selecting by n
```



Two graphy above shows among the top10 positve words and negative words, the number american poets use is amost the same. Let’s look deeper!

```
bingnegative <- get_sentiments("bing") %>%
  filter(sentiment == "negative")

wordcounts <- word_lyrics %>%
  group_by(year) %>%
  summarize(words = n())

negat <- word_lyrics %>%
  semi_join(bingnegative) %>%
  group_by(year) %>%
  summarize(negativewords = n()) %>%
  left_join(wordcounts, by = c("year")) %>%
  mutate(ratio = negativewords/words) %>%
  filter(year != 0) %>%
  ungroup()
```

```
## Joining, by = "word"
```

```
bingpositive <- get_sentiments("bing") %>%
  filter(sentiment == "positive")

posit<- word_lyrics %>%
  semi_join(bingpositive) %>%
  group_by(year) %>%
  summarize(positivewords = n()) %>%
  left_join(wordcounts, by = c("year")) %>%
  mutate(ratio = positivewords/words) %>%
  filter(year != 0) %>%
  ungroup()
```

```
## Joining, by = "word"
```

```
wordstable <- merge(negat,posit,by="year",all=T)
l <- c(1,2)
wordstable <- wordstable[-l,]
wordstable
```

	year	negativewords	words.x	ratio.x	positivewords	words.y	ratio.y
	<dbl>	<int>	<int>	<dbl>	<int>	<int>	<dbl>
3	1968	10	40	0.25000000	1	40	0.02500000
4	1970	713	6316	0.11288790	647	6316	0.10243825
5	1971	792	7538	0.10506766	639	7538	0.08477050
6	1972	795	7450	0.10671141	721	7450	0.09677852
7	1973	1175	10853	0.10826500	1022	10853	0.09416751
8	1974	1069	8835	0.12099604	727	8835	0.08228636
9	1975	512	5029	0.10180950	578	5029	0.11493339
10	1976	275	2982	0.09221999	308	2982	0.10328638
11	1977	1586	14694	0.10793521	1344	14694	0.09146590
12	1978	934	8936	0.10452104	862	8936	0.09646374


```
wordstable$posnogratio<-wordstable$positivewords/wordstable$negativewords
poyear <- wordstable$year[which(wordstable$posnograti>1)]
length(poyear)
```

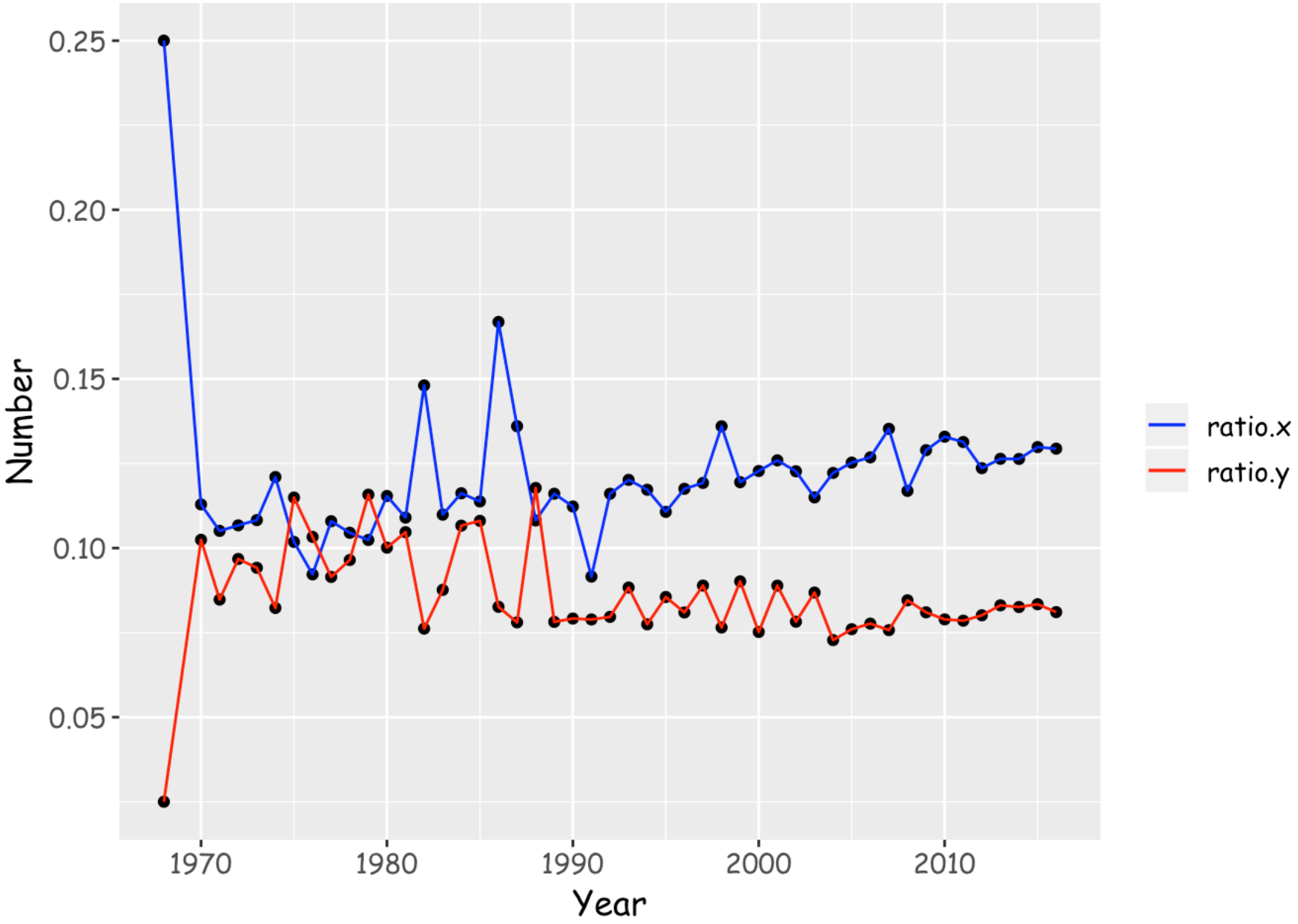
```
## [1] 4
```

