Lab 4: Text Manipulation

Name: Rufus Petrie

This week's agenda: basic string manipulations; practice reading in and summarizing real text data (Shakespeare); practice with iteration; just a little bit of regular expressions.

Some string basics

- **1a.** Define two strings variables, equal to "Statistical Computing" and 'Statistical Computing', and check whether they are equal. What do you conclude about the use of double versus single quotation marks for creating strings in R? Give an example that shows why might we prefer to use double quotation marks as the standard (think of apostrophes).

```
x <- "Statistical Computing"
y <- 'Statistical Computing'
x == y
```

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

Although the strings are equal, it's better to use double quoations so that you can include apostrophees in strings.

- **1b.** The functions `tolower()` and `toupper()` do as you'd expect: they convert strings to all lower case characters, and all upper case characters, respectively. Apply them to the strings below, as directed by the comments, to observe their behavior.

```
s1 <- "I'M NOT ANGRY I SWEAR"         # Convert to lower case
s2 <- "Mom, I don't want my veggies"  # Convert to upper case
s3 <- "Hulk, sMasH"                   # Convert to upper case
s4 <- "R2-D2 is in prime condition, a real bargain!" # Convert to lower case
s1 <- tolower(s1)
s2 <- toupper(s2)
s3 <- toupper(s3)
s4 <- tolower(s4)
```

- **1c.** Consider the string vector `presidents` of length 5 below, containing the last names of past US presidents. Define a string vector `first.letters` to contain the first letters of each of these 5 last names. Hint: use `substr()`, and take advantage of vectorization; this should only require one line of code. Define `first.letters.scrambled` to be the output of `sample(first.letters)` (the `sample()` function can be used to perform random permutations, we'll learn more about it later in the course). Lastly, reset the first letter of each last name stored in `presidents` according to the scrambled letters in `first.letters.scrambled`. Hint: use `substr()` again, and take advantage of vectorization; this should only take one line of code. Display these new last names.

```
presidents = c("Clinton", "Bush", "Reagan", "Carter", "Ford")
first.letters <- substr(presidents, 1, 1)
first.letters.scrambled <- sample(first.letters)
substr(presidents, 1, 1) <-  first.letters.scrambled
presidents
```

```
## [1] "Clinton" "Cush"    "Reagan"  "Barter"  "Ford"
```

- **1d.** Now consider the string `phrase` defined below. Using `substr()`, replace the first four characters in `phrase` by "Provide". Print `phrase` to the console, and describe the behavior you are observing. Using `substr()` again, replace the last five characters in `phrase` by "kit" (don't use the length of `phrase` as magic constant in the call to `substr()`, instead, compute the length using `nchar()`). Print `phrase` to the console, and describe the behavior you are observing.

```
phrase = "Give me a break"
substring(phrase, 1, 4) <- "Provide"
print(phrase)
```

```
## [1] "Prov me a break"
```

```
substr(phrase, nchar(phrase)-4, nchar(phrase)) <- "kit"
print(phrase)
```

```
## [1] "Prov me a kitak"
```

Replacing with substring will truncate the string you use if you don't allow enough places, and it will keep the original characters if your string is too short for the given spaces.

- **1e.** Consider the string `ingredients` defined below. Using `strsplit()`, split this string up into a string vector of length 5, with elements "chickpeas", "tahini", "olive oil", "garlic", and "salt." Using `paste()`, combine this string vector into a single string "chickpeas + tahini + olive oil + garlic + salt". Then produce a final string of the same format, but where the ingredients are sorted in alphabetical (increasing) order.

```
ingredients = "chickpeas, tahini, olive oil, garlic, salt"
ingredients <- strsplit(ingredients, split=", ")
ingredients.sorted <- sort(ingredients[[1]])
ingredients <- paste(ingredients[[1]], collapse=" + ")
ingredients.sorted <- paste(ingredients.sorted, collapse=" + ")
```

Shakespeare's complete works

Project Gutenberg offers over 50,000 free online books, especially old books (classic literature), for which copyright has expired. We're going to look at the complete works of William Shakespeare, taken from the Project Gutenberg website.

To avoid hitting the Project Gutenberg server over and over again, we've grabbed a text file from them that contains the complete works of William Shakespeare and put it on our course website. Visit http://www.stat.cmu.edu/~ryantibs/statcomp/data/shakespeare.txt in your web browser and just skim through this text file a little bit to get a sense of what it contains (a whole lot!).

Reading in text, basic exploratory tasks

- **2a.** Read in the Shakespeare data linked above into your R session with `readLines()`. Make sure you are reading the data file directly from the web (rather than locally, from a downloaded file on your computer). Call the result `shakespeare.lines`. This should be a vector of strings, each element representing a "line" of text. Print the first 5 lines. How many lines are there? How many characters in the longest line? What is the average number of characters per line? How many lines are there with zero characters (empty lines)? Hint: each of these queries should only require one line of code; for the last one, use an on-the-fly Boolean comparison and `sum()`.

```
shakespeare.lines <- readLines("http://www.stat.cmu.edu/~ryantibs/statcomp/data/shakespeare.txt")
shakespeare.lines[1:5]
```

```
## [1] "ï»¿"
## [2] "Project Gutenbergâ\200\231s The Complete Works of William Shakespeare, by"
## [3] "William Shakespeare"
## [4] ""
## [5] "This eBook is for the use of anyone anywhere in the United States and"
```

```
length(shakespeare.lines)
```

```
## [1] 147838
```

```
nchar(max(shakespeare.lines))
```

```
## [1] 70
```

```
sum(nchar(shakespeare.lines))/length(shakespeare.lines)
```

```
## [1] 37.61482
```

```
sum(nchar(shakespeare.lines)>0)
```

```
## [1] 130095
```

There are 147,838 lines. There are 70 characters in the longest line. The average characters per line is 37.61. There are 130,095 lines with characters.

- **2b.** Remove all empty lines from `shakespeare.lines` (i.e., lines with zero characters). Check that that the new length of `shakespeare.lines` makes sense to you.

```
shakespeare.lines <- shakespeare.lines[nchar(shakespeare.lines)>0]
length(shakespeare.lines)
```

```
## [1] 130095
```

- **2c.** Collapse the lines in `shakespeare.lines` into one big string, separating each line by a space in doing so, using `paste()`. Call the resulting string `shakespeare.all`. How many characters does this string have? How does this compare to the sum of characters in `shakespeare.lines`, and does this make sense to you?

```
shakespeare.all <- paste(shakespeare.lines, collapse=" ")
nchar(shakespeare.all)
```

```
## [1] 5690994
```

```
sum(nchar(shakespeare.lines))
```

```
## [1] 5560900
```

shakespeare.all has slightly more characters. This makes sense because we're adding a space with each line.

