Webscraping

library(tidyverse)

library(rvest)

## Scrape One Page

url = 'http://quotes.toscrape.com/'

getQuotes <- function(url){

read\_html

read\_html(url) -> rawPage

rawPage %>%

html\_nodes('.text') %>%

html\_text() -> quotes

rawPage %>%

scrape nodes

html\_nodes('.author') %>%

html\_text() -> authors

rawPage %>%

html\_nodes('.tags') %>%

html\_text() -> tags

rawPage %>%

scrape links

html\_nodes(".quote span a") %>%

html\_attr('href') -> authorUrls

quotes = ifelse(length(quotes) == 0,

NA

NA,

quotes)

quoteDF = data.frame(authors,

create DF

tags,

quotes)

}

getQuotes(url) -> data

# Scrape Next Pages

define base url

base\_url = 'http://quotes.toscrape.com/page/'

add page info:

chr [1:10]

urls = paste0(base\_url, 1:10, '/')

map DF

map\_df(urls, getQuotes) -> data

# Scrape Linked Pages

getAuthorUrls = function(url) {

read\_html(url) -> rawPage

rawPage %>%

scrape linked pages

html\_nodes(".quote span a") %>%

html\_attr('href') -> newUrl

create link

paste0('http://quotes.toscrape.com/’, newUrl)

}

1. Scrape Links from one page
2. Map results of all (next) pages

🡪 chr [1:50]

allAuthors = unlist(map(urls, getAuthorUrls))

# Scrape Nodes

🡪 Scrape One Page

getDealz = function(url){

read\_html

read\_html(url) -> raw\_page

raw\_page %>%

scrape node

html\_nodes('.threadGrid') -> nodes

create DF

deals = map\_df(nodes, getNodeDetails)

}

dealz = getDealz(url)