Data wrangling and Wordcloud

Akhilesh Atluri

December 14, 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)`

Install and Load Libraries

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

Getting the data from Gutenbergr

```
gutenberg_works(str_detect(title,'The Adventure'))
```

```
## # A tibble: 171 x 8
##    gutenberg_id
##           <int>
##  1           74
##  2          500
##  3          909
##  4         1194
##  5         1218
##  6         1372
##  7         1644
##  8         1661
##  9         1825
## 10         2343
## # ... with 161 more rows, and 7 more variables: title <c
## #   gutenberg_author_id <int>, language <chr>, gutenberg
## #   rights <chr>, has.text <lgl>
```

Getting the words

```
sherlock_words<-sherlock%>%
  unnest_tokens(word,text)
sherlock_words
```

```
## # A tibble: 105,426 x 2
##    gutenberg_id       word
##           <int>      <chr>
##  1         1661        the
##  2         1661 adventures
##  3         1661         of
##  4         1661   sherlock
##  5         1661     holmes
##  6         1661         by
##  7         1661        sir
##  8         1661     arthur
##  9         1661      conan
## 10         1661      doyle
## # ... with 105,416 more rows
```

Getting rid of unwanted words

```
sherlock_words<-sherlock_words%>%
  filter(!(word %in% stop_words$word))
```

grouping each word and getiing the count of each word

```
sherlock_freq<-sherlock_words%>%
  group_by(word)%>%
  summarize(count=n())
```

displaying the words on wordcloud

```
wordcloud(sherlock_freq$word,sherlock_freq$count,min.freq =
```