HW 7

SDS322E

October 12, 2022

Ethan Chang - ehc586

Please submit as a PDF or HTML file on Canvas before the due date.

For all questions, include the R commands/functions that you used to find your answer. Answers without supporting code will not receive credit.

> Review of how to submit this assignment
>
> All homework assignments will be completed using R Markdown. These `.Rmd` files consist of >text/syntax (formatted using Markdown) alongside embedded R code. When you have completed the assignment (by adding R code inside codeblocks and supporting text outside of the codeblocks), create your document as follows (assuming you are using the edupod server and submitting HTML):

> - Click the arrow next to the “Knit” button (above)
> - Choose “Knit to HTML”
> - Go to Files pane and put checkmark next to the correct HTML file
> - Click on the blue gear icon (“More”) and click Export
> - Download the file and then upload to Canvas
> - To submit a PDF, open your HTML file and print it to a pdf, then upload the pdf as your submission.

Regex and Web Scraping

In this homework we will practice our web scraping and regex skills! Recall that a regex is a special kind of string that is used for pattern matching, where certain characters in the string have a special meaning. We saw, for example, that the `.` character in a string, when interpreted as a regex, matches not just a literal period, but *any* character; so, for example:

```
my_regex <- ".og"
my_strings <- c("dog", "hog", "frog", "fr.og", "pizza")
str_view(my_strings, my_regex)
```

dog

hog

frog

fr.og

pizza

but if we want to match the literal period we need to use the regex `\.og` :

```
my_other_regex <- "\\.og"
writeLines(my_other_regex)
```

```
## \.og
```

```
str_view(my_strings, my_other_regex)
```

dog

hog

frog

fr.og

pizza

We will use this homework as an opportunity to practice some other regex patterns that are useful for routine string processing.

For a quick reference on regex and string manipulations in general, see this cheat sheet (https://raw.githubusercontent.com/rstudio/cheatsheets/main/strings.pdf).

Question 1

We can specify a *range* of values in a few different ways. One is to use square brackets: for example, `[a-z]` matches all lower case characters, `[0-9]` matches all digits, and `[a-zA-z]` matches all letters. We can also check for *one or more consecutive occurrences* of some pattern using the `+` symbol; for example, the regex `\.+` checks for one or more consecutive periods.

Now, **write a regex that matches one of more adjacent numbers/digits, and verify that it works on `string1` below using `str_view_all`. Your matches should include `12`, `47`, `7890`, etc.**

```
string1 <- "We have to extract these numbers 12, 47, 48 The integers numbers are also interesting: 189 2036 314 \',\' is a separator, so pleas
e extract these numbers 125,789,1450 and also these 564,90456 We like to offer you 7890$ per month in order to complete this task... we are jo
king"

writeLines(string1)
```

```
## We have to extract these numbers 12, 47, 48 The integers numbers are also interesting: 189 2036 314 ',' is a separator, so please extract t
hese numbers 125,789,1450 and also these 564,90456 We like to offer you 7890$ per month in order to complete this task... we are joking
```

```
## This is an example of the + regex. Note the double escape!
writeLines("\\.+")
```

```
## \.+
```

```
str_view_all(string1, "\\.+")
```

We have to extract these numbers 12, 47, 48 The integers numbers are also interesting: 189 2036 314 ',' is a separator, so please extract these numbers 125,789,1450 and also these 564,90456 We like to offer you 7890$ per month in order to complete this task... we are joking

```
# Your code here
q1_regex <- "[0-9]+"
str_view_all(string1, q1_regex)
```

We have to extract these numbers 12, 47, 48 The integers numbers are also interesting: 189 2036 314 ',' is a separator, so please extract these numbers 125,789,1450 and also these 564,90456 We like to offer you 7890$ per month in order to complete this task... we are joking

Question 2

For `string2` below, use `str_view_all()` with a regular expression to match all of the IP addresses. *Your regex does not need to work for all possible IPs, or exclude all invalid IPs, it just need to detect exactly the IPs in the string below.* Your matches should include, for example, `213.92.153.167` or `69.43.107.219`.

```
string2 <-
'Jan 13 00:48:59: DROP service 68->67(udp) from 213.92.153.167 to 69.43.107.219, prefix: "spoof iana-0/8" (in: eth0 69.43.112.233(38:f8:b7:90:
45:92):68 -> 217.70.100.113(00:21:87:79:9c:d9):67 UDP len:576 ttl:64)
Jan 13 12:02:48: ACCEPT service dns from 74.125.186.208 to firewall(pub-nic-dns), prefix: "none"  (in: eth0 74.125.186.208(00:1a:e3:52:5d:8e):
36008 -> 140.105.63.158(00:1a:9a:86:2e:62):53 UDP len:82 ttl:38)

Jan 13 17:44:52: DROP service 68->67(udp) from 172.45.240.237 to 217.70.177.60, prefix: "spoof iana-0/8" (in: eth0 216.34.90.16(00:21:91:fe:a
2:6f):68 -> 69.43.85.253(00:07:e1:7c:53:db):67 UDP len:328 ttl:64)

Jan 13 17:52:08: ACCEPT service http from 213.121.184.130 to firewall(pub-nic), prefix: "none"(in: eth0 213.121.184.130(00:05:2e:6a:a4:14):850
4 -> 140.105.63.164(00:60:11:92:ed:1b):80 TCP flags: ****S* len:52 ttl:109)'

# Your code here
#q2_regex <- "[0-9]+\\.[0-9]+\\.[0-9]+\\.[0-9]+"
q2_regex <- "([0-9]+\\.)+[0-9]+"
str_view_all(string2, q2_regex)
```

