IMT573 Problem Set 3: Scrape data and get it in shape (80pt)

Your name:

Deadline: Wed, Oct 28th 5pm

Instructions

This problem set asks you to pull data from web, and make it suitable for using later in ggplot and related functions.

1. Remember that just a numerical answer is not enough. Always comment and explain your results.
2. Be sure to include well-documented (e.g. commented) code chucks, figures, tables, and clearly written text explanations as necessary. All figures should be clearly labeled and appropriately referenced within the text.
3. Write clearly! Explain what you do and why do you do this, and what can you conclude from your results!

1 Scrape the web

This question asks you to extract for tables of data from the internet. In particular, we are going to load four datasets from `https://www.drroyspencer.com/`. Roy Spencer is a professor at University of Alabama in Huntsville who each month publishes global satellite temperature updates. Apparently he has strong opinion about global warming, and the comments underneath his posts are even more strongly opinionated. So we might consider scraping the comments too and making a study on political polarization. But we do not do it here.

Briefly, your tasks are the following:

- Load the webpage
- Find the first "UAH Global temperature update" entry on this webpage.
- Extract it's date
- Find the links to data files
- Download the data files.

There is a brief introduction to *rvest* and webscraping in R at `https://otoomet.bitbucket.io/machinelearning-R.html#6_Web_Scraping`.

More specifically:

1. (15pt) Use *rvest* library to load and parse the `https://www.drroyspencer.com/` webpage.

This is a blog that each month publishes "UAH Global Temperature Update". (It also publishes other posts and comments that are very opinionated but we skip those for now.)

2. (25pt) Find the most recent such post (i.e. the first such post on the page) and extract from its title:

   (a) Date (month and year)

   (b) Current temperature anomaly

   Hint: use element picker in the browser's developer tools and extract the corresponding header elements.

   Hint2: you can check if a string starts with another string with `startsWith`.

   Hint3: You can split a sentence into words using `strsplit` and then find which element is e.g. "2020". You can do something along these lines:

```
## Extract "+5" from this sentence:
sentence <- "The score for May 2020 is +5"
words <- strsplit(sentence, "[[:space:]]+")[[1]]  # break sentence on spaces,
                                                  # tabs and such
i <- which(words == "2020")
words[i + 2]  # we know that '+5' is two words down from '2020'

## [1] "+5"
```

3. (25pt) Toward the end of the post there are links for data: lower troposphere, mid-troposphere, tropopause, and lower stratosphere. Find this paragraph and extract the corresponding 4 links.

   Hint: find the container for the blog post (it is a *div* but what are its attributes?) and extract that container. Thereafter find all paragraphs *p* inside of that container. Then find the paragraph where there is some of the correct words/phrases. Finally, extract all *a* elements (and their *href*-s).

   Hint2: you can find which paragraph contains a given pattern using `grep`:

```
paragraphs <- c("first paragraph", "second para", "third attempt to write sth")
grep("second", paragraphs)  # 2--second one contains the word "second"

## [1] 2
```

4. (15pt) Finally, download all 4 data files and save the data for later use. Use can use something like

```
url <- "https://data.csv"
data <- readLines(url)  # read data from internet
writeLines(url, "local-data.csv")  # save locally to a local file
```

   Ensure that the data is downloaded, saved, and looks like the right data! We may need it later!

5. Challenge (not graded): find the paragraph where the current "linear warming trend" is presented, and extract that one (in C/decade).

6. Finally–how much time did you spend on this PS?