Lab8 Solutions

Vaibhav Saini - 301386847

2023-03-30

Question 1

```
# create a character vector urls to store URLS
urls <- c()

# read the HTML file into a string
html_string <- readLines("cbc.ca.2023.03.23.html", warn = FALSE)

# loop thru the doc
for (i in 1:length(html_string)) {
    # find the line with the URL
    if (grepl("href", html_string[i])) {
        # extract the URL
        url <- gsub(".*href=\"", "", html_string[i])
        url <- gsub("\".*", "", url)

        # check if the URL is http or https
        if (grepl("http", url)) {
            # check if the URL is already in the vector
            if (!url %in% urls) {
                # add the URL to the vector
                urls <- c(urls, url)
            }
        }
    }
}



# extract URLs using regular expressions
pattern <- "((http|https)://[[:graph:]]+)[[:space:]'\"\\]"
urls <- unlist(regmatches(html_string, gregexpr(pattern, html_string)))

# remove trailing whitespace and quotes from URLs
urls <- gsub("[[:space:]\\'\"]$", "", urls)

# print the length of resulting URLs
sprintf("The number of urls in the HTML are: %i",  length(urls))
```

```
## [1] "The number of urls in the HTML are: 265"
```