Jan 13 00:48:59: DROP service 68->67(udp) from 213.92.153.167 to 69.43.107.219, prefix: "spoof iana-0/8" (in: eth0 69.43.112.233(38:f8:b7:90:45:92):68 -> 217.70.100.113(00:21:87:79:9c:d9):67 UDP len:576 ttl:64) Jan 13 12:02:48: ACCEPT service dns from 74.125.186.208 to firewall(pub-nic-dns), prefix: "none" (in: eth0 74.125.186.208(00:1a:e3:52:5d:8e):36008 -> 140.105.63.158(00:1a:9a:86:2e:62):53 UDP len:82 ttl:38) Jan 13 17:44:52: DROP service 68->67(udp) from 172.45.240.237 to 217.70.177.60, prefix: "spoof iana-0/8" (in: eth0 216.34.90.16(00:21:91:fe:a2:6f):68 -> 69.43.85.253(00:07:e1:7c:53:db):67 UDP len:328 ttl:64) Jan 13 17:52:08: ACCEPT service http from 213.121.184.130 to firewall(pub-nic), prefix: "none"(in: eth0 213.121.184.130(00:05:2e:6a:a4:14):8504 -> 140.105.63.164(00:60:11:92:ed:1b):80 TCP flags: ****S* len:52 ttl:109)

Hint: a couple of useful regex patterns: - The regex `[0-9]+` detects all sequences of digits. - Wrapping a regex in `()` can be used to make groups; for example, the regex `([0-9]+\.)+` detects all sequences of (sequences of digits followed by a period) (*think about why!*) - Remember to double escape where needed, and you can check the regex that your string represents using `writeLines()`.

Question 3

For `string3` below, use `str_match_all()` (see `?str_match_all`, but this essentially just returns the individual matches in a string) with a regular expression to match sites of the form `ANTANT` where A, C, G, and T are literal, while N represents any base (A, C, G, or T). **How many matches are there?** As before, your regex does not need to be a general solution, it just needs to work with this particular string.

```
string3 <- "ATGGCAATAACCCCCCGTTTCTACTTCTAGAGGAGAAAAGTATTGACATGAGCGCTCCCGGCACAAGGGCCAAAGAAGTCTCCAATTTCTTATTTCCGAATGACATGCGTCTCCTTGCGGGTAAATCACC
GACCGCAATTCATAGAAGCCTGGGGGAACAGATAGGTCTAATTAGCTTAAGAGAGTAAATCCTGGGATCATTCAGTAGTAACCATAAACTTACGCTGGGGCTTCTTCGGCGGATTTTTACAGTTACCAACCAGGAGATTTGA
AGTAAATCAGTTGAGGATTTAGCCGCGCTATCCGGTAATCTCCAAATTAAAACATACCGTTCCATGAAGGCTAGAATTACTTACCGGCCTTTTCCATGCCTGCGCTATACCCCCCCCACTCTCCCGCTTATCCGTCCGAGCGG
AGGCAGTGCGATCCTCCGTTAAGATATTCTTACGTGTGACGTAGCTATGTATTTTGCAGAGCTGGCGAACGCGTTGAACACTTCACAGATGGTAGGGATTCGGGTAAAGGGCGTATAATTGGGGACTAACATAGGCGTAGAC
TACGATGGCGCCAACTCAATCGCAGCTCGAGCGCCCTGAATAACGTACTCATCTCAACTCATTCTCGGCAATCTACCGAGCGACTCGATTATCAACGGCTGTCTAGCAGTTCTAATCTTTTGCCAGCATCGTAATAGCCTCC
AAGAGATTGATGATAGCTATCGGCACAGAACTGAGACGGCGCCGATGGATAGCGGACTTTCGGTCAACCACAATTCCCCACGGGACAGGTCCTGCGGTGCGCATCACTCTGAATGTACAAGCAACCCAAGTGGGCCGAGCCT
GGACTCAGCTGGTTCCTGCGTGAGCTCGAGACTCGGGATGACAGCTCTTTAAACATAGAGCGGGGGCGTCGAACGGTCGAGAAAGTCATAGTACCTCGGGTACCAACTTACTCAGGTTATTGCTTGAAGCTGTACTATTTTA
GGGGGGGAGCGCTGAAGGTCTCTTCTTCTCATGACTGAACTCGCGAGGGTCGTGAAGTCGGTTCCTTCAATGGTTAAAAAACAAAGGCTTACTGTGCGCAGAGGAACGCCCATCTAGCGGCTGGCGTCTTGAATGCTCGGTC
CCCTTTGTCATTCCGGATTAATCCATTTCCCTCATTCACGAGCTTGCGAAGTCTACATTGGTATATGAATGCGACCTAGAAGAGGGCGCTTAAAATTGGCAGTGGTTGATGCTCTAAACTCCATTTGGTTTACTCGTGCATC
ACCGCGATAGGCTGACAAAGGTTTAACATTGAATAGCAAGGCACTTCCGGTCTCAATGAACGGCCGGGAAAGGTACGCGCGCGGTATGGGAGGATCAAGGGGCCAATAGAGAGGCTCCTCTCTCACTCGCTAGGAGGCAAAT
GTAAAACAATGGTTACTGCATCGATACATAAAACATGTCCATCGGTTGCCCAAAGTGTTAAGTGTCTATCACCCCTAGGGCCGTTTCCCGCATATAAACGCCAGGTTGTATCCGCATTTGATGCTACCGTGGATGAGTCTGC
GTCGAGCGCGCCGCACGAATGTTGCAATGTATTGCATGAGTAGGGTTGACTAAGAGCCGTTAGATGCGTCGCTGTACTAATAGTTGTCGACAGACCGTCGAGATTAGAAAATGGTACCAGCATTTTCGGAGGTTCTCTAACT
AGTATGGATTGCGGTGTCTTCACTGTGCTGCGGCTACCCATCGCCTGAAATCCAGCTGGTGTCAAGCCATCCCCTCTCCGGGACGCCGCATGTAGTGAAACATATACGTTGCACGGGTTCACCGCGGTCCGTTCTGAGTCGA
CCAAGGACACAATCGAGCTCCGATCCGTACCCTCGACAAACTTGTACCCGACCCCCGGAGCTTGCCAGCTCCTCGGGTATCATGGAGCCTGTGGTTCATCGCGTCCGATATCAAACTTCGTCATGATAAAGTCCCCCCCTCG
GGAGTACCAGAGAAGATGACTACTGAGTTGTGCGAT"
q3_regex <- "A[ACGT]TA[ACGT]T"
str_view_all(string3, q3_regex)
```