- **2d.** Split up `shakespeare.all` into words, using `strsplit()` with `split=" "`. Call the resulting string vector (note: here we are asking you for a vector, not a list) `shakespeare.words`. How long is this vector, i.e., how many words are there? Using the `unique()` function, compute and store the unique words as `shakespeare.words.unique`. How many unique words are there?

```
shakespeare.words <- strsplit(shakespeare.all, split=" ")
shakespeare.words <- unlist(shakespeare.words)
length(shakespeare.words)
```

```
## [1] 1370375
```

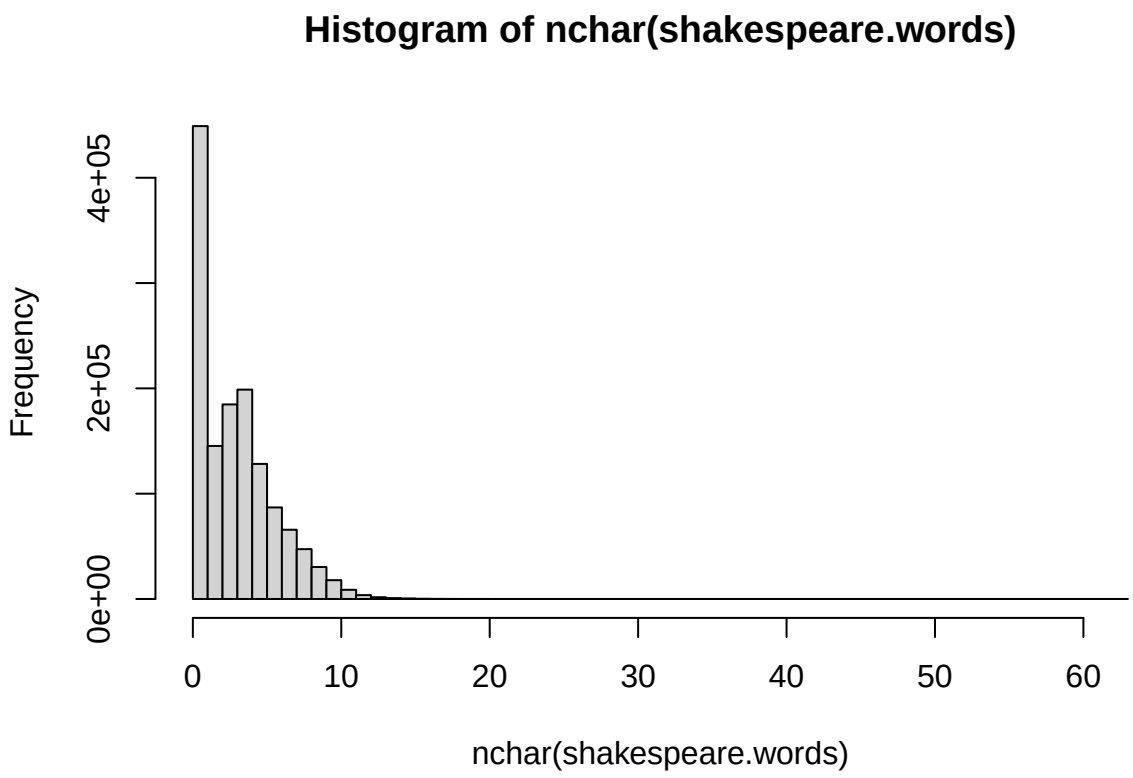
```
shakespeare.words.unique <- unique(shakespeare.words)
length(shakespeare.words.unique)
```

```
## [1] 76172
```

There are 1,370,375 words. There are 76,172 unique words.

- **2e.** Plot a histogram of the number of characters of the words in `shakespeare.words.unique`. You will have to set a large value of the `breaks` argument (say, `breaks=50`) in order to see in more detail what is going on. What does the bulk of this distribution look like to you? Why is the x-axis on the histogram extended so far to the right (what does this tell you about the right tail of the distribution)?

```
hist(nchar(shakespeare.words), breaks = 50)
```



The vast majority of words have 10 or fewer characters, but there are probably a few very long words that drag out the tail of the distribution.

- **2f.** Reminder: the `sort()` function sorts a given vector into increasing order; its close friend, the `order()` function, returns the indices that put the vector into increasing order. Both functions can take `decreasing=TRUE` as an argument, to sort/find indices according to decreasing order. See the code below for an example.

```
set.seed(0)
(x = round(runif(5, -1, 1), 2))
```

```
## [1]  0.79 -0.47 -0.26  0.15  0.82
```

```
sort(x, decreasing=TRUE)
```

```
## [1]  0.82  0.79  0.15 -0.26 -0.47
```

```
order(x, decreasing=TRUE)
```

```
## [1] 5 1 4 3 2
```

Using the `order()` function, find the indices that correspond to the top 5 longest words in `shakespeare.words.unique`. Then, print the top 5 longest words themselves. Do you recognize any of these as actual words? **Challenge**: try to pronounce the fourth longest word! What does it mean?

```
order(nchar(shakespeare.words.unique), decreasing=TRUE)[1:10]
```

```
##  [1] 73710 24445 27347 61226 75770 66628 46843 59842 60611 60908
```

```
shakespeare.words.unique[order(nchar(shakespeare.words.unique), decreasing=TRUE)][1:10]
```

```
##  [1] "_____________________________________________________________"
##  [2] "tragical-comical-historical-pastoral,"
##  [3] "both!â\200"Bardolph!â\200"Peto!â\200"Iâ\200\231ll"
##  [4] "enemies?â\200"Capulet,â\200"Montague,â\200""
##  [5] "http://www.gutenberg.org/1/0/100/"
##  [6] "six-or-seven-times-honourâ\200\231d"
##  [7] "honorificabilitudinitatibus;"
##  [8] "thisâ\200"sir-reverenceâ\200"love,"
##  [9] "study.â\200"By-and-by!â\200"God's"
## [10] "mistress!â\200"Juliet!â\200"fast,"
```

The longest actual word appears to be honorificabilitudinitatibus, which is the Latin word for being in the state of being able to achieve honors.

Computing word counts

- **3a.** Using `table()`, compute counts for the words in `shakespeare.words`, and save the result as `shakespeare.wordtab`. How long is `shakespeare.wordtab`, and is this equal to the number of unique words (as computed above)? Using named indexing, answer: how many times does the word "thou" appear? The word "rumour"? The word "gloomy"? The word "assassination"?

```
shakespeare.wordtab <- table(shakespeare.words)
shakespeare.wordtab[c("thou", "rumour", "gloomy", "assassination")]
```

```
## shakespeare.words
##          thou        rumour        gloomy assassination
##          4522             7             3             1
```

Thou appears 4,522 times, rumour appears 7 times, gloomy appears 3 times, and assassination only appears once.

