Chapter 3

Nick Lauerman

Started 11-Mar-2020

Contents

Accessing Word Data **2**

Recycling **2**

Exercises **3**

3.1 3

Exercise 3.2 5

Exercise 3.3 5

Exercise 3.4 6

```
text.v <- scan(file = "./SupportingMaterials/data/plainText/melville.txt",
               what = "character",
               sep = "\n")
start.v <- which(text.v == "CHAPTER 1. Loomings.")
end.v <- which(text.v == "orphan.")
start.metadata.v <- text.v[1:start.v-1]
end.metadata.v <- text.v[end.v+1:length(text.v)]
novel.lines.v <- text.v[start.v:end.v]
metadata.v <- c(text.v[1:(start.v-1)],
                text.v[(end.v+1):length(text.v)])
novel.v <- paste(novel.lines.v,
                 collapse = " ")
novel.lower.v <- tolower(novel.v)
moby.words.l <- strsplit(novel.lower.v,
                         "\\W")
moby.word.v <- unlist(moby.words.l)
not.blanks.v <- which(moby.word.v != "")



moby.word.v <- moby.word.v[not.blanks.v]
whale.hits.v <- length(which(moby.word.v == "whale"))
total.words.v <- length(moby.word.v)
moby.fraqs.t <- table(moby.word.v)
sorted.moby.fraq.t <- sort(moby.fraqs.t,
                           decreasing = TRUE)
```

Accessing Word Data

```
sorted.moby.fraq.t["he"]
```

```
##   he
## 1876
```

```
sorted.moby.fraq.t["she"]
```

```
## she
## 114
```

```
sorted.moby.fraq.t["him"]
```

```
##  him
## 1058
```

```
sorted.moby.fraq.t["her"]
```

```
## her
## 330
```

```
sorted.moby.fraq.t["him"] / sorted.moby.fraq.t["her"]
```

```
##      him
## 3.206061
```

```
sorted.moby.fraq.t["he"] / sorted.moby.fraq.t["she"]
```

```
##       he
## 16.45614
```

```
length(moby.word.v)
```

```
## [1] 214889
```

```
sum(sorted.moby.fraq.t)
```

