Camille Text Mining

Wuji Shan

11/29/2021

Book: Camille (LA DAME AUX CAMILIAS)
Author: Alexandre Dumas

```
Cami <- gutenberg_download(1608)
```

```
## Determining mirror for Project Gutenberg from http://www.gutenberg.org/robot/harvest
```

```
## Using mirror http://aleph.gutenberg.org
```

```
newCami <- Cami %>%
  mutate(linenumber = row_number()) %>%
  select(-gutenberg_id) %>%
  mutate(chapter = cumsum(str_detect(text,
                                     regex("^chapter [\\divxlc]",
                                           ignore_case = TRUE))))
```

```
tidy_Cami <- newCami %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words)
```

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

```
tidy_Cami %>%
  count(word, sort = TRUE)
```

```
## # A tibble: 4,134 x 2
##    word           n
##    <chr>      <int>
##  1 marguerite   453
##  2 love         220
##  3 day          175
##  4 time         149
##  5 woman        147
##  6 prudence     143
##  7 life         117
##  8 father       113
##  9 armand       106
## 10 paris         87
## # ... with 4,124 more rows
```

Sentiment Analysis

nrc

```
#textdata::lexicon_nrc(delete = TRUE)
#nrc <- textdata::lexicon_nrc()
nrc_joy <- get_sentiments("nrc") %>%
  filter(sentiment == "joy")

tidy_Cami %>%
  inner_join(nrc_joy) %>%
  count(word, sort = TRUE)
```

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

```
## # A tibble: 200 x 2
##    word       n
##    <chr>  <int>
##  1 love     220
##  2 friend    71
##  3 money     50
##  4 happy     49
##  5 pay       38
##  6 lover     37
##  7 child     34
##  8 found     34
##  9 god       34
## 10 true      29
## # ... with 190 more rows
```

bing

```
get_sentiments("bing")
```

```
## # A tibble: 6,786 x 2
##    word        sentiment
##    <chr>       <chr>
##  1 2-faces     negative
##  2 abnormal    negative
##  3 abolish     negative
##  4 abominable  negative
##  5 abominably  negative
##  6 abominate   negative
##  7 abomination negative
##  8 abort       negative
##  9 aborted     negative
## 10 aborts      negative
## # ... with 6,776 more rows
```

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

```
tidy_Cami %>%
  inner_join(bing_neg) %>%
  count(word, sort = TRUE)
```