- **3b.** How many words did Shakespeare use just once? Twice? At least 10 times? More than 100 times?

```
length(shakespeare.wordtab[shakespeare.wordtab==1])
```

```
## [1] 41843
```

```
length(shakespeare.wordtab[shakespeare.wordtab>=10])
```

```
## [1] 8187
```

```
length(shakespeare.wordtab[shakespeare.wordtab>100])
```

```
## [1] 975
```

Shakespeare used 41,843 words once, 8,187 at least ten times, and 975 words at least one hundred times.

- **3c.** Sort `shakespeare.wordtab` so that its entries (counts) are in decreasing order, and save the result as `shakespeare.wordtab.sorted`. Print the 25 most commonly used words, along with their counts. What is the most common word? Second and third most common words?

```
shakespeare.wordtab.sorted <- sort(shakespeare.wordtab, decreasing=TRUE)
shakespeare.wordtab.sorted[1:25]
```

```
## shakespeare.words
##          the      I    and     to     of      a     my     in    you     is
## 411073 25378  20629  19806  16966  16718  13657  11443  10519   9591   8335
##   that    And    not   with    his     be   your    for   have     it   this
##   8150   7769   7415   7380   6851   6411   6386   6014   5584   5242   5190
##     me     he     as
##   5107   5009   4584
```

The three most common words and the, I, and and.

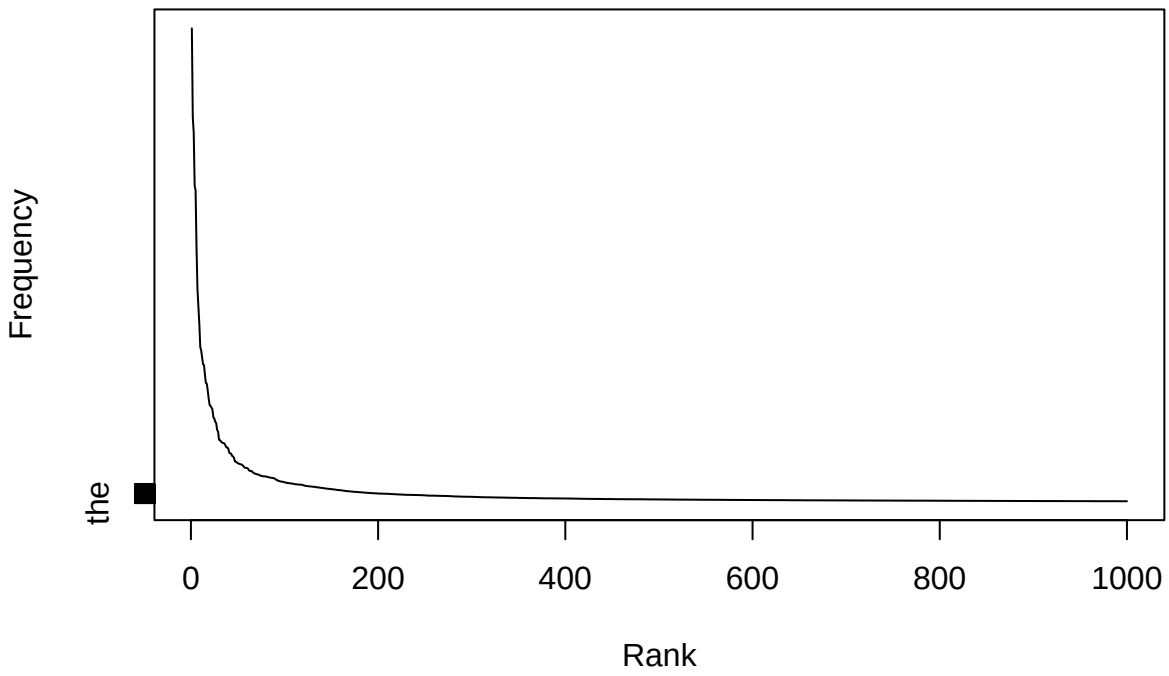
- **3d.** What you should have seen in the last question is that the most common word is the empty string "". This is just an artifact of splitting `shakespeare.all` by spaces, using `strsplit()`. Redefine `shakespeare.words` so that all empty strings are deleted from this vector. Then recompute `shakespeare.wordtab` and `shakespeare.wordtab.sorted`. Check that you have done this right by printing out the new 25 most commonly used words, and verifying (just visually) that is overlaps with your solution to the last question.

```
shakespeare.words <- shakespeare.words[shakespeare.words != ""]
shakespeare.wordtab <- table(shakespeare.words)
shakespeare.wordtab.sorted <- sort(shakespeare.wordtab, decreasing=TRUE)
shakespeare.wordtab.sorted[1:25]
```

```
## shakespeare.words
##   the     I   and    to    of     a    my    in   you    is  that   And   not
## 25378 20629 19806 16966 16718 13657 11443 10519  9591  8335  8150  7769  7415
##  with   his    be  your   for  have    it  this    me    he    as  thou
##  7380  6851  6411  6386  6014  5584  5242  5190  5107  5009  4584  4522
```

- **3e.** As done at the end of the lecture notes, produce a plot of the word counts (y-axis) versus the ranks (x-axis) in `shakespeare.wordtab.sorted`. Set `xlim=c(1,1000)` as an argument to `plot()`; this restricts the plotting window to just the first 1000 ranks, which is helpful here to see the trend more clearly. Do you see **Zipf's law** in action, i.e., does it appear that Frequency $\approx C(1/\text{Rank})^a$ (for some C, a)? **Challenge**: either programmatically, or manually, determine reasonably-well-fitting values of C, a for the Shakespeare data set; then draw the curve $y = C(1/x)^a$ on top of your plot as a red line to show how well it fits.

```
plot(1:1000, shakespeare.wordtab.sorted[1:1000], type="l",
     xlab="Rank", ylab="Frequency")
```



A tiny bit of regular expressions

- **4a.** There are a couple of issues with the way we've built our words in `shakespeare.words`. The first is that capitalization matters; from Q3c, you should have seen that "and" and "And" are counted as separate words. The second is that many words contain punctuation marks (and so, aren't really words in the first place); to see this, retrieve the count corresponding to "and," in your word table `shakespeare.wordtab`.

