Text Mining the Game of Thrones

Joseph S. Tabadero, Jr.

March 12, 2017

```
library(tidytext)
library(tm)
```

```
## Loading required package: NLP
```

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

```
##                                  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
```

```
a <- tm_map(a, removeNumbers)
a <- tm_map(a, removePunctuation)
a <- tm_map(a , stripWhitespace)
a <- tm_map(a, tolower)
adtm <-DocumentTermMatrix(a)
adtm
```

```
## <<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
```

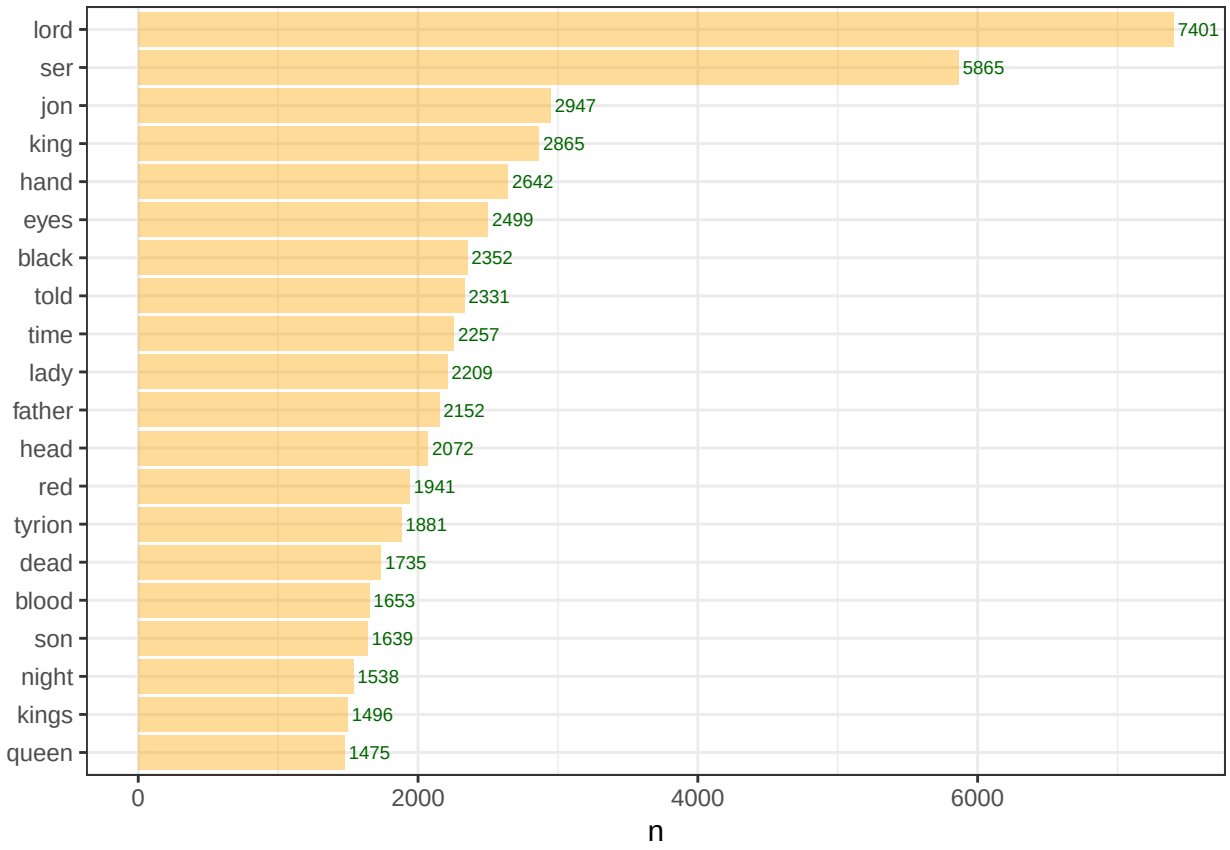
```
game <- tidy(adtm)
```

```
game <- game %>%
  anti_join(stop_words, by = c(term = "word"))

game
```

```
## # 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
```

```
game %>%  count(term, wt = count) %>%
  filter(n >= 400) %>%
  mutate(term = reorder(term, n)) %>%
  top_n(20) %>%
  ggplot(aes(term, n)) +
  geom_bar(stat = "identity", fill = "orange", alpha = 0.4, show.legend = FALSE) +
  geom_text(aes(label = n), hjust = -0.1, color = "darkgreen", size = 2.5) +
  theme_bw() +
  xlab(NULL) +
  coord_flip()
```

