Assignment 2: Web Scraping

In this assignment on web scraping you will use your browser to inspect webpages, scrape web content using R, and reorganize the scraped data. This assignment contains several new functions. Check them out using `?` and figure out how they work.

Side Quest

By now you should have received your SIM cards. In order to use them it is necessary to activate them, which involves filling out a form and bringing it one of the local Kiosks. You find the instructions on the SIM card package. Please proceed swiftly - activation may take a few days to a week. If you have not yet received your SIM card please order new ones asap.

1 Install Google Chrome

If you don't use Google Chrome please download and install it. (You can stick to your browser if you make sure that you have developer tools installed that allow you to inspect the webpage's source code.)

3 Scrape Wikipedia's Emoji page

3.1 Inspect and identify XPath

Go to https://en.wikipedia.org/wiki/Emoji, locate table of emojis, and *Inspect*. Identify XPath of the Emojis table's node.

```
XPath = '//*[@id="mw-content-text"]/table[8]'

# TRICK
# Browsers like chrome do the job of creating the XPath for you. Simply right
# click while inspecting the page's source and selecting copy/XPath.
```

3.2 Scrape using `rvest`

Install and load package `rvest`. Use `read_html`, `html_node`, and `html_table` (in that order). Before you scrape, execute `options(stringsAsFactors = T)` (to avoid factors). Consider using `magrittr`'s forward pipe operator.

```
install.packages('rvest', repos='http://ftp5.gwdg.de/pub/misc/cran/')
```

```
##
## The downloaded binary packages are in
##  /var/folders/1m/d25960px2zz234hx9g_920686jm8n4/T//Rtmpn6IEPb/downloaded_packages
```

```
library(rvest)

# avoid factors
options(stringsAsFactors = F)

# get page
url  = 'https://en.wikipedia.org/wiki/Emoji'
page = read_html(url)
```

```
# get table
table = page %>%
  html_node(xpath = XPath) %>%
  html_table()
```

Excursus: magrittr's forward pipe operator

Check out: https://cran.r-project.org/web/packages/magrittr/vignettes/magrittr.html

Example: rnorm(100,0,1) %>% mean() %>% round(2) %>% paste('was calculated using magrittr')

```
install.packages('magrittr', repos='http://ftp5.gwdg.de/pub/misc/cran/')
```

```
##
## The downloaded binary packages are in
##  /var/folders/1m/d25960px2zz234hx9g_920686jm8n4/T//Rtmpn6IEPb/downloaded_packages
```

```
library(magrittr)

# magrittr ecercise
rnorm(100,0,1) %>% mean() %>% round(2) %>% paste('was calculated using magrittr')
```

```
## [1] "0.08 was calculated using magrittr"
```

4 Process data

4.1 Reorganize data

Restructure table (here a matrix) to a `data.frame` with 3 columns (named `prefix`, `suffix`, and `emoji`) containing the row annotation (U+00Ax, U+203x, etc.), the column anntotation (0, 1, 2, etc.), and the table content, respectively. Tipp: use ´expand.grid‘. Remove empty cells, annotation cells (cells containing the column names) and Notes (at the bottom) from data frame.

Name resulting table `emoji_ids` and save it on the harddrive (`.rds` or `csv`). Note: Result should have 1088 rows.

```
# get row and colnames
rownames = table[-c(1:2),1]
colnames = c(0:9,LETTERS[1:6])

# create emoji table
emoji_ids = expand.grid(colnames,rownames)[,c(2,1)]
emoji_ids = data.frame(emoji_ids, c(t(table[-(1:2),-1])))
names(emoji_ids) = c('uni_pre','uni_suf','emoji')

# identify annotation elements and notes
eliminate = grepl('Notes\n',emoji_ids$emoji)
for(i in 1:length(colnames)) eliminate = eliminate | grepl(colnames[i],emoji_ids$emoji)
emoji_ids = emoji_ids[!eliminate,]

# remove empty cells
emoji_ids = emoji_ids[emoji_ids$emoji != '',]
```

4.2 Bind prefix and suffix

Create new variable called `unicode_string` in `emoji_ids` containing the combined pre- and suffix (i.e., replace the x in the prefix by the suffix). Use `gsub`.

```
# create unicde_string
emoji_ids$unicode_string = NA
for(i in 1:nrow(emoji_ids)) emoji_ids$unicode_string[i] = gsub('x',emoji_ids$uni_suf[i],emoji_ids$uni_pr

# save on hd
write.csv(emoji_ids,'emoji_ids.csv')
```

End