  The fix for the first issue is to convert `shakespeare.all` to all lower case characters. Hint: recall `tolower()` from Q1b. The fix for the second issue is to use the argument `split="[[:space:]]|[[:punct:]]"` in the call to `strsplit()`, when defining the words. In words, this means: *split on spaces or on punctuation marks* (more precisely, it uses what we call a **regular expression** for the `split` argument). Carry out both of these fixes to define new words `shakespeare.words.new`. Then, delete all empty strings from this vector, and compute word table from it, called `shakespeare.wordtab.new`.

```
shakespeare.words.new <- tolower(shakespeare.all)
shakespeare.words.new <- strsplit(shakespeare.words.new, split="[[:space:]]|[[:punct:]]")
shakespeare.words.new <- unlist(shakespeare.words.new)
shakespeare.words.new <- shakespeare.words.new[shakespeare.words.new != ""]
shakespeare.wordtab.new <- table(shakespeare.words.new)
```

- **4b.** Compare the length of `shakespeare.words.new` to that of `shakespeare.words`; also compare the length of `shakespeare.wordtab.new` to that of `shakespeare.wordtab`. Explain what you are observing.

```
length(shakespeare.words.new)
```

```
## [1] 985473
```

```
length(shakespeare.words)
```

```
## [1] 959302
```

```
length(shakespeare.wordtab.new)
```

```
## [1] 28365
```

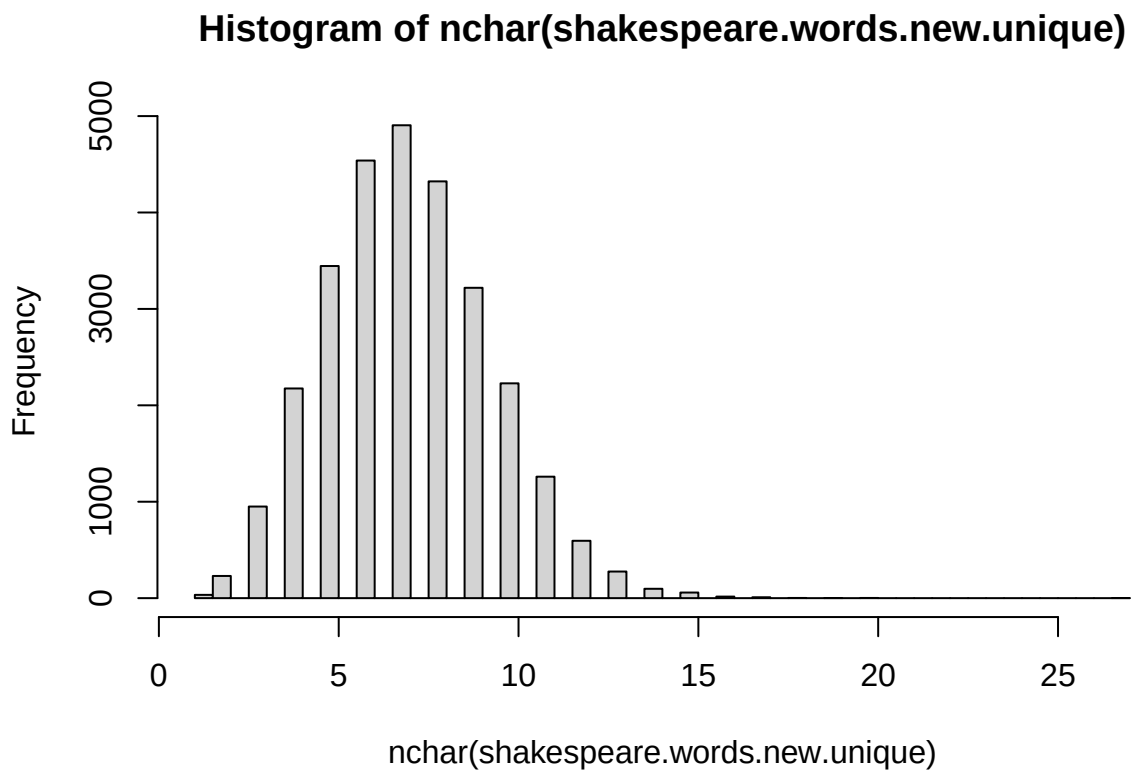
```
length(shakespeare.wordtab)
```

```
## [1] 76171
```

Lowercasing the data and doing the improved split resulted in far fewer words in the table, but the overall length of the data is still greater somehow.

- **4c.** Compute the unique words in `shakespeare.words.new`, calling the result `shakespeare.words.new.unique`. Then repeat the queries in Q2e and Q2f on `shakespeare.words.new.unique`. Comment on the histogram—is it different in any way than before? How about the top 5 longest words?

```
shakespeare.words.new.unique <- unique(shakespeare.words.new)
hist(nchar(shakespeare.words.new.unique), breaks = 50)
```




```
order(nchar(shakespeare.words.new.unique), decreasing=TRUE)[1:10]
```

```
##  [1] 19304 18376 24712 18764 11428 11474 12132 12544 18768 18855
```

```
shakespeare.words.new.unique[order(nchar(shakespeare.words.new.unique), decreasing=TRUE)][1:10]
```

```
##  [1] "honorificabilitudinitatibus" "foundâ\200"dispatchâ\200\231d"
##  [3] "thâ\200\231uncomprehensive"        "samphireâ\200"dreadful"
##  [5] "â\200\230gentlemanâ\200\231â\200""          "thâ\200\231understanding"
##  [7] "thâ\200\231offenderâ\200\231s"           "cozenageâ\200"isâ\200\231t"
##  [9] "thâ\200\231unnumberâ\200\231d"           "indistinguishâ\200\231d"
```

