Comparison Barplots with The Shunned House

Andrew Innes

December 8, 2017

Outline

Install and Load Libraries

- ```
  library(dplyr)
  ```
  ```

Install and Load Libraries

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

Install and Load Libraries

- `library(dplyr)`
- `library(tidytext)`
- `library(gutenbergr)`

Install and Load Libraries

- `library(dplyr)`
- `library(tidytext)`
- `library(gutenbergr)`
- `library(ggplot2)`

Install and Load Libraries

- `library(dplyr)`
- `library(tidytext)`
- `library(gutenbergr)`
- `library(ggplot2)`
- `library(stringr)`

Access Project Gutenberg

```
df<-gutenberg_works(str_detect(title,
                               'The Shunned House'))
df$gutenberg_id

## [1] 31469

df$title

## [1] "The Shunned House"
```

Download The Shunned House

```
House<-gutenberg_download(31469)
colnames(House)

## [1] "gutenberg_id" "text"

substr(House$text[400],1,21)

## [1] "of mold in that regio"
```

Unpack the Words

```
shunned_words<-House%>%
  unnest_tokens(word,text)
colnames(shunned_words)

## [1] "gutenberg_id" "word"

shunned_words[398:400,]

## # A tibble: 3 x 2
##   gutenberg_id  word
##          <int> <chr>
## 1        31469  farm
## 2        31469    or
## 3        31469  semi
```

The Bing Lexicon

```
bing<-get_sentiments('bing')
colnames(bing)

## [1] "word"      "sentiment"

bing[398:400,]

## # A tibble: 3 x 2
##          word sentiment
##         <chr>     <chr>
## 1     awkward  negative
## 2 awkwardness  negative
## 3      awsome  positive
```

The Inner Join

```
shunned_words<-inner_join(shunned_words,bing)
shunned_words$gutenberg_id<-NULL
shunned_words[398:400,]

## # A tibble: 3 x 2
##       word sentiment
##      <chr>     <chr>
## 1 ignorant  negative
## 2 smelling  negative
## 3  shunned  negative
```

Top Ten Positive Words I

```
shunned_pos<-shunned_words%>%
  filter(sentiment=='positive')%>%
  group_by(word)%>%
  summarize(count=n(),sentiment=first(sentiment))%>%
  arrange(count)%>%
  top_n(10,wt=count)
```

Top Ten Positive Words II

```
shunned_pos

## # A tibble: 13 x 3
##          word count sentiment
##         <chr> <int>     <chr>
##  1     enough     4  positive
##  2       good     4  positive
##  3        led     4  positive
##  4     master     4  positive
##  5     strong     4  positive
##  6  strongest     4  positive
##  7     proper     5  positive
##  8 providence     8  positive
##  9       well     9  positive
## 10      mercy    10  positive
## 11      great    12  positive
## 12    benefit    13  positive
## 13       like    17  positive
```

Top Ten Negative Words I

```
shunned_neg<-shunned_words%>%
  filter(sentiment=='negative')%>%
  group_by(word)%>%
  summarize(count=n(),sentiment=first(sentiment))%>%
  arrange(count)%>%
  filter(word!='miss')%>%
  top_n(10,wt=count)
```

Top Ten Negative Words II

```
shunned_neg

## # A tibble: 14 x 3
##         word count sentiment
##        <chr> <int>     <chr>
##  1   broken     5  negative
##  2     evil     5  negative
##  3 horrible     5  negative
##  4 peculiar     5  negative
##  5 sinister     5  negative
##  6    smell     5  negative
##  7 terrible     5  negative
##  8    weird     5  negative
##  9  hideous     7  negative
## 10    queer     8  negative
## 11    death     9  negative
## 12  strange    10  negative
## 13     died    11  negative
```

The Comparison Bar Plot I

```
shunned_pos$word<-factor(shunned_pos$word,
                         levels=shunned_pos$word)
shunned_neg$word<-factor(shunned_neg$word,
                         levels=shunned_neg$word)
shunned_comp<-rbind(shunned_pos,shunned_neg)
plot<-ggplot()+
  geom_bar(data=shunned_comp,
           aes(x=word,y=count, fill=sentiment,
               color=sentiment),stat='identity')+
  coord_flip()+
  facet_wrap(~sentiment,scales='free_y')+
  scale_fill_manual(values=c('black','#ea6205'))+
  scale_color_manual(values=c('#ea6205','black'))
```

The Comparison Bar Plot II