ATGGCAATAACCCCCCGTTTCTACTTCTAGAGGAGAAAAGTATTGACATGAGCGCTCCCGGCACAAGGGCCAAAGAAGTCTCCAATTTCTTATTTCCGAATGACATGCGTCTCCTTGCGGGTAAATCACCGACCGCAATT

```
str_match_all(string3, q3_regex)
```

```
## [[1]]
##      [,1]
## [1,] "AGTATT"
## [2,] "AGTAGT"
## [3,] "ATTACT"
## [4,] "ACTATT"
## [5,] "ATTAAT"
## [6,] "ACTAAT"
## [7,] "ACTAGT"
## [8,] "ACTACT"
```

Answer: There are 8 matches for this specified pattern.

Question 4

Let's scrape the text of the first chapter of Moby Dick from an ebook website

```
url <- "https://standardebooks.org/ebooks/herman-melville/moby-dick/text/chapter-1"
moby_html <- read_html(url)
moby_html
```

```
## {html_document}
## <html xmlns="http://www.w3.org/1999/xhtml" xmlns:epub="http://www.idpf.org/2007/ops" epub:prefix="z3998: http://www.daisy.org/z3998/2012/vo
cab/structure/, se: https://standardebooks.org/vocab/1.0" xml:lang="en-US" lang="en-US">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
## [2] <body epub:type="bodymatter z3998:fiction">\n<header><nav><ul>\n<li><a hr ...
```

The text `<p>My Text Here</p>` in an HTML file represents a *paragraph* (`p` for paragraph). Using `moby_ps <- html_nodes(moby_html, "p")` we can grab all of the paragraph nodes from `moby_html` , and then print using `print(moby_ps)` .

How many paragraphs appear in `moby_ps` , and what do they represent?

```
## your code here
moby_ps <- html_nodes(moby_html, "p")
print(moby_ps)
```

```
## {xml_nodeset (17)}
##  [1] <p>Call me Ishmael. Some years ago—never mind how long precisely—havin ...
##  [2] <p>There now is your insular city of the Manhattoes, belted round by wha ...
##  [3] <p>Circumambulate the city of a dreamy Sabbath afternoon. Go from Corlea ...
##  [4] <p>But look! here come more crowds, pacing straight for the water, and s ...
##  [5] <p>Once more. Say you are in the country; in some high land of lakes. Ta ...
##  [6] <p>But here is an artist. He desires to paint you the dreamiest, shadies ...
##  [7] <p>Now, when I say that I am in the habit of going to sea whenever I beg ...
##  [8] <p>No, when I go to sea, I go as a simple sailor, right before the mast, ...
##  [9] <p>What of it, if some old hunks of a sea-captain orders me to get a bro ...
## [10] <p>Again, I always go to sea as a sailor, because they make a point of p ...
## [11] <p>Finally, I always go to sea as a sailor, because of the wholesome exe ...
## [12] <p>\n\t\t\t\t\t\t<strong>“Grand Contested Election for the Presidency of t ...
## [13] <p>“Whaling voyage by one Ishmael.</p>
## [14] <p>\n\t\t\t\t\t<em>“Bloody battle in Afghanistan.”</em>\n\t\t\t\t</p>
## [15] <p>Though I cannot tell why it was exactly that those stage managers, th ...
## [16] <p>Chief among these motives was the overwhelming idea of the great whal ...
## [17] <p>By reason of these things, then, the whaling voyage was welcome; the  ...
```

