Untitled

vidhu

February 12, 2019

Text Analysis Load the libraries

```
library(tidytext)
library(wordcloud2)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Read the data

```
train<- read.csv("drugsComTrain_raw.csv", stringsAsFactors = FALSE)
```

```
head(train$review)
```

```
## [1] "\"It has no side effect, I take it in combination of Bystolic 5 Mg and Fish Oil\""
## [2] "\"My son is halfway through his fourth week of Intuniv. We became concerned when he bega
n this last week, when he started taking the highest dose he will be on. For two days, he could
hardly get out of bed, was very cranky, and slept for nearly 8 hours on a drive home from school
vacation (very unusual for him.) I called his doctor on Monday morning and she said to stick it
out a few days. See how he did at school, and with getting up in the morning. The last two days
have been problem free. He is MUCH more agreeable than ever. He is less emotional (a good thin
g), less cranky. He is remembering all the things he should. Overall his behavior is better. \nW
e have tried many different medications and so far this is the most effective.\""
## [3] "\"I used to take another oral contraceptive, which had 21 pill cycle, and was very happy
- very light periods, max 5 days, no other side effects. But it contained hormone gestodene, whi
ch is not available in US, so I switched to Lybrel, because the ingredients are similar. When my
other pills ended, I started Lybrel immediately, on my first day of period, as the instructions
said. And the period lasted for two weeks. When taking the second pack- same two weeks. And now,
with third pack things got even worse- my third period lasted for two weeks and now it&#039;s th
e end of the third week- I still have daily brown discharge.\nThe positive side is that I didn&#
039;t have any other side effects. The idea of being period free was so tempting... Alas.\""
## [4] "\"This is my first time using any form of birth control. I&#039;m glad I went with the p
atch, I have been on it for 8 months. At first It decreased my libido but that subsided. The onl
y downside is that it made my periods longer (5-6 days to be exact) I used to only have periods
for 3-4 days max also made my cramps intense for the first two days of my period, I never had cr
amps before using birth control. Other than that in happy with the patch\""
## [5] "\"Suboxone has completely turned my life around.  I feel healthier, I&#039;m excelling a
t my job and I always have money in my pocket and my savings account.  I had none of those befor
e Suboxone and spent years abusing oxycontin.  My paycheck was already spent by the time I got i
t and I started resorting to scheming and stealing to fund my addiction.  All that is history.
If you&#039;re ready to stop, there&#039;s a good chance that suboxone will put you on the path
of great life again.  I have found the side-effects to be minimal compared to oxycontin.  I&#03
9;m actually sleeping better.   Slight constipation is about it for me.  It truly is amazing. Th
e cost pales in comparison to what I spent on oxycontin.\""
## [6] "\"2nd day on 5mg started to work with rock hard erections however experianced headache,
lower bowel preassure. 3rd day erections would wake me up &amp; hurt! Leg/ankles aches   severe
lower bowel preassure like you need to go #2 but can&#039;t! Enjoyed the initial rockhard erecti
ons but not at these side effects or $230 for months supply! I&#039;m 50 &amp; work out 3Xs a we
ek. Not worth side effects!\""
```

Make a copy of train data

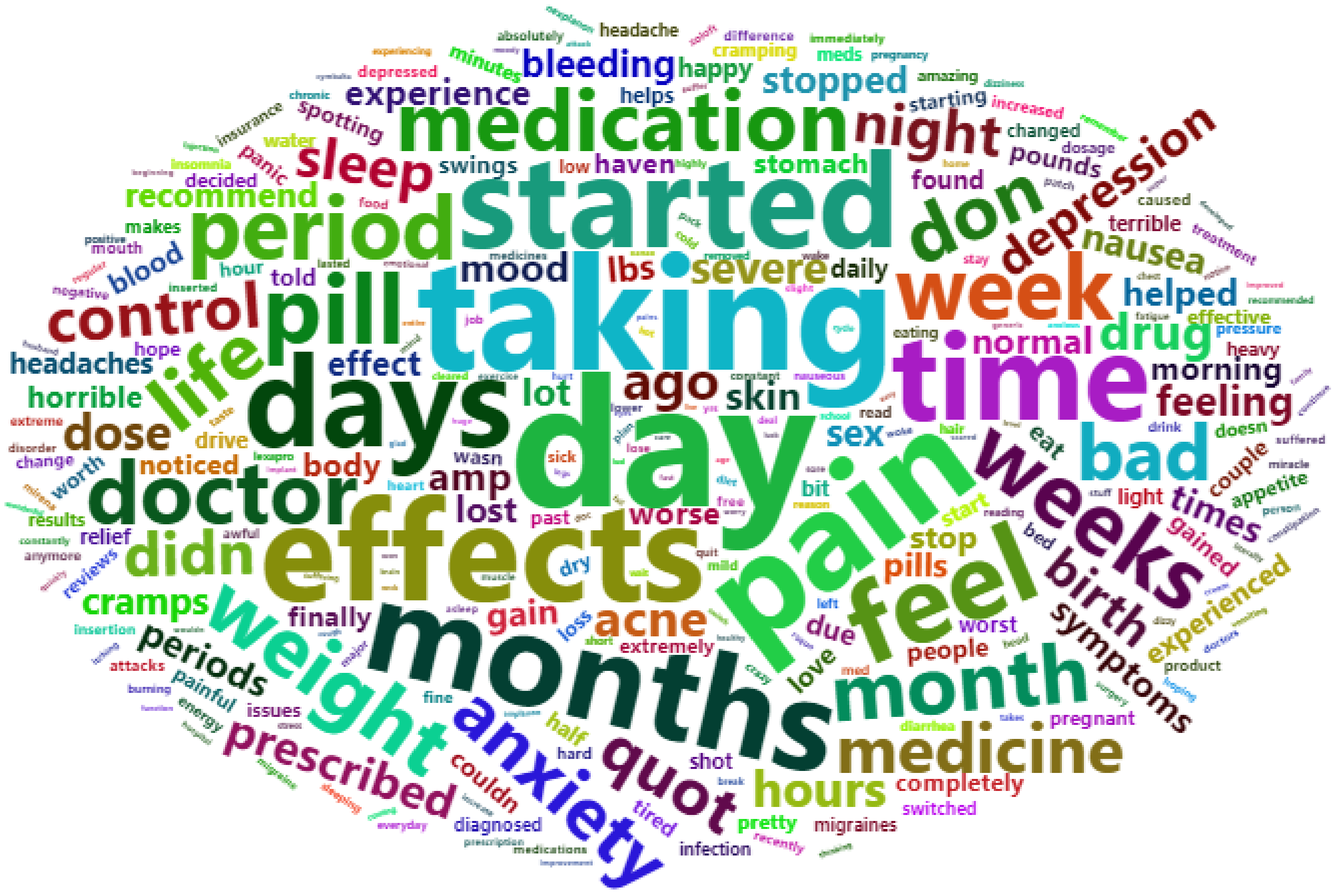
```
cleantrain = train
# function to replace the absurd and missing values to NULL.
Missing = function(x) gsub("^[0-9]+.*", NA, x)
cleantrain$condition = sapply(cleantrain$condition,Missing)
cleantrain$condition = replace(cleantrain$condition, cleantrain$condition == "", NA)
```