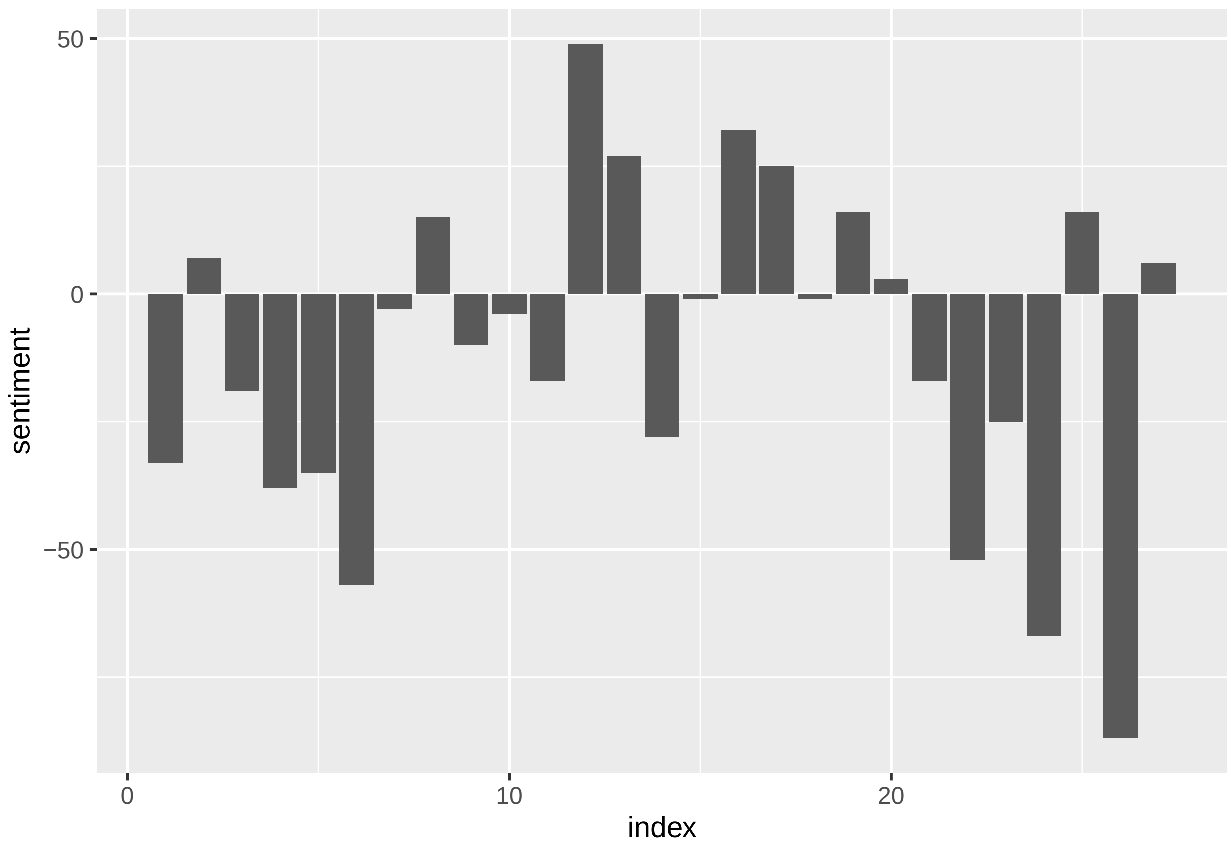
```
## Joining, by = "word"
```

```
## # A tibble: 567 x 2
##    word         n
##    <chr>    <int>
##  1 poor        54
##  2 mistress    47
##  3 dead        33
##  4 rue         33
##  5 doubt       32
##  6 sad         29
##  7 death       28
##  8 die         28
##  9 spite       26
## 10 fear        24
## # ... with 557 more rows
```

```
Cami_sentiment <- tidy_Cami %>%
  inner_join(get_sentiments("bing")) %>%
  count(index = chapter, sentiment) %>%
  pivot_wider(names_from = sentiment, values_from = n, values_fill = 0) %>%
  mutate(sentiment = positive - negative)
```

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

```
ggplot(Cami_sentiment, aes(index, sentiment)) +
  geom_col(show.legend = FALSE)
```