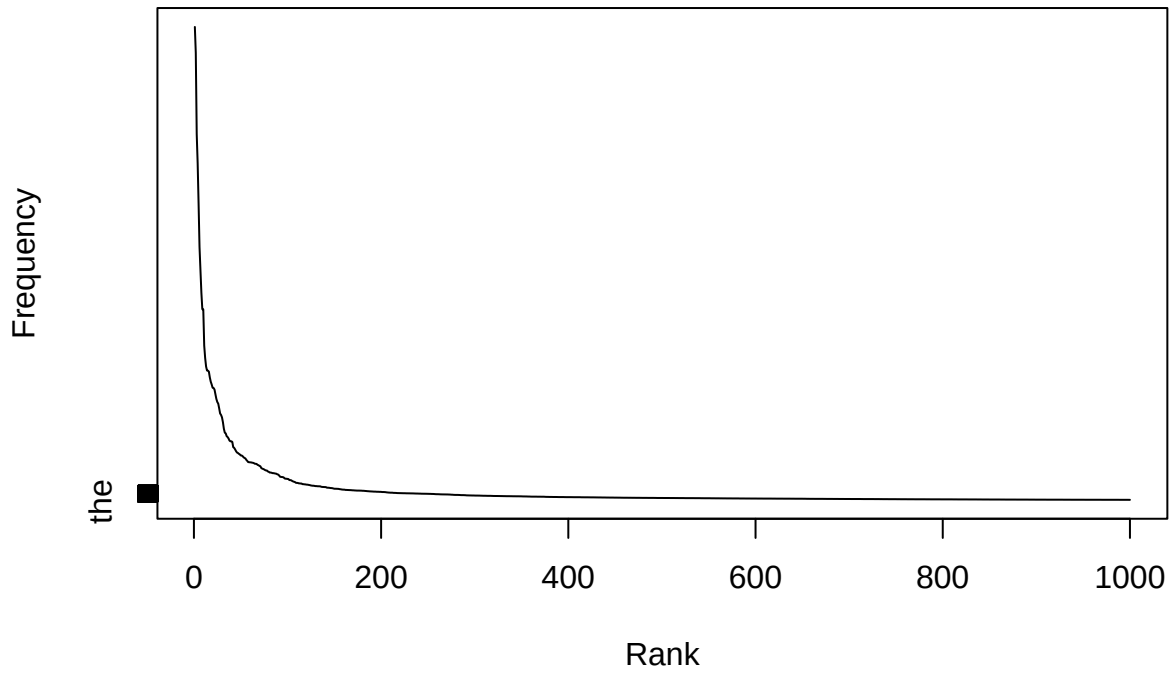
The histogram has a much smaller tail than before. Honorificabilitudinitatibus is still the longest word by far, and many of the other noisy words dropped out of the top.

- **4d.** Sort `shakespeare.wordtab.new` so that its entries (counts) are in decreasing order, and save the result as `shakespeare.wordtab.sorted.new`. Print out the 25 most common words and their counts, and compare them (informally) to what you saw in Q3d. Also, produce a plot of the new word counts, as you did in Q3e. Does Zipf's law look like it still holds?

```
shakespeare.wordtab.new.sorted <- sort(shakespeare.wordtab.new, decreasing=TRUE)
shakespeare.wordtab.new.sorted[1:25]
```

```
## shakespeare.words.new
##   the   and     i    to    of     a   you    my    in  that    is   not  with
## 30002 28362 23225 21395 18833 16087 14627 13198 12158 12146  9859  9079  8537
##    me   for    it     d   his    be    he  this  your   but     s  have
##  8280  8273  8210  7848  7575  7401  7188  7179  7076  6776  6483  6291
```

```
plot(1:1000, shakespeare.wordtab.new.sorted[1:1000], type="l",
     xlab="Rank", ylab="Frequency")
```



The top ranking words are fairly similar to the ones from before. Zipf's law still appears to hold.

Where are Shakespeare's plays, in this massive text?

- **5a.** Let's go back to `shakespeare.lines`. Take a look at lines 19 through 23 of this vector: you should see a bunch of spaces preceding the text in lines 21, 22, and 23. Redefine `shakespeare.lines` by setting it equal to the output of calling the function `trimws()` on `shakespeare.lines`. Print out lines 19 through 23 again, and describe what's happened.

```
shakespeare.lines[19:23]
```

```
## [1] "SHAKESPEARE ***"
## [2] "The Complete Works of William Shakespeare"
## [3] "by William Shakespeare"
## [4] "      Contents"
## [5] "                THE SONNETS"
```

```
shakespeare.lines <- trimws(shakespeare.lines)
shakespeare.lines[19:23]
```

```
## [1] "SHAKESPEARE ***"
## [2] "The Complete Works of William Shakespeare"
## [3] "by William Shakespeare"
## [4] "Contents"
## [5] "THE SONNETS"
```

This function trimmed the whitespace (indentations) in the lines.

- **5b.** Visit http://www.stat.cmu.edu/~ryantibs/statcomp/data/shakespeare.txt in your web browser and just skim through this text file. Near the top you'll see a table of contents. Note that "THE SONNETS" is the first play, and "VENUS AND ADONIS" is the last. Using `which()`, find the indices of the lines in `shakespeare.lines` that equal "THE SONNETS", report the index of the *first* such occurence, and store it as `toc.start`. Similarly, find the indices of the lines in `shakespeare.lines` that equal "VENUS AND ADONIS", report the index of the *first* such occurence, and store it as `toc.end`.

```
toc.start <- which(shakespeare.lines == "THE SONNETS")[1]
toc.end <- which(shakespeare.lines == "VENUS AND ADONIS")[1]
```

- **5c.** Define `n = toc.end - toc.start + 1`, and create an empty string vector of length `n` called `titles`. Using a `for()` loop, populate `titles` with the titles of Shakespeare's plays as ordered in the table of contents list, with the first being "THE SONNETS", and the last being "VENUS AND ADONIS". Print out the resulting `titles` vector to the console. Hint: if you define the counter variable `i` in your `for()` loop to run between 1 and `n`, then you will have to index `shakespeare.lines` carefully to extract the correct titles. Think about the following. When `i=1`, you want to extract the title of the first play in `shakespeare.lines`, which is located at index `toc.start`. When `i=2`, you want to extract the title of the second play, which is located at index `toc.start + 1`. And so on.

```
n <- toc.end - toc.start + 1
titles <- vector(length = n)
for(i in 1:n){
  titles[i] <- shakespeare.lines[toc.start+ i-1]
}
```

- **5d.** Use a `for()` loop to find out, for each play, the index of the line in `shakespeare.lines` at which this play begins. It turns out that the *second* occurence of "THE SONNETS" in `shakespeare.lines` is where this play actually begins (this first ocurrence is in the table of contents), and so on, for

each play title. Use your `for()` loop to fill out an integer vector called `titles.start`, containing the indices at which each of Shakespeare's plays begins in `shakespeare.lines`. Print the resulting vector `titles.start` to the console.

```
titles.start <- vector(length = n)
for(i in 1:n){
  titles.start[i] <- which(shakespeare.lines == titles[i])[2]
}
```

- **5e.** Define `titles.end` to be an integer vector of the same length as `titles.start`, whose first element is the second element in `titles.start` minus 1, whose second element is the third element in `titles.start` minus 1, and so on. What this means: we are considering the line before the second play begins to be the last line of the first play, and so on. Define the last element in `titles.end` to be the length of `shakespeare.lines`. You can solve this question either with a `for()` loop, or with proper indexing and vectorization. **Challenge**: it's not really correct to set the last element in `titles.end` to be length of `shakespeare.lines`, because there is a footer at the end of the Shakespeare data file. By looking at the data file visually in your web browser, come up with a way to programmatically determine the index of the last line of the last play, and implement it.

```
titles.end <- vector(length = n)
for(i in 1:(n-1)){
  titles.end[i] <- titles.start[i+1]-1
}
titles.end[n] <- length(shakespeare.lines)
```

- **5f.** In Q5d, you should have seen that the starting index of Shakespeare's 38th play "THE TWO NOBLE KINSMEN" was computed to be `NA`, in the vector `titles.start`. Why? If you run `which(shakespeare.lines == "THE TWO NOBLE KINSMEN")` in your console, you will see that there is only one occurence of "THE TWO NOBLE KINSMEN" in `shakespeare.lines`, and this occurs in the table of contents. So there was no second occurence, hence the resulting `NA` value.