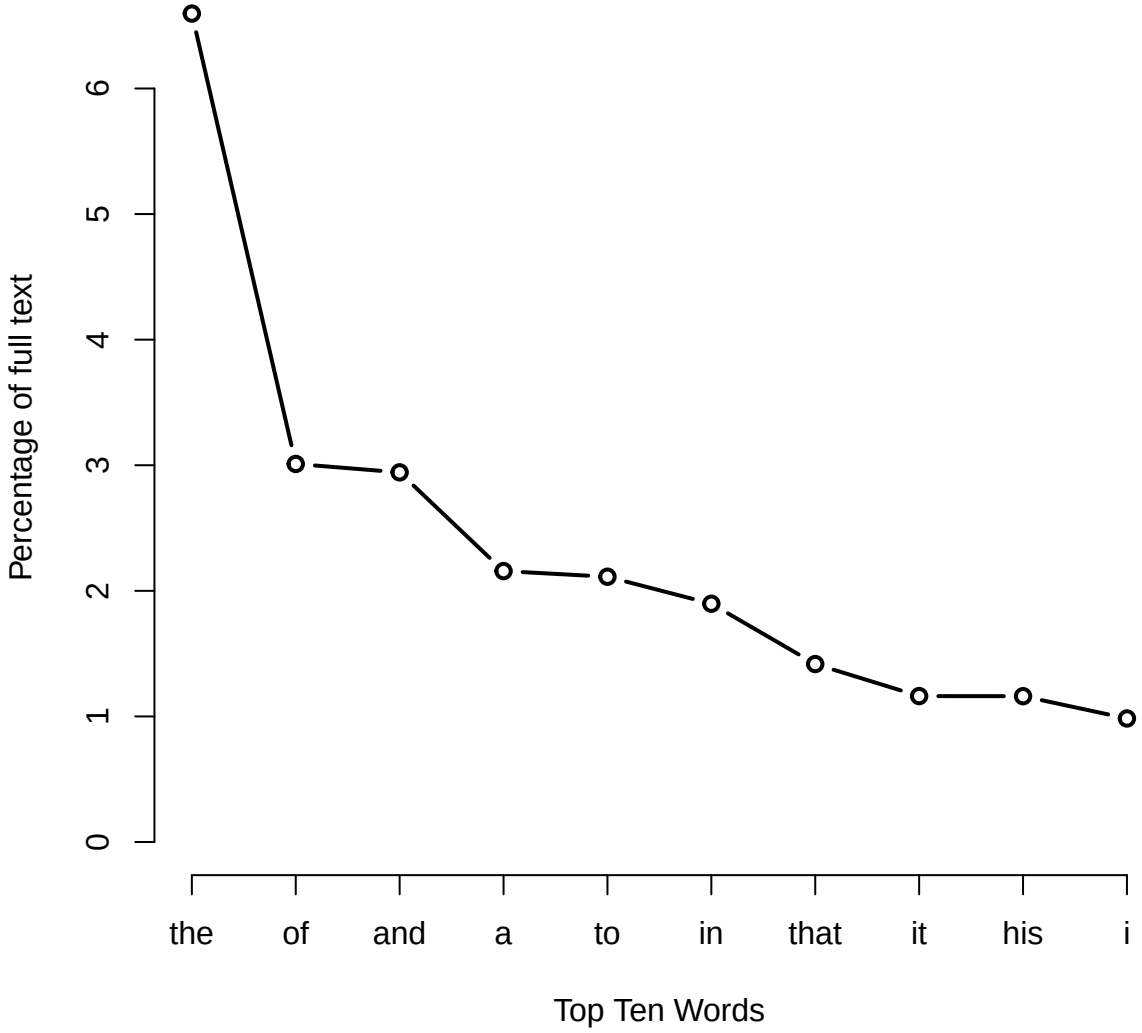
```
## [1] 214889
```

Recycling

```
sorted.moby.rel.freqs.t <- 100 * (sorted.moby.fraq.t/sum(sorted.moby.fraq.t))
sorted.moby.rel.freqs.t["the"]
```

```
##      the
## 6.596429
```

```
#, fig.cap="section 3.2"}
plot(sorted.moby.rel.freqs.t[1:10],
     type = "b",
     xlab = "Top Ten Words",
     ylab = "Percentage of full text",
     xaxt = "n")
axis(1, 1:10,
     labels = names(sorted.moby.rel.freqs.t[1:10]))
```



Exercises

3.1

```
austen.text.v <- scan(file = "./SupportingMaterials/data/plainText/austen.txt",
               what = "character",
               sep = "\n")
austen.first <- which(austen.text.v == "CHAPTER 1")
austen.last <- which(austen.text.v == "THE END")
austen.last <- austen.last - 1
austen.lines <- austen.text.v[austen.first:austen.last]
austen.text <- paste(austen.lines,
                     collapse = " ")
```

