

# IMT573 Lab 4: Get game scores from the web

Your name:

Deadline: Thu, Oct 22nd 5pm

## Instructions

1. Ensure your result is readable!
2. Ensure the code chunks are visible!
3. Upload both .html and .rmd!

This lab asks you to scrape a few game scores from a college football website. There are two steps:

- Find the correct html element using your browser developer tools
- Extract those elements using `html_node` (for a single one) or `html_nodes` (for multiple ones).

Extract text using `html_text`

## 1 Scrape recent college football scores

Before you start, take a look at the football scores page <https://www.sports-reference.com/cfb/>. Open your browsers developer tools and see what kind of elements do you see.

And now the real thing:

1. Load and parse the college football page <https://www.sports-reference.com/cfb/>
2. Find the table “Recent Games” a little down on the page. What is the element that contains the “Recent Games” text?
3. This element is contained inside a DIV. What is the class of that DIV?  
Note: DIV is a vertical block of text on the page.
4. Extract that DIV. Call this “summaries”

Hint: use `html_node(tag.class)` to extract element of *tag* with class *class*

Now we only work on the the “summaries”. We do not use the rest of the page any more.

5. Underneath the title “Recent Games” you see a lot of scores like

Georgia (3) 24  
Alabama (2) 41

6. Inside of which elements are those texts embedded?
7. Extract all those elements, and texts from therein. Print those texts.  
Optional: You can collapse the tabs and spaces into just a single space. You can do this with:

```
gsub("[[:space:]]+", " ", text)
```

for instance

```
gsub("[[:space:]]+", " ", "a b c\t\t d")  
  
## [1] "a b c d"
```

You can also remove leading and trailing whitespace with ‘trimws’:

```
trimws(" a \t ")  
  
## [1] "a"
```

Congrats! You are done! Enjoy football in the virus-season :-)

## 2 Challenge (not graded)

If you feel this was too simple then there is a more serious task:

- Take these game scores and put them in a data frame. The result should look like:

| Team 1         | Score 1 | Team 2        | Score 2 |
|----------------|---------|---------------|---------|
| Georgia        | 24      | Alabama       | 41      |
| North Carolina | 28      | Florida State | 31      |
| ...            |         |               |         |