  But now take a look at line 118,463 in `shakespeare.lines`: you will see that it is "THE TWO NOBLE KINSMEN:", so this is really where the second play starts, but because of colon ":" at the end of the string, this doesn't exactly match the title "THE TWO NOBLE KINSMEN", as we were looking for. The advantage of using the `grep()` function, versus checking for exact equality of strings, is that `grep()` allows us to match substrings. Specifically, `grep()` returns the indices of the strings in a vector for which a substring match occurs, e.g.,

```
grep(pattern="cat",
     x=c("cat", "canned goods", "batman", "catastrophe", "tomcat"))
```

```
## [1] 1 4 5
```

```
so we can see that in this example, `grep()` was able to find substring matches to "cat" in the first, 
```

```
titles.start <- vector(length = n)
for(i in 1:n){
  titles.start[i] <-  grep(pattern=titles[i], x=shakespeare.lines)[2]
}

titles.end <- vector(length = n)
for(i in 1:(n-1)){
  titles.end[i] <- titles.start[i+1]-1
}
titles.end[n] <- length(shakespeare.lines)

titles.start
```

```
##  [1]     67   2378   5311   9142  11773  13703  17591  21386  26645  30390
## [11]  33615  36903  39958  43249  46413  49896  52681  55428  60108  62924
## [21]  65463  68320  71021  73767  75997  79470  83084  86328  89287  93443
## [31]  97536 101206 103641 106199 108939 113683 116176 118464 122683 126021
## [41] 126352 126557 126627 128535
```

```
titles.end
```

```
##  [1]   2377   5310   9141  11772  13702  17590  21385  26644  30389  33614
## [11]  36902  39957  43248  46412  49895  52680  55427  60107  62923  65462
## [21]  68319  71020  73766  75996  79469  83083  86327  89286  93442  97535
## [31] 101205 103640 106198 108938 113682 116175 118463 122682 126020 126351
## [41] 126556 126626 128534 130095
```

Extracting and analysing a couple of plays

- **6a.** Let's look at two of Shakespeare's most famous tragedies. Programmatically find the index at which "THE TRAGEDY OF HAMLET, PRINCE OF DENMARK" occurs in the `titles` vector. Use this to find the indices at which this play starts and ends, in the `titles.start` and `titles.end` vectors, respectively. Call the lines of text corresponding to this play `shakespeare.lines.hamlet`. How many such lines are there? Do the same, but now for the play "THE TRAGEDY OF ROMEO AND JULIET", and call the lines of text corresponding to this play `shakespeare.lines.romeo`. How many such lines are there?

```
n <- which(titles == "THE TRAGEDY OF HAMLET, PRINCE OF DENMARK")
start <- titles.start[n]
end <- titles.end[n]
shakespeare.lines.hamlet <- shakespeare.lines[start:end]
length(shakespeare.lines.hamlet)
```

```
## [1] 5259
```

```
n <- which(titles == "THE TRAGEDY OF ROMEO AND JULIET")
start <- titles.start[n]
end <- titles.end[n]
shakespeare.lines.romeo <- shakespeare.lines[start:end]
length(shakespeare.lines.romeo)
```

```
## [1] 4093
```

Hamlet is 5,429 lines long, and Romeo & Juliet is 4,093 lines long.

- **6b.** Repeat the analysis, outlined in Q4, on `shakespeare.lines.hamlet`. (This should mostly just involve copying and pasting code as needed.) That is, to be clear: * collapse `shakespeare.lines.hamlet` into one big string, separated by spaces; * convert this string into all lower case characters; * divide this string into words, by splitting on spaces or on punctuation marks, using `split="[[:space:]]|[[:punct:]]"` in the call to `strsplit()`; * remove all empty words (equal to the empty string `""`), and report how many words remain; * compute the unique words, report the number of unique words, and plot a histogram of their numbers of characters; * report the 5 longest words; * compute a word table, and report the 25 most common words and their counts; * finally, produce a plot of the word counts verus rank.

```
shakespeare.lines.hamlet <- paste(shakespeare.lines.hamlet, collapse=" ")
shakespeare.lines.hamlet <- tolower(shakespeare.lines.hamlet)
shakespeare.lines.hamlet <- strsplit(shakespeare.lines.hamlet, split="[[:space:]]|[[:punct:]]")
shakespeare.lines.hamlet <- unlist(shakespeare.lines.hamlet)
```