Answer: 17 paragraphs appear in moby_ps; they represent all of the paragraphs in text that appear in the first chapter of Moby Dick from the ebook website.

Question 5

After getting the paragraphs of the first chapter in the previous question, get the corresponding text with `html_text()` . Then, **do the following:** - collapse all paragraphs into a single character vector using `str_flatten()` , separating the paragraphs with a two newlines (represented by `\n\n` in an `R` string). - Save the result as `moby_text` . - Use `writeLines()` to make sure that things work as intended (i.e., does it appear to just print the first chapter?).

```
moby_text <- html_text(moby_ps) %>% str_flatten("\n\n")
writeLines(moby_text)
```

```
## Call me Ishmael. Some years ago—never mind how long precisely—having little or no money in my purse, and nothing particular to interest me
on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and regulat
ing the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find
myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get
such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and methodically
knocking people’s hats off—then, I account it high time to get to sea as soon as I can. This is my substitute for pistol and ball. With a phil
osophical flourish Cato throws himself upon his sword; I quietly take to the ship. There is nothing surprising in this. If they but knew it, a
lmost all men in their degree, some time or other, cherish very nearly the same feelings towards the ocean with me.
##
## There now is your insular city of the Manhattoes, belted round by wharves as Indian isles by coral reefs—commerce surrounds it with her sur
f. Right and left, the streets take you waterward. Its extreme downtown is the battery, where that noble mole is washed by waves, and cooled b
y breezes, which a few hours previous were out of sight of land. Look at the crowds of water-gazers there.
##
## Circumambulate the city of a dreamy Sabbath afternoon. Go from Corlears Hook to Coenties Slip, and from thence, by Whitehall, northward. Wh
at do you see?—Posted like silent sentinels all around the town, stand thousands upon thousands of mortal men fixed in ocean reveries. Some le
aning against the spiles; some seated upon the pier-heads; some looking over the bulwarks of ships from China; some high aloft in the rigging,
as if striving to get a still better seaward peep. But these are all landsmen; of week days pent up in lath and plaster—tied to counters, nail
ed to benches, clinched to desks. How then is this? Are the green fields gone? What do they here?
##
## But look! here come more crowds, pacing straight for the water, and seemingly bound for a dive. Strange! Nothing will content them but the
extremest limit of the land; loitering under the shady lee of yonder warehouses will not suffice. No. They must get just as nigh the water as
they possibly can without falling in. And there they stand—miles of them—leagues. Inlanders all, they come from lanes and alleys, streets and
avenues—north, east, south, and west. Yet here they all unite. Tell me, does the magnetic virtue of the needles of the compasses of all those
ships attract them thither?
##
## Once more. Say you are in the country; in some high land of lakes. Take almost any path you please, and ten to one it carries you down in a
dale, and leaves you there by a pool in the stream. There is magic in it. Let the most absentminded of men be plunged in his deepest reveries—
stand that man on his legs, set his feet a-going, and he will infallibly lead you to water, if water there be in all that region. Should you e
ver be athirst in the great American desert, try this experiment, if your caravan happen to be supplied with a metaphysical professor. Yes, as
everyone knows, meditation and water are wedded forever.
##
## But here is an artist. He desires to paint you the dreamiest, shadiest, quietest, most enchanting bit of romantic landscape in all the vall
ey of the Saco. What is the chief element he employs? There stand his trees, each with a hollow trunk, as if a hermit and a crucifix were with
in; and here sleeps his meadow, and there sleep his cattle; and up from yonder cottage goes a sleepy smoke. Deep into distant woodlands winds
a mazy way, reaching to overlapping spurs of mountains bathed in their hillside blue. But though the picture lies thus tranced, and though thi
s pine-tree shakes down its sighs like leaves upon this shepherd’s head, yet all were vain, unless the shepherd’s eye were fixed upon the magi
c stream before him. Go visit the Prairies in June, when for scores on scores of miles you wade knee-deep among Tiger-lilies—what is the one c
harm wanting?—Water—there is not a drop of water there! Were Niagara but a cataract of sand, would you travel your thousand miles to see it? W
hy did the poor poet of Tennessee, upon suddenly receiving two handfuls of silver, deliberate whether to buy him a coat, which he sadly neede
d, or invest his money in a pedestrian trip to Rockaway Beach? Why is almost every robust healthy boy with a robust healthy soul in him, at so
me time or other crazy to go to sea? Why upon your first voyage as a passenger, did you yourself feel such a mystical vibration, when first to
ld that you and your ship were now out of sight of land? Why did the old Persians hold the sea holy? Why did the Greeks give it a separate dei
ty, and own brother of Jove? Surely all this is not without meaning. And still deeper the meaning of that story of Narcissus, who because he c
ould not grasp the tormenting, mild image he saw in the fountain, plunged into it and was drowned. But that same image, we ourselves see in al
l rivers and oceans. It is the image of the ungraspable phantom of life; and this is the key to it all.
##
## Now, when I say that I am in the habit of going to sea whenever I begin to grow hazy about the eyes, and begin to be over conscious of my l
ungs, I do not mean to have it inferred that I ever go to sea as a passenger. For to go as a passenger you must needs have a purse, and a purs
e is but a rag unless you have something in it. Besides, passengers get seasick—grow quarrelsome—don’t sleep of nights—do not enjoy themselves
much, as a general thing;—no, I never go as a passenger; nor, though I am something of a salt, do I ever go to sea as a Commodore, or a Captai