```
# function to expand contractions in an English-language source
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)
}
# function to remove special characters
removeSpecialChars <- function(x) gsub("[^a-zA-Z]", " ", x)

# Fix contractions and remove special characters from data
cleantrain$review <- sapply(cleantrain$review, fix.contractions)
cleantrain$review <- sapply(cleantrain$review, removeSpecialChars)
```

```
train_tidy <- cleantrain %>%
  unnest_tokens(word, review)%>%
  filter(!nchar(word) < 3) %>%
  anti_join(stop_words)
```

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

```
t <- train_tidy %>%
  count(word, sort = TRUE)

wordcloud2(t[1:300, ], size = .5)
```



```
train_bing <- train_tidy %>%
  inner_join(get_sentiments("bing"))
```

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

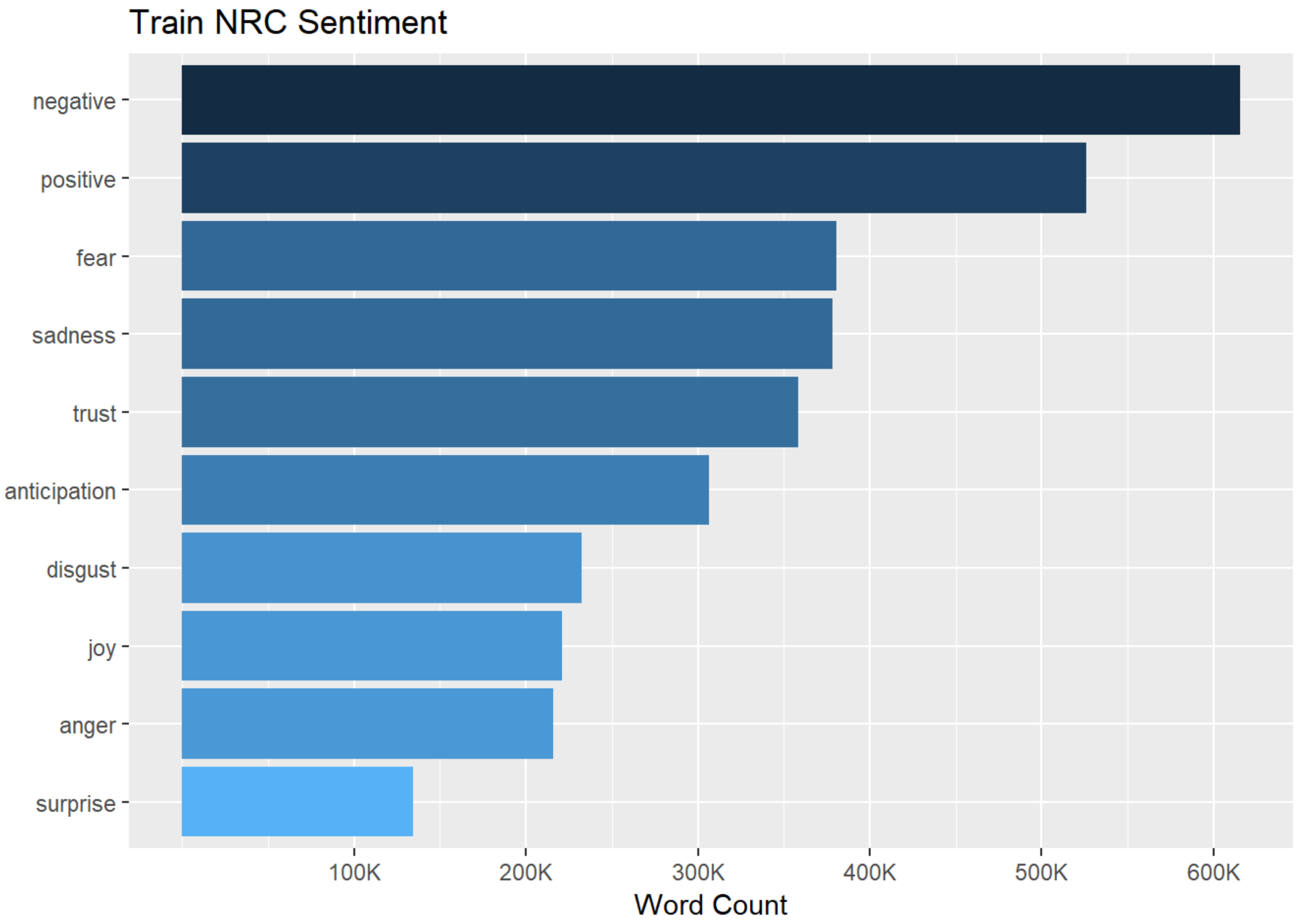
```
train_nrc <- train_tidy %>%
  inner_join(get_sentiments("nrc"))
```

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

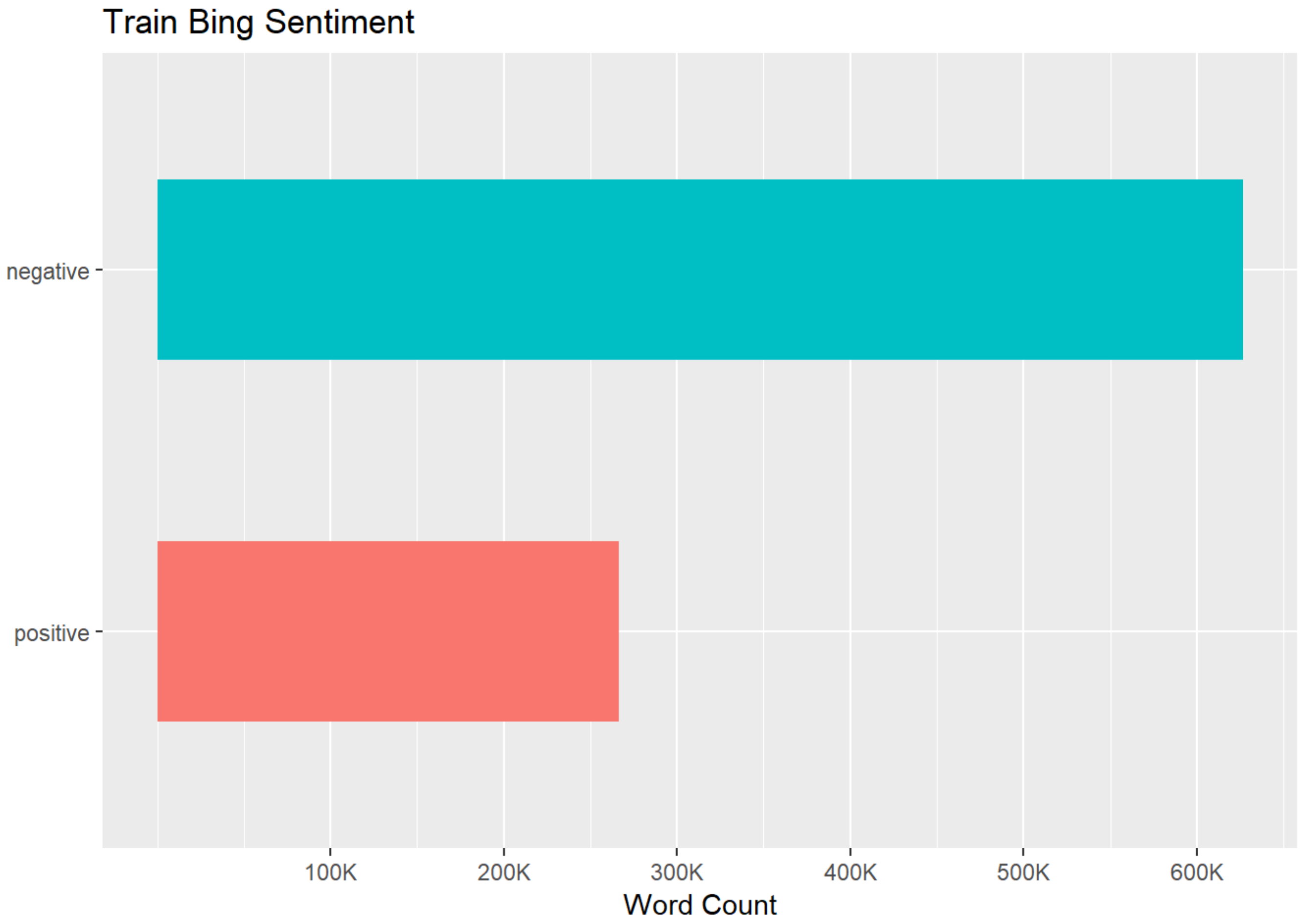
```
train_nrc_sub <- train_tidy %>%
  inner_join(get_sentiments("nrc")) %>%
  filter(!sentiment %in% c("positive", "negative"))
```

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