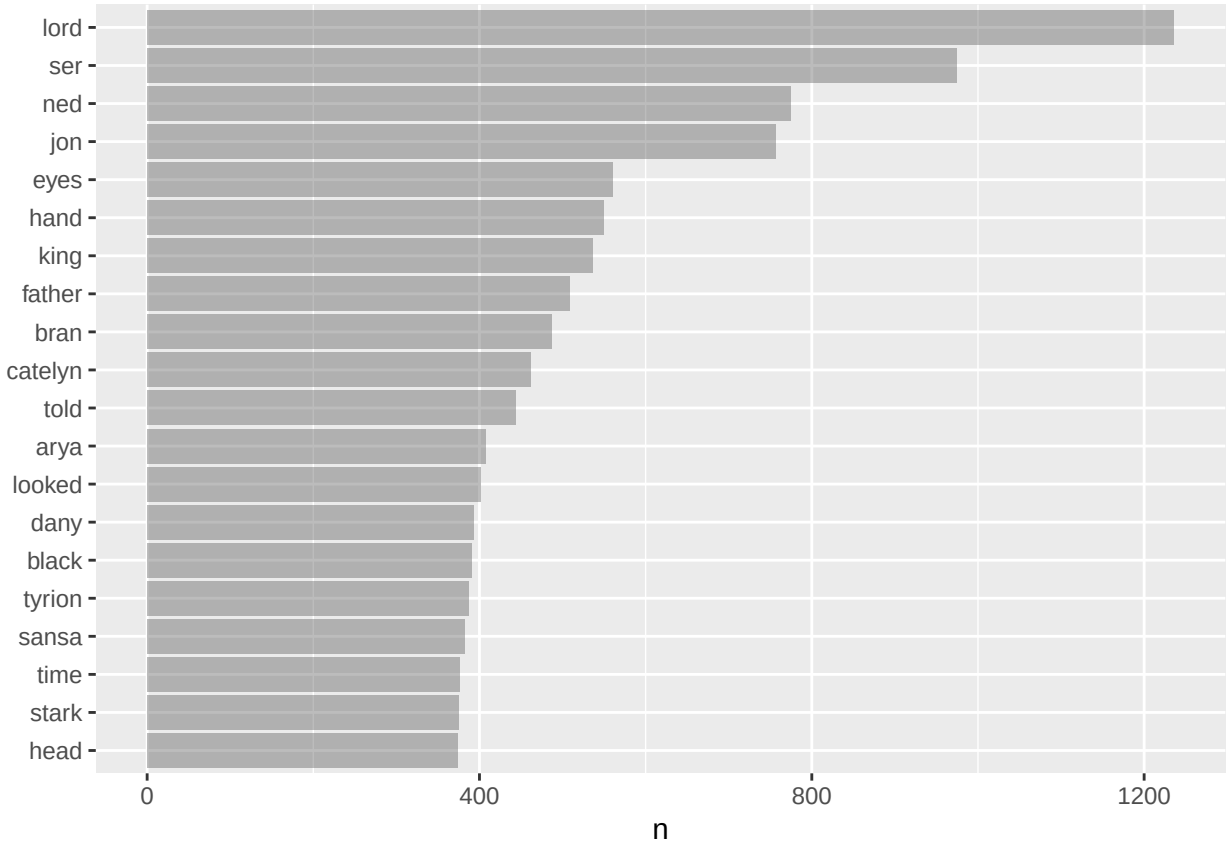
```
## Selecting by n
```



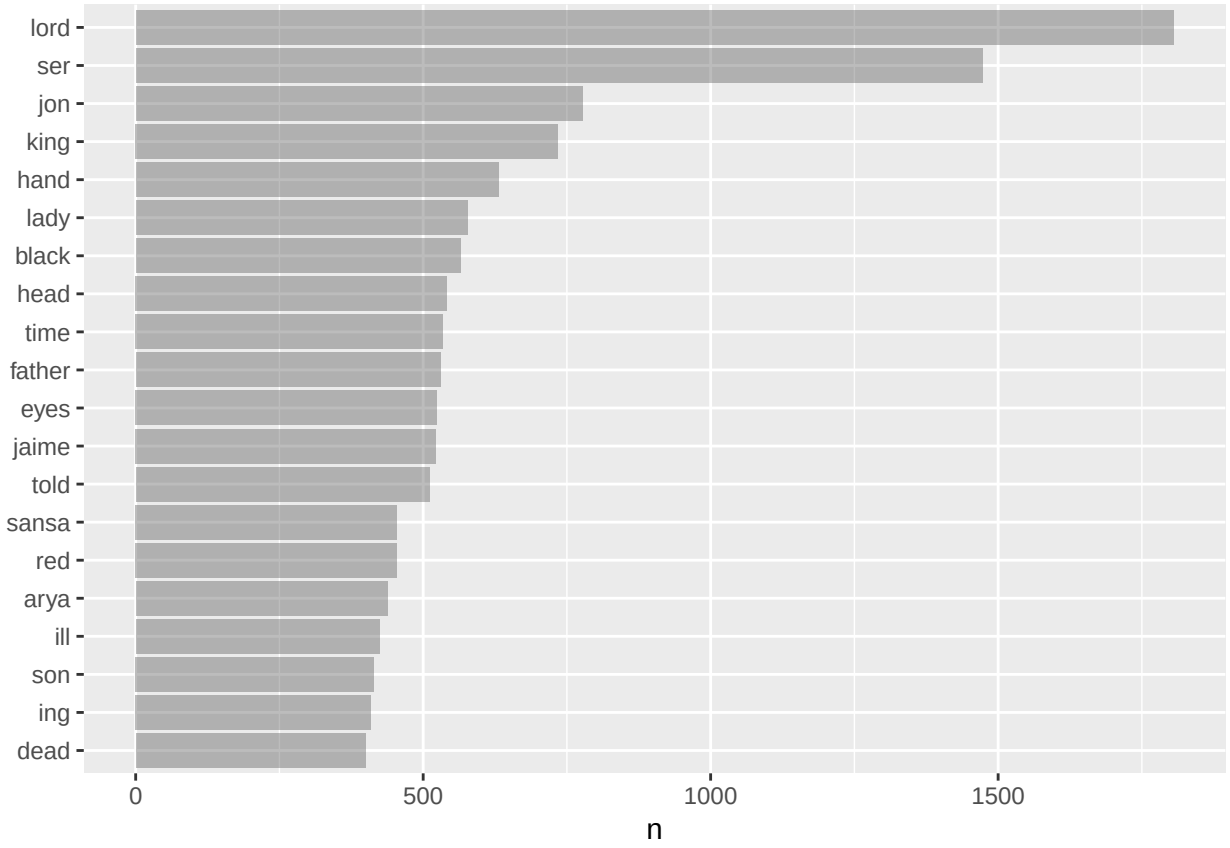

```
book <- unique(game$document)

plot_game <- function(book) {
out <- game %>% group_by(document) %>%
  count(term, wt = count) %>%
  filter(n >= 200, document == book) %>%
  mutate(term = reorder(term, n)) %>%
  top_n(20)
out %>%  ggplot(aes(term, n)) +
  geom_bar(stat = "identity", alpha = 0.4, show.legend = FALSE) +
  xlab(NULL) +
  coord_flip()
}
lapply(book, plot_game)
```