```
neyear <- wordstable$year[which(wordstable$posnograti<1)]
length(neyear)
```

```
## [1] 44
```

Op! Surprise! Among 48 years, there is only 4 years that positive words ratio is greater than the negative words ratio!

```
ggplot(wordstable, aes(x=year)) +
            geom_point(aes(y=ratio.x), ) +
            geom_line(aes(y=ratio.x, , color="ratio.x")) +
            geom_point(aes(y=ratio.y)) +
            geom_line(aes(y=ratio.y, color="ratio.y"))+
  scale_colour_manual("",values = c("ratio.x" = "blue","ratio.y" = "red"))+
  xlab("Year")+ylab("Number")+
  theme(text=element_text(size=13, family="Comic Sans MS"))
```



The ratio.x means the negative words/total words; the ratio.y means the positive words/total words. In the graph, it seems that there is a trend that more and more negative words are used in the lyrics. The positive words ratio is kind of stable. Let's do a linear regression test to see whether there is a linear relationship between negative words and years.

```
summary(lm(wordstable$ratio.x~wordstable$year))
```

```
##
## Call:
## lm(formula = wordstable$ratio.x ~ wordstable$year)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.030260 -0.011304 -0.003936  0.003326  0.130204
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -5.407e-02  4.829e-01  -0.112    0.911
## wordstable$year  8.835e-05  2.423e-04   0.365    0.717
##
## Residual standard error: 0.02332 on 46 degrees of freedom
## Multiple R-squared:  0.002881,   Adjusted R-squared:  -0.0188
## F-statistic: 0.1329 on 1 and 46 DF,  p-value: 0.7171
```

P value is quite large, which shows there is no significant evidence that negative words ratio has a positive linear relationship with years.

If the world without "LOVE"

"love" seems like an outliear. The number of "LOVE" used in lyrics is 6 times bigger than the

second-used word. What would happen if there is no “LOVE”