afinn

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

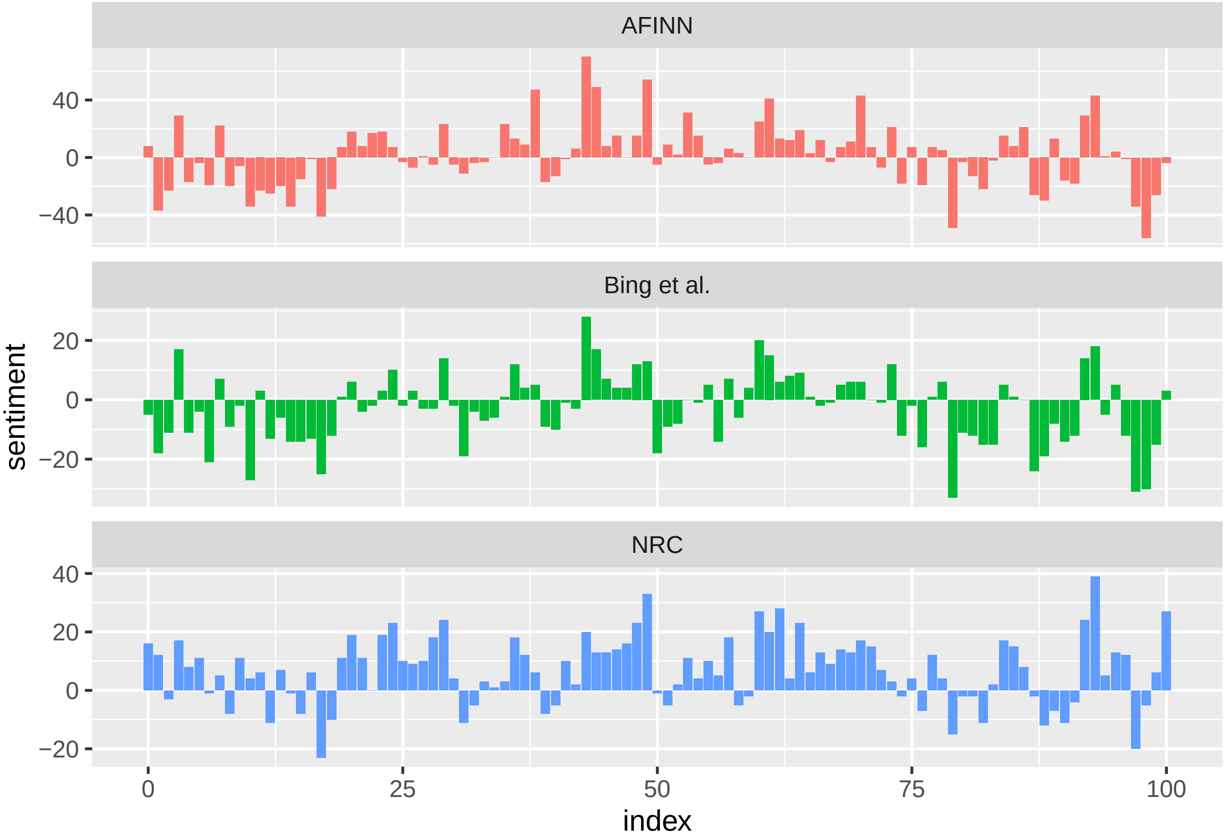
```
## Joining, by = "word"
```

compare the three sentiment dictionaries

```
bing_and_nrc <- bind_rows(
  tidy_Cami %>%
    inner_join(get_sentiments("bing")) %>%
    mutate(method = "Bing et al."),
  tidy_Cami %>%
    inner_join(get_sentiments("nrc") %>%
                 filter(sentiment %in% c("positive",
                                         "negative"))
    ) %>%
    mutate(method = "NRC")) %>%
  count(method, index = linenumber %/% 80, sentiment) %>%
  pivot_wider(names_from = sentiment,
              values_from = n,
              values_fill = 0) %>%
  mutate(sentiment = positive - negative)
```