n, or a Cook. I abandon the glory and distinction of such offices to those who like them. For my part, I abominate all honorable respectable t
oils, trials, and tribulations of every kind whatsoever. It is quite as much as I can do to take care of myself, without taking care of ships,
barques, brigs, schooners, and whatnot. And as for going as cook—though I confess there is considerable glory in that, a cook being a sort of
officer on shipboard—yet, somehow, I never fancied broiling fowls;—though once broiled, judiciously buttered, and judgmatically salted and pep
pered, there is no one who will speak more respectfully, not to say reverentially, of a broiled fowl than I will. It is out of the idolatrous
dotings of the old Egyptians upon broiled ibis and roasted river horse, that you see the mummies of those creatures in their huge bake-houses
the pyramids.
##
## No, when I go to sea, I go as a simple sailor, right before the mast, plumb down into the forecastle, aloft there to the royal masthead. Tr
ue, they rather order me about some, and make me jump from spar to spar, like a grasshopper in a May meadow. And at first, this sort of thing
is unpleasant enough. It touches one’s sense of honor, particularly if you come of an old established family in the land, the Van Rensselaers,
or Randolphs, or Hardicanutes. And more than all, if just previous to putting your hand into the tar-pot, you have been lording it as a countr
y schoolmaster, making the tallest boys stand in awe of you. The transition is a keen one, I assure you, from a schoolmaster to a sailor, and
requires a strong decoction of Seneca and the Stoics to enable you to grin and bear it. But even this wears off in time.
##
## What of it, if some old hunks of a sea-captain orders me to get a broom and sweep down the decks? What does that indignity amount to, weigh
ed, I mean, in the scales of the New Testament? Do you think the archangel Gabriel thinks anything the less of me, because I promptly and resp
ectfully obey that old hunks in that particular instance? Who ain’t a slave? Tell me that. Well, then, however the old sea-captains may order
me about—however they may thump and punch me about, I have the satisfaction of knowing that it is all right; that everybody else is one way or
other served in much the same way—either in a physical or metaphysical point of view, that is; and so the universal thump is passed round, and
all hands should rub each other’s shoulder-blades, and be content.
##
## Again, I always go to sea as a sailor, because they make a point of paying me for my trouble, whereas they never pay passengers a single pe
nny that I ever heard of. On the contrary, passengers themselves must pay. And there is all the difference in the world between paying and bei
ng paid. The act of paying is perhaps the most uncomfortable infliction that the two orchard thieves entailed upon us. But being paid—what wil
l compare with it? The urbane activity with which a man receives money is really marvellous, considering that we so earnestly believe money to
be the root of all earthly ills, and that on no account can a monied man enter heaven. Ah! how cheerfully we consign ourselves to perdition!
##
```

```
## Finally, I always go to sea as a sailor, because of the wholesome exercise and pure air of the forecastle deck. For as in this world, head
winds are far more prevalent than winds from astern (that is, if you never violate the Pythagorean maxim), so for the most part the Commodore
on the quarterdeck gets his atmosphere at second hand from the sailors on the forecastle. He thinks he breathes it first; but not so. In much
the same way do the commonalty lead their leaders in many other things, at the same time that the leaders little suspect it. But wherefore it
was that after having repeatedly smelt the sea as a merchant sailor, I should now take it into my head to go on a whaling voyage; this the inv
isible police officer of the Fates, who has the constant surveillance of me, and secretly dogs me, and influences me in some unaccountable wa
y—he can better answer than anyone else. And, doubtless, my going on this whaling voyage, formed part of the grand programme of Providence tha
t was drawn up a long time ago. It came in as a sort of brief interlude and solo between more extensive performances. I take it that this part
of the bill must have run something like this:
##
##
##                  “Grand Contested Election for the Presidency of the United States.
##
##
## “Whaling voyage by one Ishmael.
##
##
##                  “Bloody battle in Afghanistan.”
##
##
## Though I cannot tell why it was exactly that those stage managers, the Fates, put me down for this shabby part of a whaling voyage, when ot
hers were set down for magnificent parts in high tragedies, and short and easy parts in genteel comedies, and jolly parts in farces—though I c
annot tell why this was exactly; yet, now that I recall all the circumstances, I think I can see a little into the springs and motives which b
eing cunningly presented to me under various disguises, induced me to set about performing the part I did, besides cajoling me into the delusi
on that it was a choice resulting from my own unbiased free-will and discriminating judgment.
##
## Chief among these motives was the overwhelming idea of the great whale himself. Such a portentous and mysterious monster roused all my curi
osity. Then the wild and distant seas where he rolled his island bulk; the undeliverable, nameless perils of the whale; these, with all the at
tending marvels of a thousand Patagonian sights and sounds, helped to sway me to my wish. With other men, perhaps, such things would not have
been inducements; but as for me, I am tormented with an everlasting itch for things remote. I love to sail forbidden seas, and land on barbaro
us coasts. Not ignoring what is good, I am quick to perceive a horror, and could still be social with it—would they let me—since it is but wel
l to be on friendly terms with all the inmates of the place one lodges in.
##
## By reason of these things, then, the whaling voyage was welcome; the great floodgates of the wonder-world swung open, and in the wild conce
its that swayed me to my purpose, two and two there floated into my inmost soul, endless processions of the whale, and, mid most of them all,
one grand hooded phantom, like a snow hill in the air.
```