```
delove <- which(word_lyrics$word=="love")
newword_lyrics <- word_lyrics[-delove,]
```

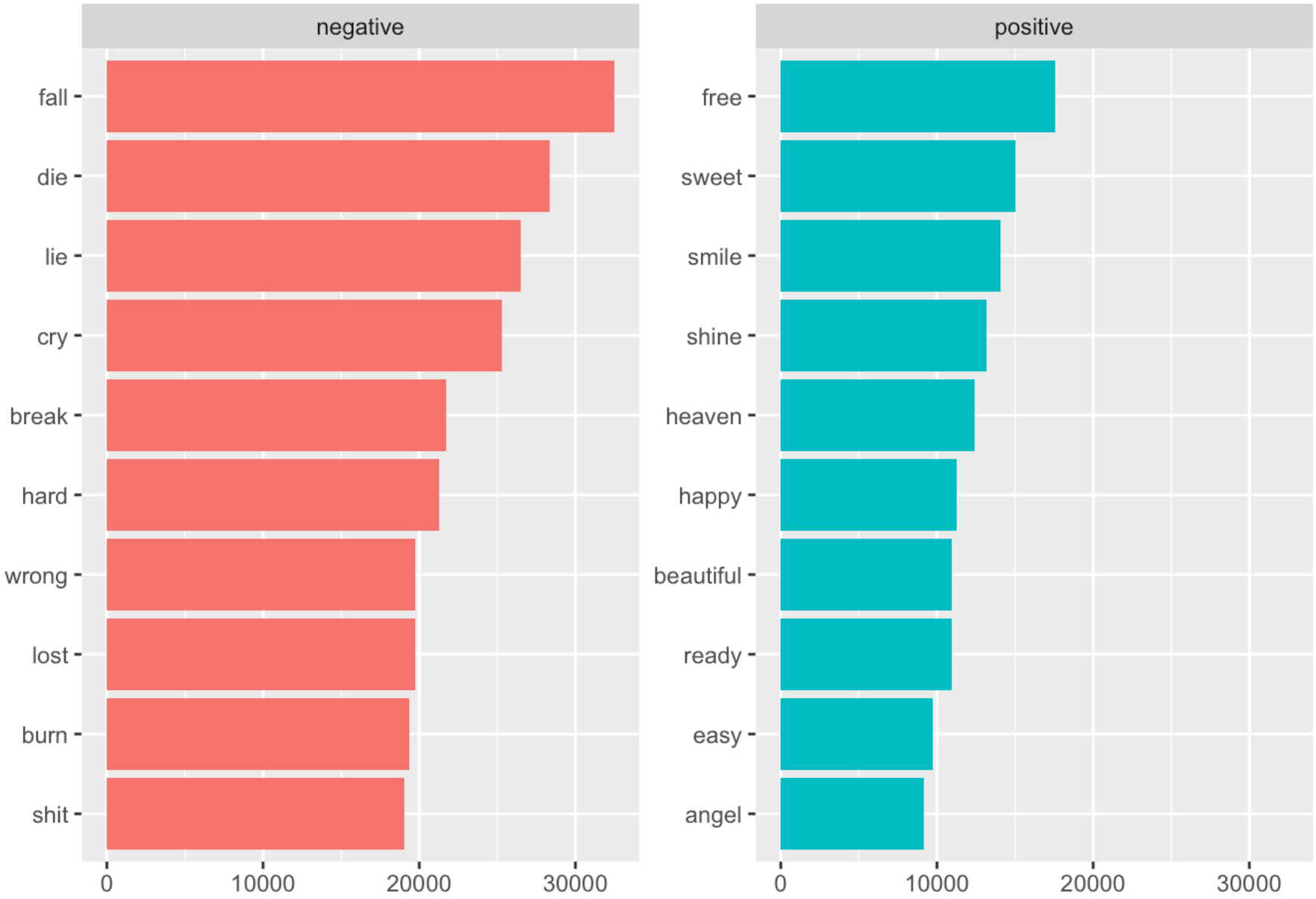
```
emotion_word_counts <- newword_lyrics %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()
```