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

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



later

```
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
```

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

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

2.4 Most common positive and negative words

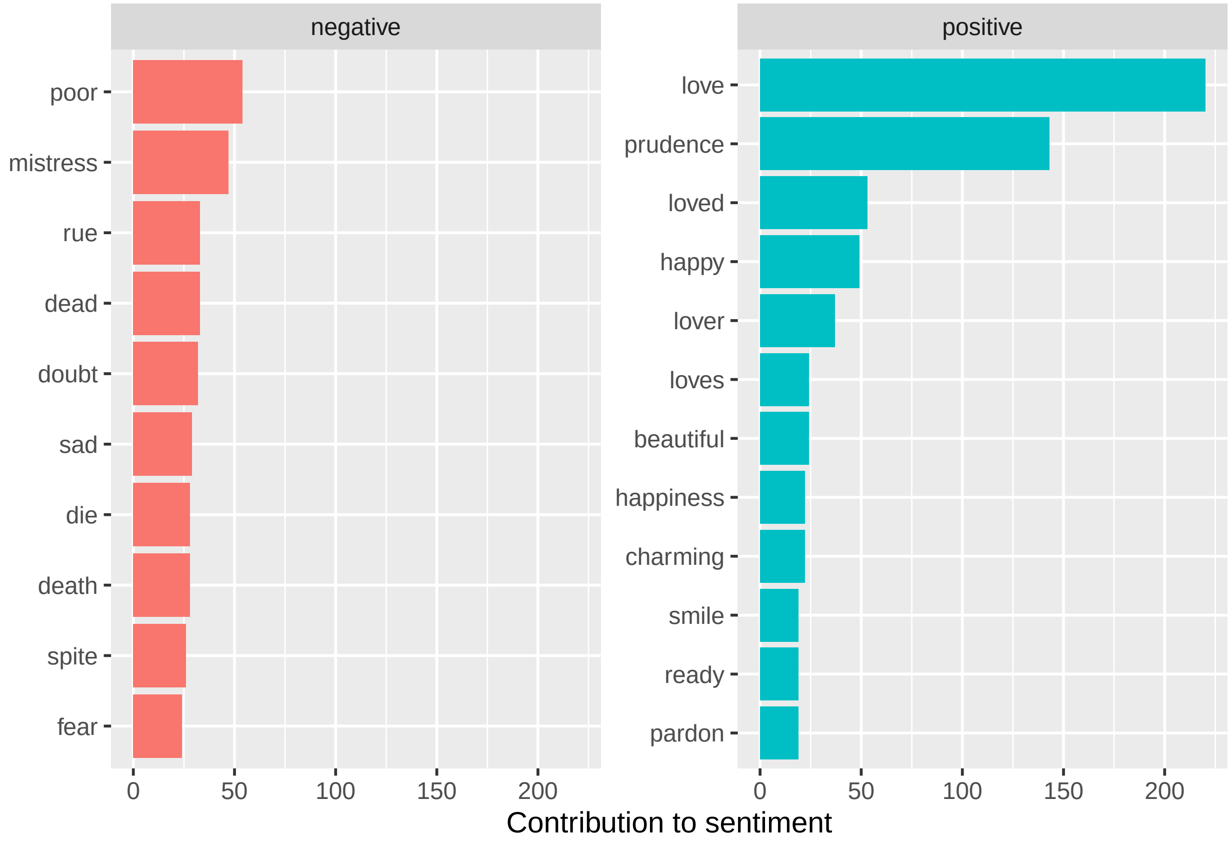
```r
bing_word_counts <- tidy_Cami %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  ungroup()
```

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

```r
bing_word_counts
```

```
## # A tibble: 920 x 3
##    word     sentiment     n
##    <chr>    <chr>     <int>
##  1 love     positive    220
##  2 prudence positive    143
##  3 poor     negative     54
##  4 loved    positive     53
##  5 happy    positive     49
##  6 mistress negative     47
##  7 lover    positive     37
##  8 dead     negative     33
##  9 rue      negative     33
## 10 doubt    negative     32
## # ... with 910 more rows
```

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



2.5 Wordclouds

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

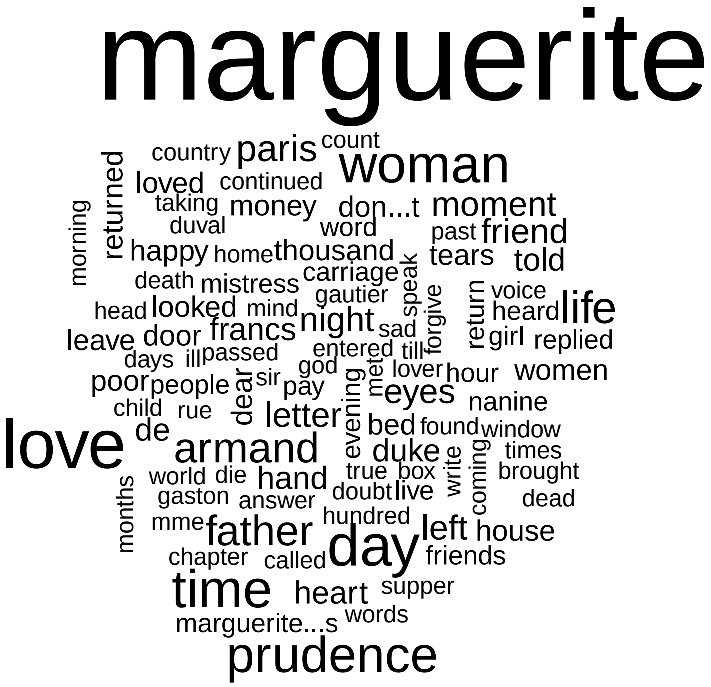
```
## Joining, by = "word"
```

```
## Warning in strwidth(words[i], cex = size[i], ...): conversion failure on 'don't'
## in 'mbcsToSbcs': dot substituted for <e2>
```

```
## Warning in strwidth(words[i], cex = size[i], ...): conversion failure on 'don't'
## in 'mbcsToSbcs': dot substituted for <80>
```

