Assignment_4_task2

Guangze Yu

12/8/2021

Task Two: bag of words analysis

The book that I chose is Women of Achievement by Benjamin Griffith Brawley.

The first step for analysis is to have a tidy text. Three original lexicons that include in "Text Mining with R" are AFINN, bing and nrc. The extra lexicon is loughran.

We clean the text in a tidy format with one word per row and then are ready to do the sentiment analysis. The output seperate each sentences into one word per row. This can help to future anlysis.

```
text_df <- Achievement%>%
  mutate(linenumber = row_number(),
         chapter = cumsum(str_detect(text,
                                     regex("^chapter [\\divxlc]",
                                           ignore_case = TRUE)))) %>%
  unnest_tokens(word, text)
```

Small sections of text may not have enough words in them to get a good estimate of sentiment while really large sections can wash out narrative structure. We decide to choose 80- line as index to track the net sentiment. Four lexicons are used at below part. The extra dictionary,"loughran", was developed based on analyses of financial reports, and intentionally avoids words like "share" and "fool", as well as subtler terms like "liability" and "risk" that may not have a negative meaning in a financial context.

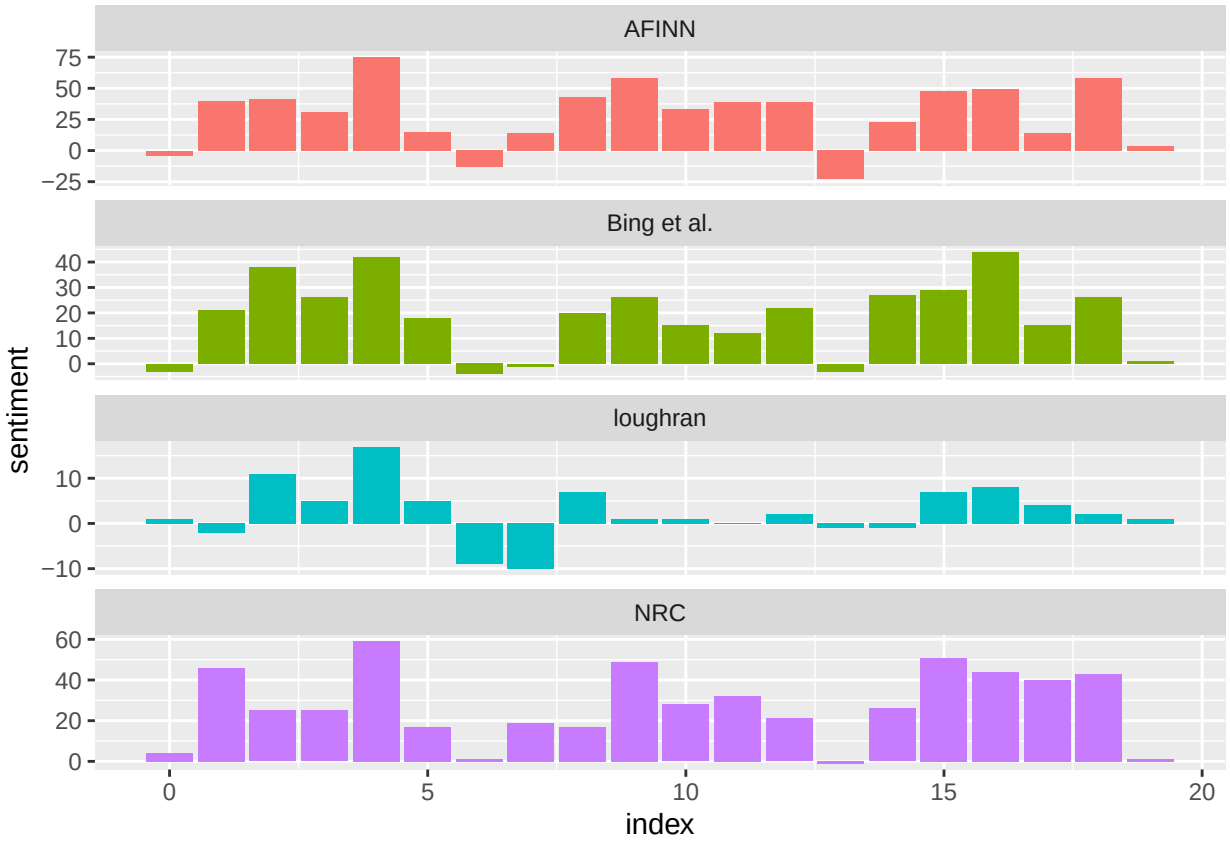
```
afinn <- text_df %>%
  inner_join(get_sentiments("afinn")) %>%
  group_by(index = linenumber %/% 80) %>%
  summarise(sentiment = sum(value)) %>%
  mutate(method = "AFINN")


bing_and_nrc_lou <- bind_rows(
  text_df %>%
    inner_join(get_sentiments("bing")) %>%
    mutate(method = "Bing et al."),
  text_df %>%
    inner_join(get_sentiments("nrc") %>%
                 filter(sentiment %in% c("positive",
                                         "negative"))
    ) %>%
    mutate(method = "NRC"),
  text_df %>%
    inner_join(get_sentiments("loughran") %>%
                 filter(sentiment %in% c("positive",
                                         "negative"))
```

```
    ) %>%
    mutate(method = "loughran")) %>%
  count(method, index = linenumber %/% 80, sentiment) %>%
  pivot_wider(names_from = sentiment,
              values_from = n,
              values_fill = 0) %>%
  mutate(sentiment = positive - negative)
```

Based on above sentiment analysis, we plot four different lexicons. We can obviously find that loughran sentiment score is lower than other three lexicon. The reason behind is that loughran is specially aimed for financial context. My book is a novel which contain lots of sentiment word. This novel talks about Black women's life in America and achieve themselves. The key tone should contain happy and also difficulties of their life. So, the visulazation matches with the plotline of this book.

```
bind_rows(afinn,
          bing_and_nrc_lou) %>%
  ggplot(aes(index, sentiment, fill = method)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~method, ncol = 1, scales = "free_y")
```



The result for the NRC lexicon biased so high in sentiment compared to the Bing et al. The reason is that bing contains more positive word. Also, loughran contain less sentiment in score is because it contain more negative words.

```
get_sentiments("nrc") %>%
  filter(sentiment %in% c("positive", "negative")) %>%
  count(sentiment)
```

```
## # A tibble: 2 x 2
```

```
##   sentiment     n
##   <chr>     <int>
## 1 negative   3318
## 2 positive   2308
```

```
get_sentiments("loughran") %>%
  filter(sentiment %in% c("positive", "negative")) %>%
  count(sentiment)
```

```
## # A tibble: 2 x 2
##   sentiment     n
##   <chr>     <int>
## 1 negative   2355
## 2 positive    354
```

```
get_sentiments("bing") %>%
  count(sentiment)
```

```
## # A tibble: 2 x 2
##   sentiment     n
##   <chr>     <int>
## 1 negative   4781
## 2 positive   2005
```