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

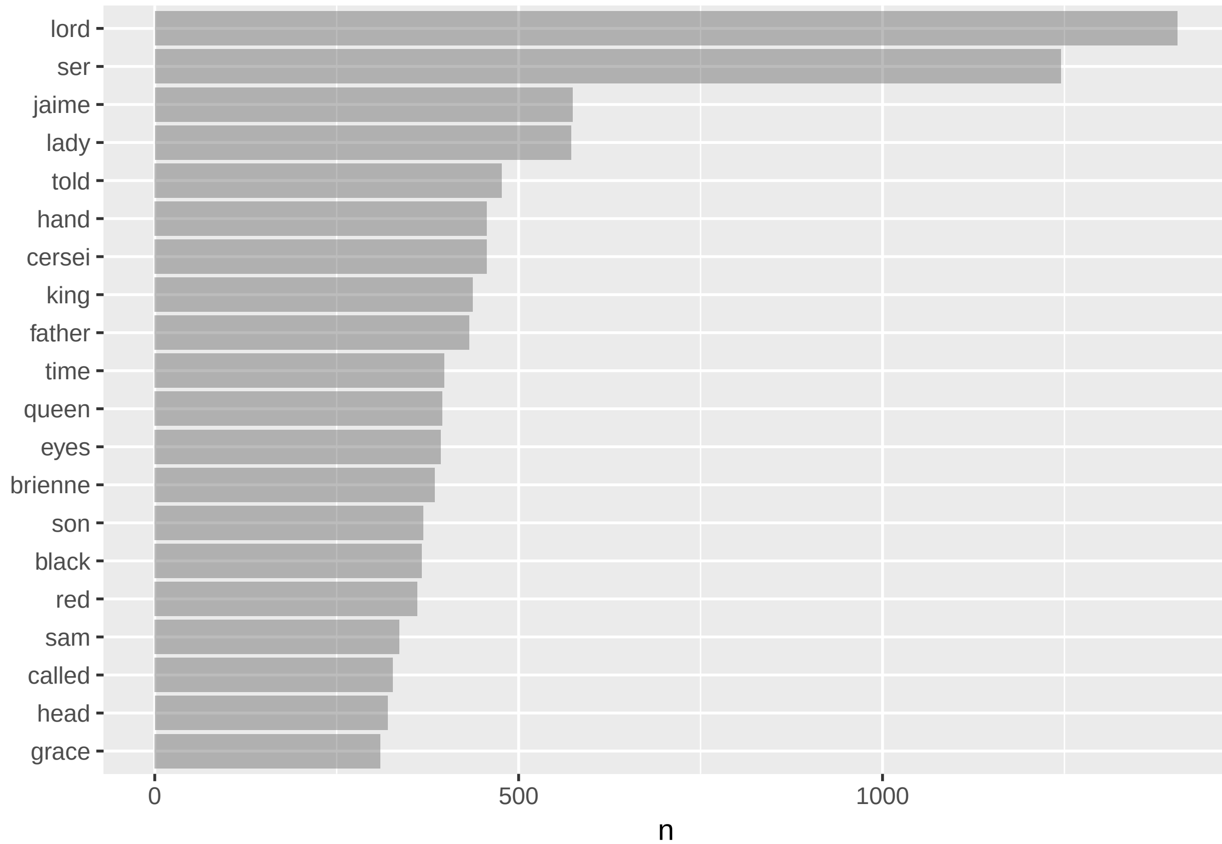
```
## [[1]]
```


lord
ser
ned
jon
eyes
hand
king
father
bran
catelyn
told
arya
looked
dany
black
tyrion
sansa
time
stark
head
0
400
800
1200
n

```
##
## [[2]]
```


lord
ser
jon
king
hand
lady
black
head
time
father
eyes
jaime
told
sansa
red
arya
ill
son
ing
dead
0
500
1000
1500
n

