Homework 2, Number 1

Aaron Coates

5/4/2019

```
setwd("~/Documents/GitHub/MMSS_311_2")

library(xml2)
library(rvest)
library(tm, tidytext)
```

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

```
library(stringr)
```

1.1 (a.)

```
no <- read_html('/Users/aaroncoates/Desktop/cries.htm')
```

1.1 (b.)

```
nodes <- html_nodes(no, '.mw-category-group+ .mw-category-group a')
```

1.1 (c.)

```
country <- html_text(nodes)
url <- html_attr(nodes, "href")
fullurl <- url_absolute(url, 'https://en.wikipedia.org/wiki/Category:Member_states_of_the_Association_o
combined <- cbind(country, fullurl)
finaldata <- as.data.frame(combined, stringsAsFactors = F)
finaldata
```

```
##        country                                    fullurl
## 1       Brunei      https://en.wikipedia.org/wiki/Brunei
## 2     Cambodia    https://en.wikipedia.org/wiki/Cambodia
## 3    Indonesia   https://en.wikipedia.org/wiki/Indonesia
## 4         Laos        https://en.wikipedia.org/wiki/Laos
## 5     Malaysia    https://en.wikipedia.org/wiki/Malaysia
## 6      Myanmar     https://en.wikipedia.org/wiki/Myanmar
## 7  Philippines https://en.wikipedia.org/wiki/Philippines
## 8    Singapore   https://en.wikipedia.org/wiki/Singapore
## 9     Thailand    https://en.wikipedia.org/wiki/Thailand
## 10     Vietnam     https://en.wikipedia.org/wiki/Vietnam
```

1.1(d.)

```
for(i in 1:10){
  finaldata$text[i] <- finaldata$fullurl[i] %>%
    read_html() %>%
    html_nodes('p+ ul li , p') %>%
    html_text() %>%
  paste(collapse = ' ')
}
```