Question 6

Our goal now is to extract a data frame consisting of (i) a list of words and (ii) the number of times that word appears in the chapter. To do this, we need to get down to single string that just contains words separated by spaces, without any punctuation.

Create a new object `moby_text_2` containing just the words, with all punctuation stripped out by doing the following:

- Remove all newlines we introduced by replacing all instances of `\n\n` with a space (use `str_replace_all`).
- Replace any emdash `â€”` (note: it's not a hyphen! copy and paste the character in this file if you don't know how to make an emdash symbol on your computer) with a space using `str_replace_all("â€”", " ")`.
- Use a regular expression that removes anything that is *not* an uppercase letter (`[A-Z]`), lowercase letter `[a-z]`, a digit `[0-9]`, a *regular* dash `-` or a space .
- Convert to all lowercase using `str_to_lower()`.

Now, write the resulting string using `writeLines()` and then report the length of the string using `str_length()`.

```
## Your code here
moby_text_2 <- moby_text %>% str_replace_all("\n+", " ") %>% str_replace_all("—", " ") %>% str_remove_all("[^a-zA-Z0-9\\-\\s]") %>% str_remove
_all("\\t+") %>% str_to_lower()
writeLines(moby_text_2)
```

call me ishmael some years ago never mind how long precisely having little or no money in my purse and nothing particular to interest me on shore i thought i would sail about a little and see the watery part of the world it is a way i have of driving off the spleen and regulating t he circulation whenever i find myself growing grim about the mouth whenever it is a damp drizzly november in my soul whenever i find myself in voluntarily pausing before coffin warehouses and bringing up the rear of every funeral i meet and especially whenever my hypos get such an upp er hand of me that it requires a strong moral principle to prevent me from deliberately stepping into the street and methodically knocking peo ples hats off then i account it high time to get to sea as soon as i can this is my substitute for pistol and ball with a philosophical flouri sh cato throws himself upon his sword i quietly take to the ship there is nothing surprising in this if they but knew it almost all men in the ir degree some time or other cherish very nearly the same feelings towards the ocean with me there now is your insular city of the manhattoes belted round by wharves as indian isles by coral reefs commerce surrounds it with her surf right and left the streets take you waterward its e xtreme downtown is the battery where that noble mole is washed by waves and cooled by breezes which a few hours previous were out of sight of land look at the crowds of water-gazers there circumambulate the city of a dreamy sabbath afternoon go from corlears hook to coenties slip and from thence by whitehall northward what do you see posted like silent sentinels all around the town stand thousands upon thousands of mortal m en fixed in ocean reveries some leaning against the spiles some seated upon the pier-heads some looking over the bulwarks of ships from china some high aloft in the rigging as if striving to get a still better seaward peep but these are all landsmen of week days pent up in lath and p laster tied to counters nailed to benches clinched to desks how then is this are the green fields gone what do they here but look here come mo re crowds pacing straight for the water and seemingly bound for a dive strange nothing will content them but the extremest limit of the land l oitering under the shady lee of yonder warehouses will not suffice no they must get just as nigh the water as they possibly can without fallin g in and there they stand miles of them leagues inlanders all they come from lanes and alleys streets and avenues north east south and west ye t here they all unite tell me does the magnetic virtue of the needles of the compasses of all those ships attract them thither once more say y ou are in the country in some high land of lakes take almost any path you please and ten to one it carries you down in a dale and leaves you t here by a pool in the stream there is magic in it let the most absentminded of men be plunged in his deepest reveries stand that man on his le gs set his feet a-going and he will infallibly lead you to water if water there be in all that region should you ever be athirst in the great american desert try this experiment if your caravan happen to be supplied with a metaphysical professor yes as everyone knows meditation and w ater are wedded forever but here is an artist he desires to paint you the dreamiest shadiest quietest most enchanting bit of romantic landscap e in all the valley of the saco what is the chief element he employs there stand his trees each with a hollow trunk as if a hermit and a cruci fix were within and here sleeps his meadow and there sleep his cattle and up from yonder cottage goes a sleepy smoke deep into distant woodlan ds winds a mazy way reaching to overlapping spurs of mountains bathed in their hillside blue but though the picture lies thus tranced and thou gh this pine-tree shakes down its sighs like leaves upon this shepherds head yet all were vain unless the shepherds eye were fixed upon the ma gic stream before him go visit the prairies in june when for scores on scores of miles you wade knee-deep among tiger-lilies what is the one c harm wanting water there is not a drop of water there were niagara but a cataract of sand would you travel your thousand miles to see it why d id the poor poet of tennessee upon suddenly receiving two handfuls of silver deliberate whether to buy him a coat which he sadly needed or inv est his money in a pedestrian trip to rockaway beach why is almost every robust healthy boy with a robust healthy soul in him at some time or other crazy to go to sea why upon your first voyage as a passenger did you yourself feel such a mystical vibration when first told that you an d your ship were now out of sight of land why did the old persians hold the sea holy why did the greeks give it a separate deity and own broth er of jove surely all this is not without meaning and still deeper the meaning of that story of narcissus who because he could not grasp the t ormenting mild image he saw in the fountain plunged into it and was drowned but that same image we ourselves see in all rivers and oceans it i s the image of the ungraspable phantom of life and this is the key to it all now when i say that i am in the habit of going to sea whenever i begin to grow hazy about the eyes and begin to be over conscious