```
train_nrc %>%
  group_by(sentiment) %>%
  summarise(word_count = n()) %>%
  ungroup() %>%
  mutate(sentiment = reorder(sentiment, word_count)) %>%
  #Use `fill = -word_count` to make the larger bars darker
  ggplot(aes(sentiment, word_count, fill = -word_count)) +
  geom_col() +
  guides(fill = FALSE) + #Turn off the legend
    labs(x = NULL, y = "Word Count") +
  scale_y_continuous(breaks = seq(100000, 700000,100000), labels= c("100K","200K","300K", "400K"
, "500K", "600K", "700K")) +
  ggtitle("Train NRC Sentiment") +
  coord_flip()
```



```
train_bing %>%
  group_by(sentiment) %>%
  summarise(word_count = n()) %>%
  ungroup() %>%
  mutate(sentiment = reorder(sentiment, word_count)) %>%
  ggplot(aes(sentiment, word_count, fill = sentiment)) +
  geom_col(width=0.5) +
  guides(fill = FALSE) +
  labs(x = NULL, y = "Word Count") +
  scale_y_continuous(breaks = seq(100000, 700000,100000), labels= c("100K","200K","300K", "400K"
, "500K", "600K", "700K")) +
  ggtitle("Train Bing Sentiment") +
  coord_flip()
```



sentiments according to drug and condition

```
train_bing %>% na.omit %>%
  group_by(condition, drugName,sentiment) %>%
  summarise(word_count = n()) %>%
  ungroup() %>%
  group_by(condition)
```

```
## # A tibble: 13,671 x 4
## # Groups:   condition [788]
##    condition                drugName       sentiment word_count
##    <chr>                    <chr>          <chr>          <int>
##  1 Abdominal Distension     Bethanechol    negative           3
##  2 Abdominal Distension     Bethanechol    positive           1
##  3 Abdominal Distension     Urecholine     negative           3
##  4 Abdominal Distension     Urecholine     positive           1
##  5 Abnormal Uterine Bleeding Alesse        negative          25
##  6 Abnormal Uterine Bleeding Alesse        positive          13
##  7 Abnormal Uterine Bleeding Alyacen 1 / 35 negative          4
##  8 Abnormal Uterine Bleeding Alyacen 1 / 35 positive          1
##  9 Abnormal Uterine Bleeding Amethia Lo     negative          4
## 10 Abnormal Uterine Bleeding Amethia Lo     positive          2
## # ... with 13,661 more rows
```