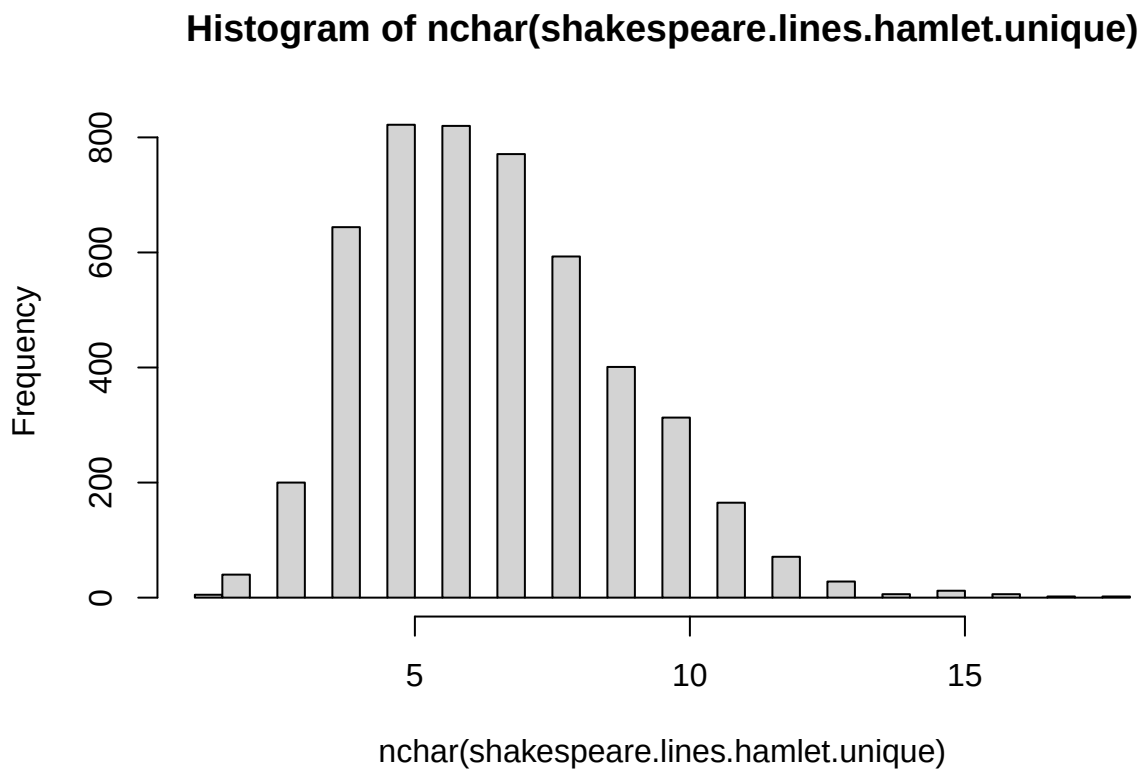
```
shakespeare.lines.hamlet <- shakespeare.lines.hamlet[shakespeare.lines.hamlet != ""]
length(shakespeare.lines.hamlet)
```

```
## [1] 32193
```

```
shakespeare.lines.hamlet.unique <- unique(shakespeare.lines.hamlet)
length(shakespeare.lines.hamlet.unique)
```

```
## [1] 4901
```

```
hist(nchar(shakespeare.lines.hamlet.unique), breaks = 50)
```




```
shakespeare.lines.hamlet.unique[order(nchar(shakespeare.lines.hamlet.unique),
                                      decreasing = TRUE)][1:5]
```

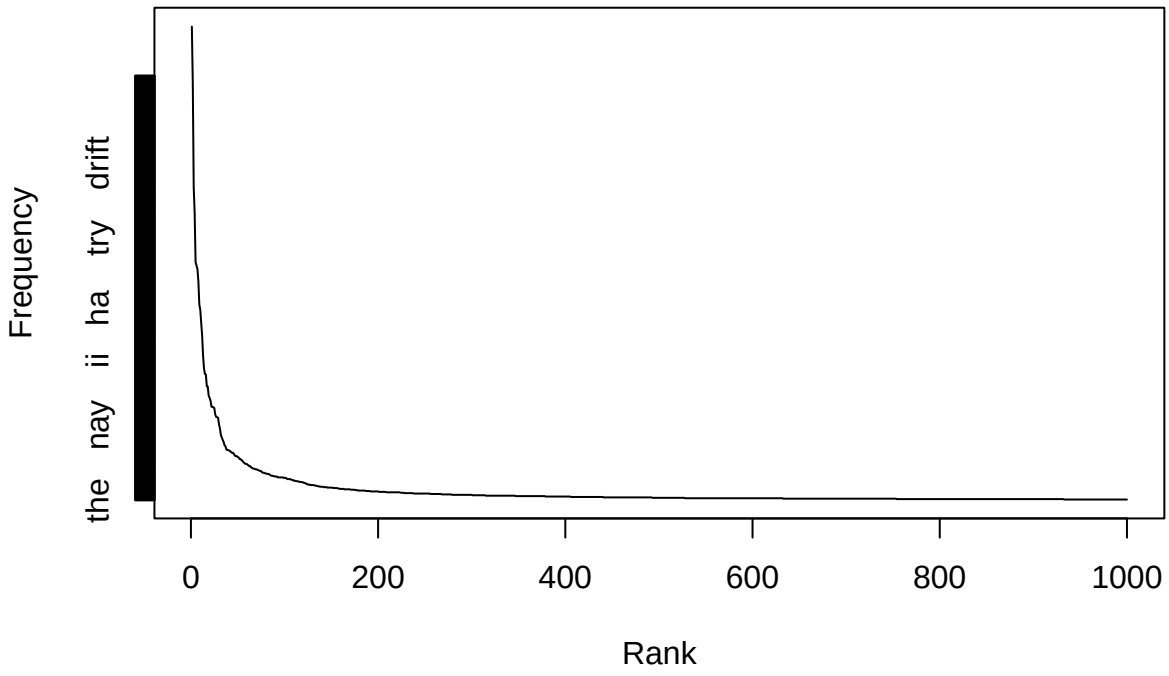
```
## [1] "â\200\230gentlemanâ\200\231â\200"" "thâ\200\231understanding" "thâ\200\231offenderâ\200\231s"
## [4] "cozenageâ\200"isâ\200\231t"  "thâ\200\231extravagant"
```

```
shakespeare.lines.hamlet.wordtab <- table(shakespeare.lines.hamlet)
shakespeare.lines.hamlet.wordtab.sorted <- sort(shakespeare.lines.hamlet.wordtab, decreasing = TRUE)
shakespeare.lines.hamlet.wordtab.sorted[1:25]
```

```
## shakespeare.lines.hamlet
##    the    and     to     of      a    you      i     my hamlet     in     it
##   1115    982    737    677    561    553    545    516    462    449    418
##   that     is    not   this    his    but   with    for   your     me     be
##    388    341    311    298    297    269    269    247    242    235    221
##   lord     as     he
```

```
##     221     220     218
```

```
plot(1:1000, shakespeare.lines.hamlet.wordtab.sorted[1:1000], type="l",
     xlab="Rank", ylab="Frequency")
```



- **6c.** Repeat the same task as in the last part, but on `shakespeare.lines.romeo`. (Again, this should just involve copying and pasting code as needed. P.S. Isn't this getting tiresome? You'll be happy when we learn functions, next week!) Comment on any similarities/differences you see in the answers.