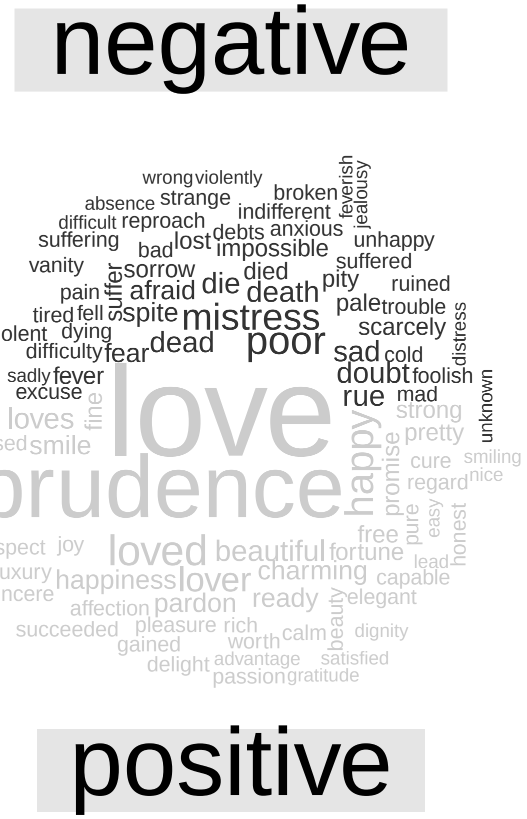
```
## Warning in strwidth(words[i], cex = size[i], ...): conversion failure on 'don't'
## in 'mbcsToSbcs': dot substituted for <99>
```

```
## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
## rotWord * : conversion failure on 'don't' in 'mbcsToSbcs': dot substituted for
## <e2>
## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
## rotWord * : conversion failure on 'don't' in 'mbcsToSbcs': dot substituted for
## <80>

## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
## rotWord * : conversion failure on 'don't' in 'mbcsToSbcs': dot substituted for
## <99>

## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
## rotWord * : font metrics unknown for Unicode character U+2019

## Warning in strwidth(words[i], cex = size[i], ...): conversion failure on
## 'marguerite's' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in strwidth(words[i], cex = size[i], ...): conversion failure on
## 'marguerite's' in 'mbcsToSbcs': dot substituted for <80>

## Warning in strwidth(words[i], cex = size[i], ...): conversion failure on
## 'marguerite's' in 'mbcsToSbcs': dot substituted for <99>

## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt
## = rotWord * : conversion failure on 'marguerite's' in 'mbcsToSbcs': dot
## substituted for <e2>

## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt
## = rotWord * : conversion failure on 'marguerite's' in 'mbcsToSbcs': dot
## substituted for <80>

## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt
## = rotWord * : conversion failure on 'marguerite's' in 'mbcsToSbcs': dot
## substituted for <99>

## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
## rotWord * : font metrics unknown for Unicode character U+2019
```



```
tidy_Cami %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = TRUE) %>%
  acast(word ~ sentiment, value.var = "n", fill = 0) %>%
  comparison.cloud(colors = c("gray20", "gray80"),
                   max.words = 100)
```

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



Find the number of negative words in each chapter and divide by the total words in each chapter. Which chapter has the highest proportion of negative words?

```
wordcounts <- tidy_Cami %>%
  group_by(chapter) %>%
  summarize(words = n())

tidy_Cami %>%
  semi_join(bing_neg) %>%
  group_by(chapter) %>%
  summarize(negativewords = n()) %>%
  left_join(wordcounts, by = c("chapter")) %>%
  mutate(ratio = negativewords/words) %>%
  filter(chapter != 0) %>%
  slice_max(ratio, n = 1) %>%
  ungroup()
```

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

```
## # A tibble: 1 x 4
##   chapter negativewords words ratio
##     <int>         <int> <int> <dbl>
## 1      24           139   906 0.153
```

Chapter 24 has the highest proportion of negative words.