of my lungs i do not mean to have it inferred that i ever go to sea as a pass enger for to go as a passenger you must needs have a purse and a purse is but a rag unless you have something in it besides passengers get sea sick grow quarrelsome dont sleep of nights do not enjoy themselves much as a general thing no i never go as a passenger nor though i am someth ing of a salt do i ever go to sea as a commodore or a captain or a cook i abandon the glory and distinction of such offices to those who like them for my part i abominate all honorable respectable toils trials and tribulations of every kind whatsoever it is quite as much as i can do to take care of myself without taking care of ships barques brigs schooners and whatnot and as for going as cook though i confess there is con siderable glory in that a cook being a sort of officer on shipboard yet somehow i never fancied broiling fowls though once broiled judiciously buttered and judgmatically salted and peppered there is no one who will speak more respectfully not to say reverentially of a broiled fowl tha n i will it is out of the idolatrous dotings of the old egyptians upon broiled ibis and roasted river horse that you see the mummies of those creatures in their huge bake-houses the pyramids no when i go to sea i go as a simple sailor right before the mast plumb down into the forecas tle aloft there to the royal masthead true they rather order me about some and make me jump from spar to spar like a grasshopper in a may mead ow and at first this sort of thing is unpleasant enough it touches ones sense of honor particularly if you come of an old established family i n the land the van rensselaers or randolphs or hardicanutes and more than all if just previous to putting your hand into the tar-pot you have been lording it as a country schoolmaster making the tallest boys stand in awe of you the transition is a keen one i assure you from a schoolm aster to a sailor and requires a strong decoction of seneca and the stoics to enable you to grin and bear it but even this wears off in time w hat of it if some old hunks of a sea-captain orders me to get a broom and sweep down the decks what does that indignity amount to weighed i me an in the scales of the new testament do you think the archangel gabriel thinks anything the less of me because i promptly and respectfully ob ey that old hunks in that particular instance who aint a slave tell me that well then however the old sea-captains may order me about however they may thump and punch me about i have the satisfaction of knowing that it is all right that everybody else is one way or other served in mu ch the same way either in a physical or metaphysical point of view that is and so the universal thump is passed round and all hands should rub each others shoulder-blades and be content again i always go to sea as a sailor because they make a point of paying me for my trouble whereas they never pay passengers a single penny that i ever heard of on the contrary passengers themselves must pay and there is all the difference i n the world between paying and being paid the act of paying is perhaps the most uncomfortable infliction that the two orchard thieves entailed upon us but being paid what will compare with it the urbane activity with which a man receives money is really marvellous considering that we so earnestly believe money to be the root of all earthly ills and that on no account can a monied man enter heaven ah how cheerfully we consig n ourselves to perdition finally i always go to sea as a sailor because of the wholesome exercise and pure air of the forecastle deck for as i n this world head winds are far more prevalent than winds from astern that is if you never violate the pythagorean maxim so for the most part the commodore on the quarterdeck gets his atmosphere at second hand from the sailors on the forecastle he thinks he breathes it first but not so in much the same way do the commonalty lead their leaders in many other things at the same time that the leaders little suspect it but wher efore it was that after having repeatedly smelt the sea as a merchant sailor i should now take it into my head to go on a whaling voyage this the invisible police officer of the fates who has the constant surveillance of me and secretly dogs me and influences me in some unaccountable way he can better answer than anyone else and doubtless my going on this whaling voyage formed part of the grand programme of providence that was drawn up a long time ago it came in as a sort of brief interlude and solo between more extensive performances i take it that this part of the bill must have run something like this grand contested election for the presidency of the united states whaling voyage by one ishmael blo ody battle in afghanistan though i cannot tell why it was exactly that those stage managers the fates put me down for this shabby part of a w haling voyage when others were set down for magnificent parts in high tragedies and short and easy parts in genteel comedies and jolly parts i n farces though i cannot tell why this was exactly yet now that i recall all the circumstances i think i can see a little into the springs and motives which being cunningly presented to me under various disguises induced me to set about performing the part i did besides cajoling me in to the delusion that it was a choice resulting from my own unbiased free-will and discriminating judgment chief among these motives was the ov erwhelming idea of the great whale himself such a portentous and mysterious monster roused all my curiosity then the wild and distant seas whe re he rolled his island bulk the undeliverable nameless perils of the whale these with all the attending marvels of a thousand patagonian sigh

```
ts and sounds helped to sway me to my wish with other men perhaps such things would not have been inducements but as for me i am tormented wit
h an everlasting itch for things remote i love to sail forbidden seas and land on barbarous coasts not ignoring what is good i am quick to per
ceive a horror and could still be social with it would they let me since it is but well to be on friendly terms with all the inmates of the pl
ace one lodges in by reason of these things then the whaling voyage was welcome the great floodgates of the wonder-world swung open and in the
wild conceits that swayed me to my purpose two and two there floated into my inmost soul endless processions of the whale and mid most of them
all one grand hooded phantom like a snow hill in the air
```

```
str_length(moby_text_2)
```

```
## [1] 11848
```