```
## Joining, by = "word"
```

without “LOVE” the image is more clear that there are more negative words used in the lyrics.

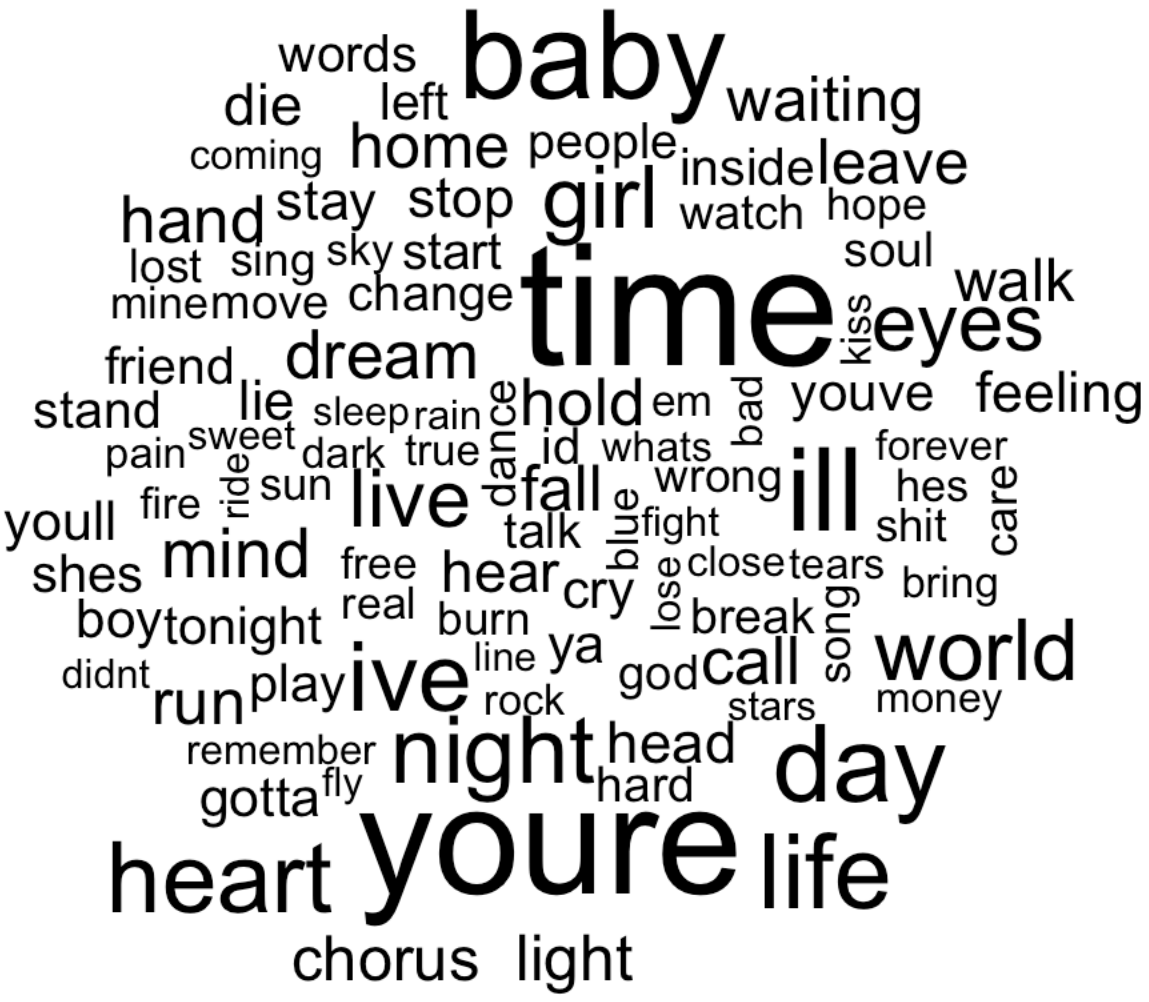
```
emotion_word_counts %>%
  group_by(sentiment) %>%
  top_n(10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(y = "Contribution to sentiment",
       x = NULL) +
  coord_flip()
```

```
## Selecting by n
```




```
library(wordcloud)

newword_lyrics %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 100))
```




```
bingnegative <- get_sentiments("bing") %>%
  filter(sentiment == "negative")

wordcounts <- newword_lyrics %>%
  group_by(year) %>%
  summarize(words = n())

negat <- newword_lyrics %>%
  semi_join(bingnegative) %>%
  group_by(year) %>%
  summarize(negativewords = n()) %>%
  left_join(wordcounts, by = c("year")) %>%
  mutate(ratio = negativewords/words) %>%
  filter(year != 0) %>%
  ungroup()
```

```
## Joining, by = "word"
```

```
bingpositive <- get_sentiments("bing") %>%
  filter(sentiment == "positive")

posit<- newword_lyrics %>%
  semi_join(bingpositive) %>%
  group_by(year) %>%
  summarize(positivewords = n()) %>%
  left_join(wordcounts, by = c("year")) %>%
  mutate(ratio = positivewords/words) %>%
  filter(year != 0) %>%
  ungroup()
```

```
## Joining, by = "word"
```

```
wordstable <- merge(negat,posit,by="year",all=T)
l <- c(1,2)
wordstable <- wordstable[-l,]
```

```
wordstable$posnogratio<-wordstable$positivewords/wordstable$negativewords
poyear <- wordstable$year[which(wordstable$posnograti>1)]
length(poyear)
```

```
## [1] 0
```

```
neg <- wordstable$year[which(wordstable$posnograti<1)]
length(neg)
```

```
## [1] 47
```

No doubt, without “LOVE”, each year, nagative words ratio is greater than positive words ratio.

Conclusions

American poets use “LOVE” a lot, the frequency is 6 times bigger than the second-used word. It gives us a misconception

that most american poets are optimism, however, when we consider the whole positive words ratio VS nagative words ratio, most american poets are pessimism!