## Data wrangling and Wordcloud

Akhilesh Atluri

December 14, 2017

#### Outline

Install and Load Libraries

Getting the data from Gutenbergr

Getting the words

Getting rid of unwanted words

grouping each word and getiing the count of each word

displaying the words on wordlcoud

► library(dplyr)

- ► library(dplyr)
- ► library(tidytext)

- ► library(dplyr)
- ► library(tidytext)
- ► library(gutenbergr)

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

- ► library(dplyr)
- ▶ library(tidytext)
- ► library(gutenbergr)
- ► library(ggplot2)
- ► library(stringr)

- ► 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
               500
##
##
   3
               909
##
              1194
              1218
##
   5
##
   6
              1372
## 7
              1644
              1661
## 8
##
              1825
              2343
## 10
## # ... with 161 more rows, and 7 more variables: title <
## #
       gutenberg_author_id <int>, language <chr>, gutenberg
```

rights (chr) has tout (1g1)

## 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
##
              1661 adventures
    3
##
              1661
                            of
##
              1661 sherlock
##
    5
              1661
                       holmes
##
              1661
                            by
##
              1661
                           sir
              1661
                       arthur
##
##
              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 =

# door rose words left sirible found appeared word appearad waiting half paper business singular photograph eyes sitting an strange passed drove reason photograph death. front found john letter word entered rate son leave brought hand held white hat floor goose square go sistersimon black spoke square go father heart miss square go father heart miss square go for good father heart miss square go for good father heart miss square go for good for good father heart miss square go for good for good father heart miss square go for good for good father heart miss square go for good father hear night remarked papers strong office read note ran understand bell dress married set instant chair colonel morning woman fear St returned gentlemansuddenly steps pounds inspector watson windows laid people light gried friend companion round