Answer: The length of the string is 11848.

Question 7

Finally, split the resulting `moby_text_2` into a list of words using `str_split()` (split at one *or more spaces*). Pipe the result into `unlist()` to convert the resulting list into a character vector, and save it as `word`. Lastly, convert it into a dataframe with one word per row by running `moby_words <- data.frame(word = word)`. **How many unique words are there in the dataset?**

```
## Your code here
word <- moby_text_2 %>% str_split(" +") %>% unlist()
moby_words <- data.frame(word = word)
moby_words %>% unique() %>% count()
```

```
##     n
## 1 850
```

Answer: There are 850 unique words in the dataset.

Question 8

Let's use `mutate()` with `str_length()` to compute the length of each word (call the new variable `length`). **What is the longest word in chapter 1?** *Note: the longest word should have 15 characters including a hyphen.*

```
# Your code goes here
moby_words %>% mutate(length = str_length(word)) %>% slice_max(length, n = 1)
```

```
##              word length
## 1 shoulder-blades     15
```

Answer: The longest word in chapter 1 is "shoulder-blades".

Question 9

A *stop word* are words that are typically filtered out before processing text data, as they are deemed to be irrelevant; words like "and", "a", "to", and so forth. The `tidytext` package (install if you don't have it) includes a list of stop words:

```
library(tidytext)
```

Use `anti_join()` on `moby_words` and `stop_words` (from `tidytext`) to remove common uninformative words. **Compute the most frequently appearing word from the words that remain after removing stop words. What is it?**

```
## Your code here
anti_join(moby_words, stop_words) %>% group_by(word) %>% count() %>% arrange(-n)
```

```
## Joining, by = "word"
```

```
## # A tibble: 620 × 2
## # Groups:   word [620]
##    word          n
##    <chr>     <int>
##  1 sea          10
##  2 water         7
##  3 land          6
##  4 time          6
##  5 voyage        6
##  6 sailor        5
##  7 stand         5
##  8 whaling       5
##  9 money         4
## 10 passenger     4
## # … with 610 more rows
## # ℹ Use `print(n = ...)` to see more rows
```

Answer: The most frequently appearing word, excluding stop words, is “sea”.

```
##                                     sysname
##                                     "Linux"
##                                     release
##                           "4.15.0-193-generic"
##                                     version
## "#204-Ubuntu SMP Fri Aug 26 19:20:21 UTC 2022"
##                                    nodename
##                   "educcomp01.ccbb.utexas.edu"
##                                     machine
##                                    "x86_64"
##                                       login
##                                   "unknown"
##                                        user
##                                    "ehc586"
##                                effective_user
##                                    "ehc586"
```