anna karenina

William Taylor Bickelmann

December 16, 2017

Abstract

I will be analyzing One of Leo Tolstoy's anna karenina

1 "Tolstoy's anna karenina"

First to get the packages into the session.

```
library(tidytext)
library(tm)
library(wordcloud)
library(stringr)
library(dplyr)
library(knitr)
library(gutenbergr)
```

Then we use gutenbergr to extract the data into a data frame.

```
gutenberg_works(str_detect(author, "Tolstoy"))
```

```
## # A tibble: 41 x 8
##    gutenberg_id                                 title            author
##           <int>                                 <chr>             <chr>
##  1          243  The Forged Coupon, and Other Stories Tolstoy, Leo, graf
##  2          689 The Kreutzer Sonata and Other Stories Tolstoy, Leo, graf
##  3          985                        Father Sergius Tolstoy, Leo, graf
##  4          986                        Master and Man Tolstoy, Leo, graf
##  5         1399                         Anna Karenina Tolstoy, Leo, graf
##  6         1938                          Resurrection Tolstoy, Leo, graf
##  7         2142                             Childhood Tolstoy, Leo, graf
##  8         2450                               Boyhood Tolstoy, Leo, graf
##  9         2600                         War and Peace Tolstoy, Leo, graf
## 10         2637                                 Youth Tolstoy, Leo, graf
## # ... with 31 more rows, and 5 more variables: gutenberg_author_id <int>,
## #   language <chr>, gutenberg_bookshelf <chr>, rights <chr>,
## #   has_text <lgl>
```

```
df <- gutenberg_download(1399)
```

2 Cleaning the Text

Now to break up the dataframe into individual

```
words_df <- df%>%
  unnest_tokens(word, text)

head(words_df)

## # A tibble: 6 x 2
##   gutenberg_id       word
##          <int>      <chr>
## 1         1399       anna
## 2         1399   karenina
## 3         1399         by
## 4         1399        leo
## 5         1399    tolstoy
## 6         1399 translated
```

Now to get rid of stop words

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

words_df <- words_df%>%
  filter(!word == "thy" & !word == "thou" & !word == "thee")
```

Next is to use dplyr to create a count for each words

```
words_free <- words_df%>%
  group_by(word)%>%
  summarise(count = n())%>%
  arrange(-count)
#make a count of the word

head(words_free)

## # A tibble: 6 x 2
##      word count
##     <chr> <int>
## 1   levin  1517
## 2 vronsky   776
```

```
## 3     anna   741
## 4  alexey   629
## 5   kitty   598
## 6    time   564
```

3 Wordcloud

Next step is the wordcloud

```r
wordcloud(words_free$word, words_free$count, min.freq = 25)
```