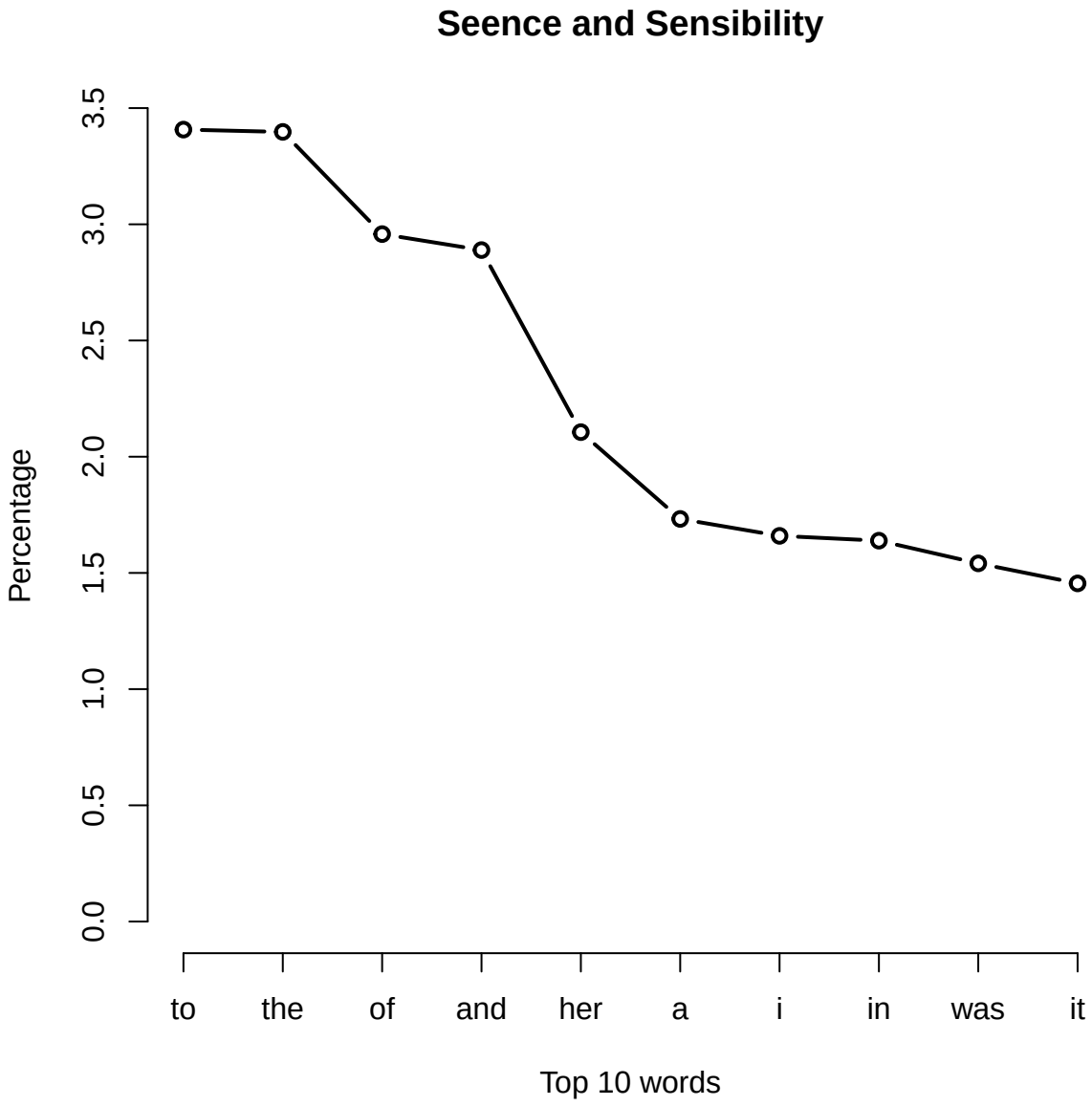
```
austen.lower <- tolower(austen.text)
austen.words <- strsplit(austen.lower,
                         "\\W")
austen.words.v <- unlist(austen.words)
not.blanks <- which(austen.words.v != "")
austen.words.v <- austen.words.v[not.blanks]

austen.sorted.freq.table <- sort(table(austen.words.v),
                                 decreasing = TRUE)
head(austen.sorted.freq.table,
     n = 10)
```

```
## austen.words.v
##   to  the   of  and  her    a    i   in  was   it
## 4115 4103 3572 3489 2543 2092 2004 1979 1861 1757
```

```
austen.sorted.rel.freq.table <- 100*(austen.sorted.freq.table /
                                       sum(austen.sorted.freq.table))
```

```
#, fig.cap="exercise 3.1}
plot(austen.sorted.rel.freq.table[1:10],
     type = "b",
     main = "Seence and Sensibility",
     xlab = "Top 10 words",
     ylab = "Percentage",
     xaxt = "n")
axis(1, 1:10,labels = names(austen.sorted.rel.freq.table[1:10]))
```



Exercise 3.2

```
unique(c(names(sorted.moby.rel.freqs.t)[1:10],
         names(austen.sorted.rel.freq.table)[1:10]))
```

```
##  [1] "the"  "of"   "and"  "a"    "to"   "in"   "that" "it"   "his"  "i"
## [11] "her"  "was"
```

Exercise 3.3

```
names(austen.sorted.rel.freq.table[
  which(names(austen.sorted.rel.freq.table[1:10])
              %in% names(sorted.moby.rel.freqs.t[1:10]))
```

```
])
```

```
## [1] "to"  "the" "of"  "and" "a"   "i"   "in"  "it"
```

Exercise 3.4

```
presentAusten <- which(names(austen.sorted.rel.freq.table[1:10])
              %in% names(sorted.moby.rel.freqs.t[1:10]))
names(austen.sorted.rel.freq.table[1:10])[-presentAusten]
```

```
## [1] "her" "was"
```

```
presentMoby<- which(names(sorted.moby.rel.freqs.t[1:10])
              %in% names(austen.sorted.rel.freq.table[1:10]))
names(sorted.moby.rel.freqs.t[1:10])[-presentMoby]
```

```
## [1] "that" "his"
```