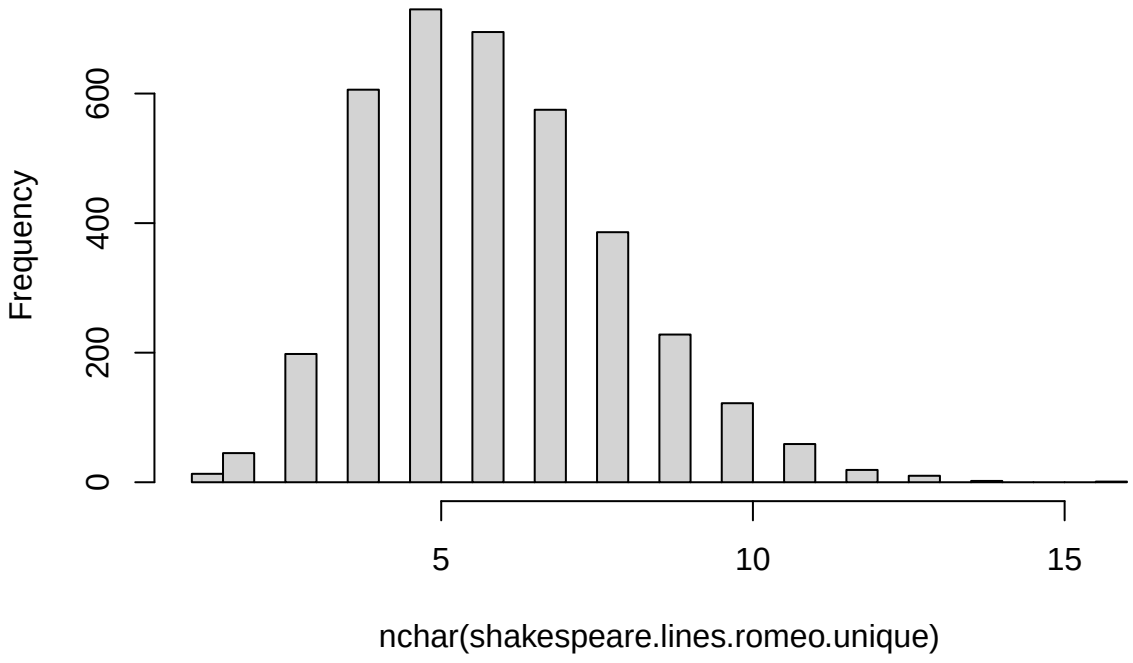
```
shakespeare.lines.romeo <- paste(shakespeare.lines.romeo, collapse=" ")
shakespeare.lines.romeo <- tolower(shakespeare.lines.romeo)
shakespeare.lines.romeo <- strsplit(shakespeare.lines.romeo, split="[[:space:]]|[[:punct:]]")
shakespeare.lines.romeo <- unlist(shakespeare.lines.romeo)
shakespeare.lines.romeo <- shakespeare.lines.romeo[shakespeare.lines.romeo != ""]
length(shakespeare.lines.romeo)
```

```
## [1] 26901
```

```
shakespeare.lines.romeo.unique <- unique(shakespeare.lines.romeo)
length(shakespeare.lines.romeo.unique)
```

```
## [1] 3689
```

```
hist(nchar(shakespeare.lines.romeo.unique), breaks = 50)
```



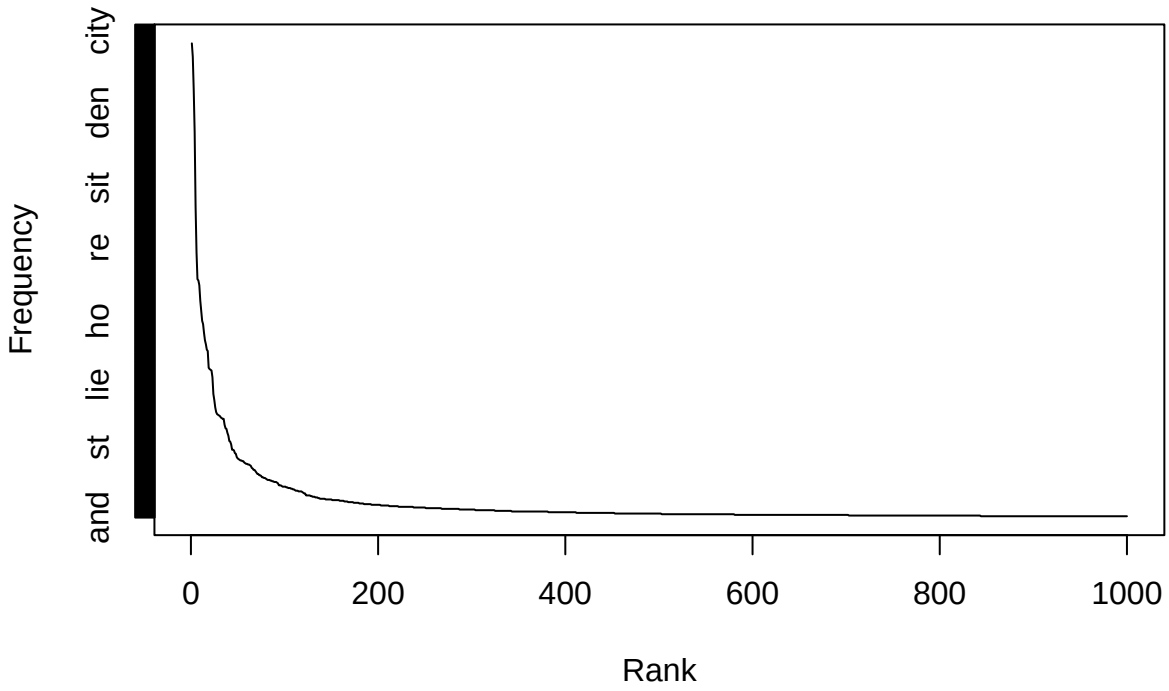

```
shakespeare.lines.romeo.unique[order(nchar(shakespeare.lines.romeo.unique),
                                     decreasing = TRUE)][1:5]
```

```
## [1] "reverenceâ\200"love" "distemperature"   "unthankfulness"   "interchanging"
## [5] "â\200"underneath"
```

```
shakespeare.lines.romeo.wordtab <- table(shakespeare.lines.romeo)
shakespeare.lines.romeo.wordtab.sorted <- sort(shakespeare.lines.romeo.wordtab, decreasing = TRUE)
shakespeare.lines.romeo.wordtab.sorted[1:25]
```

```
## shakespeare.lines.romeo
##    and    the      i     to      a     of     my   that     is     in  romeo
##    707    687    638    575    464    396    356    354    347    324    309
##    you      s   thou     me    not   with      d     it    for   this    â\200"
##    294    289    276    265    259    251    249    224    222    221    220
##     be juliet    but
##    210    185    176
```

```
plot(1:1000, shakespeare.lines.romeo.wordtab.sorted[1:1000], type="l",
     xlab="Rank", ylab="Frequency")
```



- **Challenge.** Using a `for()` loop and the `titles.start`, `titles.end` vectors constructed above, answer the following questions. What is Shakespeare's longest play (in terms of the number of words)? What is Shakespeare's shortest play? In which play did Shakespeare use his longest word (in terms of the number of characters)? Are there any plays in which "the" is not the most common word?