Text Mining the Game of Thrones

Joseph S. Tabadero, Jr.

March 12, 2017

library(tidytext)  
library(tm)

## Loading required package: NLP

got <-Corpus(DirSource("C:/Users/jtabadero-pc/Documents/martin"), readerControl = list(language="en"))  
summary(got)

## Length Class Mode  
## Book 1 - A Game of Thrones.txt 2 PlainTextDocument list  
## Book 2 - A Clash of Kings.txt 2 PlainTextDocument list  
## Book 3 - A Storm of Swords.txt 2 PlainTextDocument list  
## Book 4 - A Feast for Crows.txt 2 PlainTextDocument list  
## Book 5 - A Dance With Dragons.txt 2 PlainTextDocument list

got <- tm\_map(got, removeNumbers)  
got <- tm\_map(got, removePunctuation)  
got <- tm\_map(got , stripWhitespace)  
got <- tm\_map(got, tolower)  
got\_dtm <-DocumentTermMatrix(got)  
got\_dtm

## <<DocumentTermMatrix (documents: 5, terms: 30468)>>  
## Non-/sparse entries: 77433/74907  
## Sparsity : 49%  
## Maximal term length: 53  
## Weighting : term frequency (tf)

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

##   
## Attaching package: 'ggplot2'

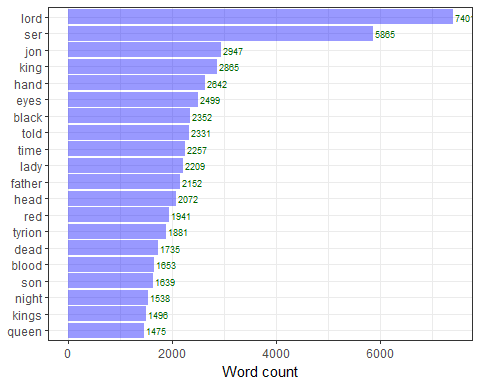
## The following object is masked from 'package:NLP':  
##   
## annotate

got\_tidy <- tidy(got\_dtm)  
  
got\_tidy <- got\_tidy %>%  
 anti\_join(stop\_words, by = c(term = "word"))  
  
got\_tidy

## # A tibble: 74,890 × 3  
## document term count  
## <chr> <chr> <dbl>  
## 1 Book 1 - A Game of Thrones.txt aback 5  
## 2 Book 3 - A Storm of Swords.txt aback 7  
## 3 Book 4 - A Feast for Crows.txt aback 8  
## 4 Book 5 - A Dance With Dragons.txt aback 4  
## 5 Book 1 - A Game of Thrones.txt abandon 3  
## 6 Book 2 - A Clash of Kings.txt abandon 4  
## 7 Book 3 - A Storm of Swords.txt abandon 8  
## 8 Book 4 - A Feast for Crows.txt abandon 5  
## 9 Book 5 - A Dance With Dragons.txt abandon 19  
## 10 Book 1 - A Game of Thrones.txt abandoned 8  
## # ... with 74,880 more rows

got\_tidy %>% count(term, wt = count) %>%  
 filter(n >= 400) %>%  
 mutate(term = reorder(term, n)) %>%  
 top\_n(20) %>%   
 ggplot(aes(term, n)) +  
 geom\_bar(stat = "identity", fill = "blue", alpha = 0.4, show.legend = FALSE) +  
 geom\_text(aes(label = n), hjust = -0.1, color = "darkgreen", size = 2.5) +  
 theme\_bw() +  
 xlab(NULL) + ylab("Word count") +  
 coord\_flip()

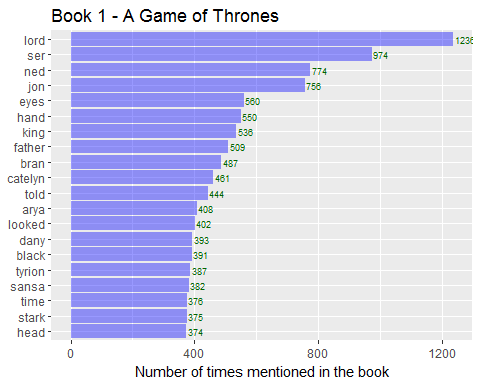
## Selecting by n



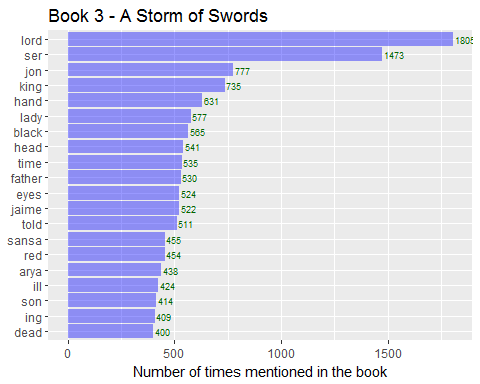
book <- unique(got\_tidy$document)  
  
plot\_game <- function(book) {  
out <- got\_tidy %>% group\_by(document) %>%  
 count(term, wt = count) %>%  
 filter(document == book) %>%  
 mutate(term = reorder(term, n)) %>%  
 top\_n(20)   
out %>% ggplot(aes(term, n)) +  
 geom\_bar(stat = "identity", alpha = 0.4, fill = "blue", show.legend = FALSE) +  
 xlab(NULL) + ylab("Number of times mentioned in the book") +   
 geom\_text(aes(label = n), hjust = -0.1, color = "darkgreen", size = 2.5) +  
 ggtitle(substr(basename(book), 1, nchar(basename(book)) - 4)) +  
 coord\_flip()  
}  
lapply(book, plot\_game)

## Selecting by n  
## Selecting by n  
## Selecting by n  
## Selecting by n  
## Selecting by n

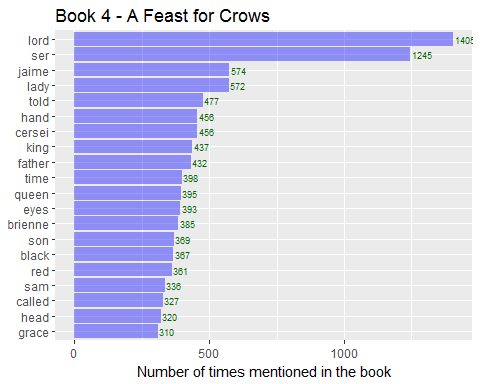
## [[1]]



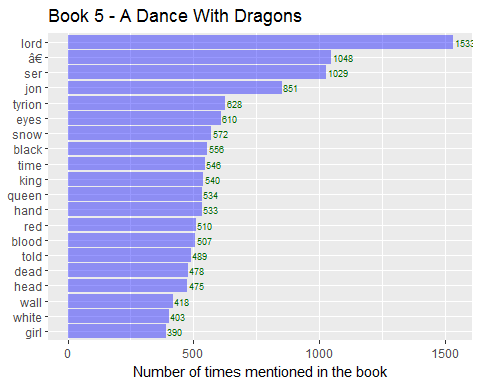
##   
## [[2]]



##   
## [[3]]



##   
## [[4]]



##   
## [[5]]