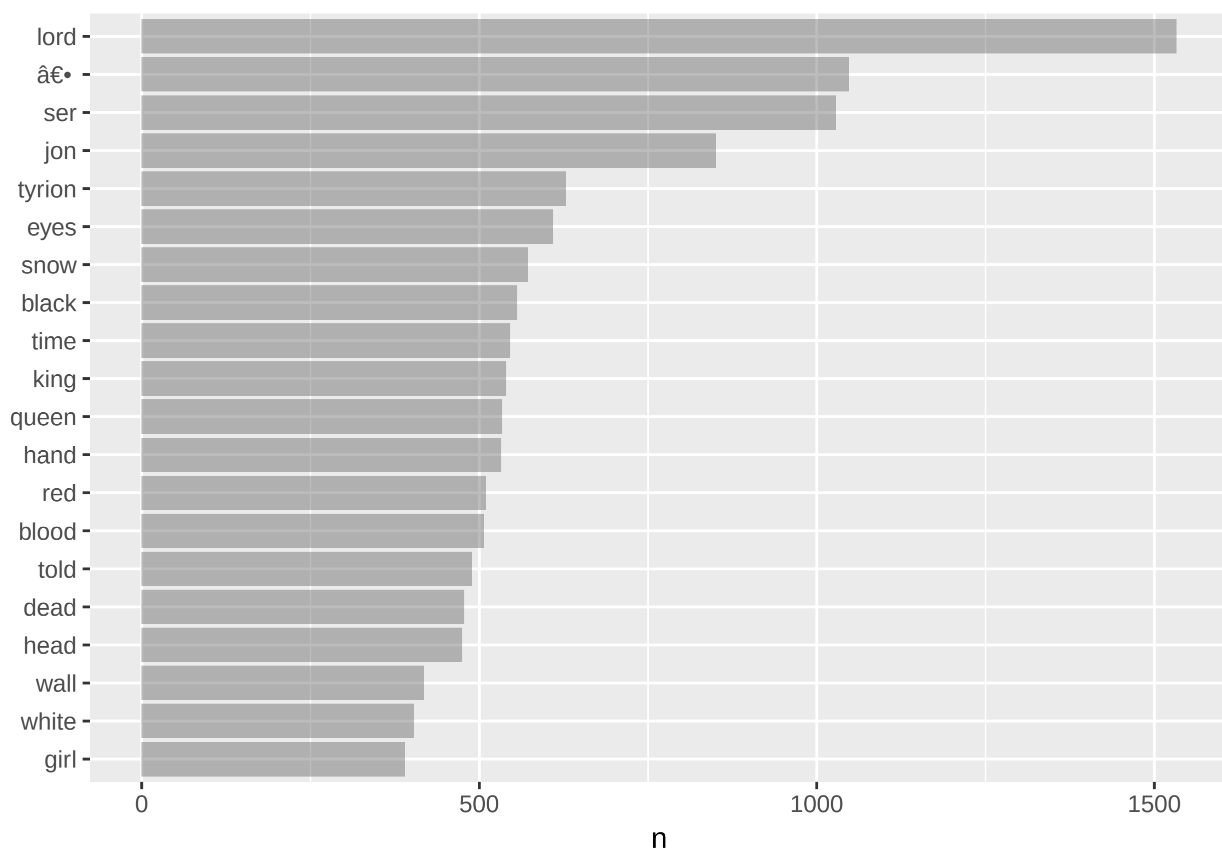
```
##
## [[3]]
```



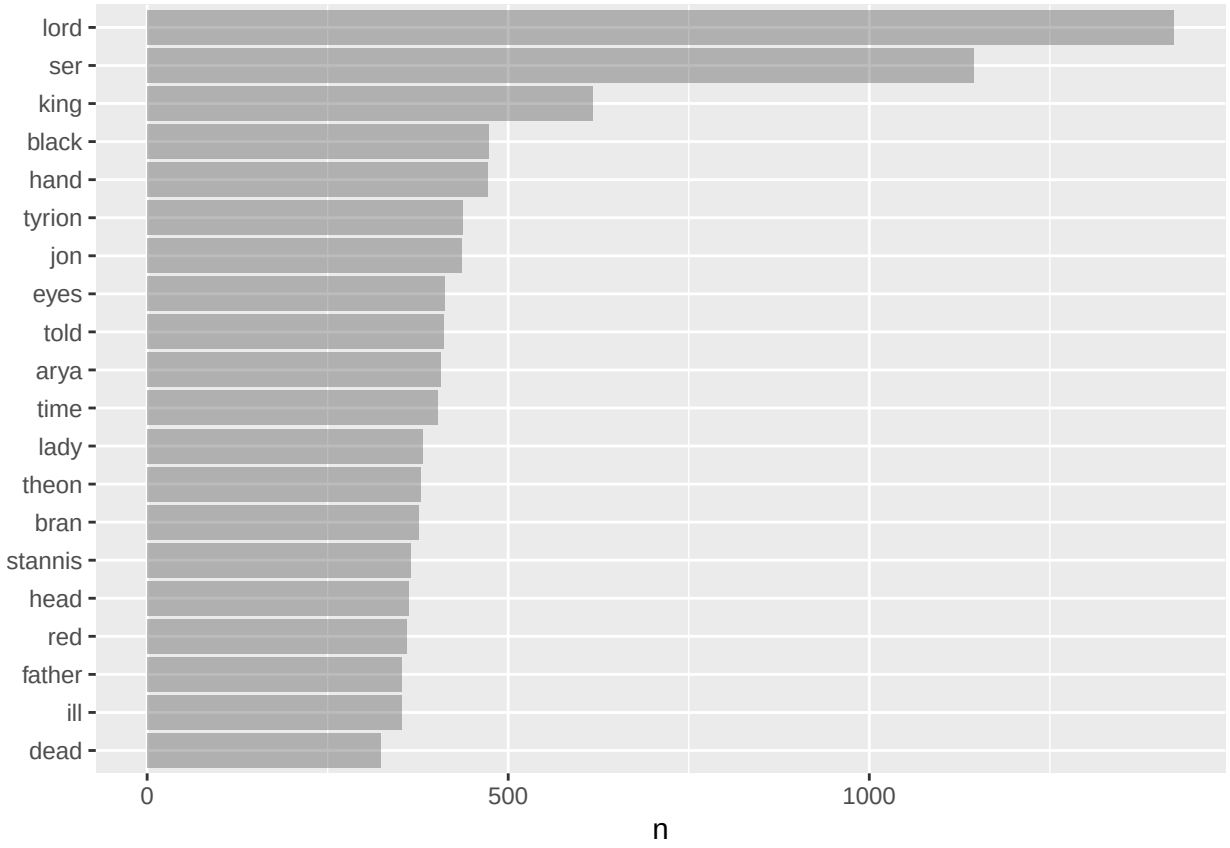

```
##
## [[4]]
```

```
## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d
```

```
## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9d

## Warning in grid.Call.graphics(L_text, as.graphicsAnnot(x$label), x$x, x
## $y, : font width unknown for character 0x9d
```




```
##
## [[5]]
```


lord
ser
king
black
hand
tyrion
jon
eyes
told
arya
time
lady
theon
bran
stannis
head
red
father
ill
dead
0
500
1000
n