Agenda

• Worked example from previous lecture
• Extracting matches with `regmatches()`
• Worked example: scraping web links
• Tagged expressions (capture groups)
Example: E-mail
Find the e-mail addresses

B
Anthony Brockwell
Adjunct Associate Professor

C
Jodi Casabianca
Postdoctoral Fellow
Office: Wean Hall 8119
Phone: (412) 268-6328
Email: jodicasa@andrew.cmu.edu

D
Bernie Devlin
Adjunct Associate Professor
Office: University of Pittsburgh
Phone: (412) 624-1432

Home Page

E
Where are the e-mail addresses?

<ul class="peoplelisting">
<li><a name="B"><br /></a>
<h3>B</h3>
<p></p>
<ul>
<li>Anthony Brockwell<br />
Adjunct Associate Professor</li>
</ul>
</li>
<li><a name="C"><br /></a>
<h3>C</h3>
<p></p>
<ul>
<li>Jodi Casabianca<br />
Postdoctoral Fellow<br />
Office: Wean Hall 8119<br />
Phone: (412) 268-6328<br />
Email: jodicasa @ andrew.cmu.edu</li>
</ul>
</li>
<li><a name="D"><br /></a>
<h3>D</h3>
<p></p>
<ul>
<li>Bernie Devlin<br />
Adjunct Associate Professor<br />
Office: University of Pittsburgh<br />
Phone: (412) 624-1432</li>
<li>George T. Duncan<br />
Emeritus Professor<br />
Office: Hamburg Hall 2102D<br />
Phone: (412) 268-2172<br />
Email: gd17 @ andrew.cmu.edu<br />
<a href="http://duncan.heinz.cmu.edu/GeorgeWeb/">Home Page</a></li>
</ul>
</li>
</ul>
Where are the e-mail addresses?
What’s an e-mail address?

vqv@stat.cmu.edu
cshalizi@stat.cmu.edu

[[:alnum:]-_.]+@[[:alnum:]-_.]+.[[:alpha:]]+
Steps

- Read web page into R – `readLines()`
- Test e-mail pattern* – `grep()`
- Match e-mail pattern – `gregexpr()`
- Extract matches – `regmatches()`

* optional – *just to make sure you didn’t mess up*
Live Demo
• Use `regexpr()`, `gregexpr()`, or `regexec()` to find matching locations

• Then use `regmatches()` to extract the matching substrings
**regmatches()**

\[
y \leftarrow \text{regmatches}(x, m, \text{invert} = \text{FALSE})
\]

- **x** string that you searched
- **m** match data returned from `regexpr()`, `gregexpr()`, or `regexec()`
- **invert** extract non-matched substrings?
# Get indices and lengths of matches
m <- gregexpr(pattern = emailpattern2, cmustat)

# Extract substrings
e-mails <- regmatches(cmustat, m)
Example: Web links
Brittanie Boone (Econ-Stat, 2012) has been chosen as a 2011-2012 Andrew Carnegie Society Scholar. This award is for seniors who "embody high standards of academic excellence combined with multi-dimensional characteristics such as volunteerism, involvement in student organizations, participation in sports or the arts and leadership." Only nine H&SS students were chosen as ACS scholars this year. Besides receiving monetary awards, ACS Scholars work together during the school year to contribute to philanthropic activities on campus.

Congratulations to Brittanie!

Applications are invited for possible tenure-track, lecturer, and visiting positions. Carnegie Mellon offers a collegial faculty environment, emphasizing a combination of disciplinary and cross-disciplinary research and teaching. All areas of statistics are welcome, and joint appointments with other units in the Pittsburgh area are possible. We especially encourage women and minorities to apply. For further information, please contact

Founded in 1966, Carnegie Mellon's Department of Statistics evolved separately from the traditional umbrella of Mathematical Sciences in a University environment that emphasized computation, the understanding of human behavior and decision-making, and cross-disciplinary research. This led the Department to define a path for itself by

- Business & Travel Policy - READ ME FIRST
- Reimbursement Procedures
- Reimbursement FAQs
Goal

• Extract all links on the CMU Stat Dept Homepage
What's a link?

Example matches

\begin{verbatim}
<a href = "http://www.stat.cmu.edu/">
</a>
\end{verbatim}

\begin{verbatim}
<a href="seminar_calendar" style=".."> 
</a>
\end{verbatim}

Regular expression

\begin{verbatim}
<\s*\s+.\s*<\s*href<\s*="\s*[\^]"*[\^>]*>>
\end{verbatim}

\texttt{s} is shorthand for \texttt{[[:space:]]}
Live Demo
Problem

These are some matching substrings

[37] "<a href="http://bayesml.stat.cmu.edu">"

How do we get just the part next to href, between the quotation marks?
We want the green part

Example matches

\[\text{\textless a href = "http://www.stat.cmu.edu/"}\textgreater
\text{\textless a href="seminar_calendar" style=".."}\textgreater

Regular expression

\[\text{\textless\textbackslash s*\textbackslash a\textbackslash s+.\textbackslash s*href\textbackslash s*=}\textbackslash s*\textquoteright [^\textquoteright][][^\textgreater][]\textgreater\textbackslash s*\textgreater\]
Tagged Expressions

- Enclose *capture group* with parentheses
- Can be extracted separately using `regexec()`
Example matches

```html
<a href = "http://www.stat.cmu.edu/">
<a href="seminar_calendar" style="..">
```

Regular expression

```regex
\[s*a\s+.\*\s*href\s*="\s*"([\^\"]\*)"[\^>\]\*>
```

*capture group*
– enclosed in parentheses
\textbf{regexec()}

- Like \texttt{regexpr()}, except it also returns positions of substrings corresponding to capture groups

- \textbf{Note!} \texttt{regexec()} returns at most 1 match (and corresponding groups) for each element of \( x \)

- Use a two-pass approach (with \texttt{gregexpr()}) when there are multiple matches
```
> pat <- '<<\s*a\s+.\s*href\s*\s*=\s*"([\^]*)"[^>]*>'
> str <- '<a href="http://facebook.com">'
> m <- regexec(pat, str, ignore.case = TRUE)
> print(m)
[[1]]
[1] 1 10
attr("match.length")
[1] 30 19

> regmatches(str, m)
[[1]]
```
Live Demo
Summary

- Use capture groups (parentheses) when you want to capture a substring within a match
- Not restricted to web-related applications
- See class webpage for R code used in live demos
- Next: Importing data from websites