Meet a New R Package: scrapeR

Joe Walsh

May 1, 2013

What scrapeR Does

Directly

- downloads & preprocesses webpages
- saves as a list
- one function: scrape()

What scrapeR Does

Directly

- downloads & preprocesses webpages
- saves as a list
- one function: scrape()

- Dependency package: XML

What scrapeR Does

Directly

- downloads & preprocesses webpages
- saves as a list
- one function: scrape()

- Dependency package: XML

Indirectly

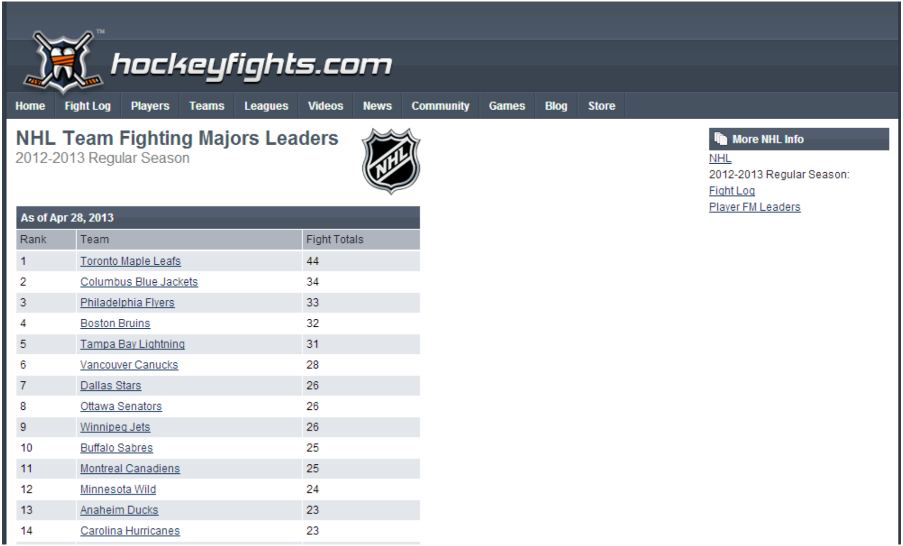
- only requires three lines of code
- can be automated
- dynamic reports
- improved replicability

Why I Like It

- huge amounts of data
- new data being released
- can stop static reports
- eases replication

Example: Cherry Hypothesis

Example: Cherry Hypothesis



NHL Team Fighting Majors Leaders

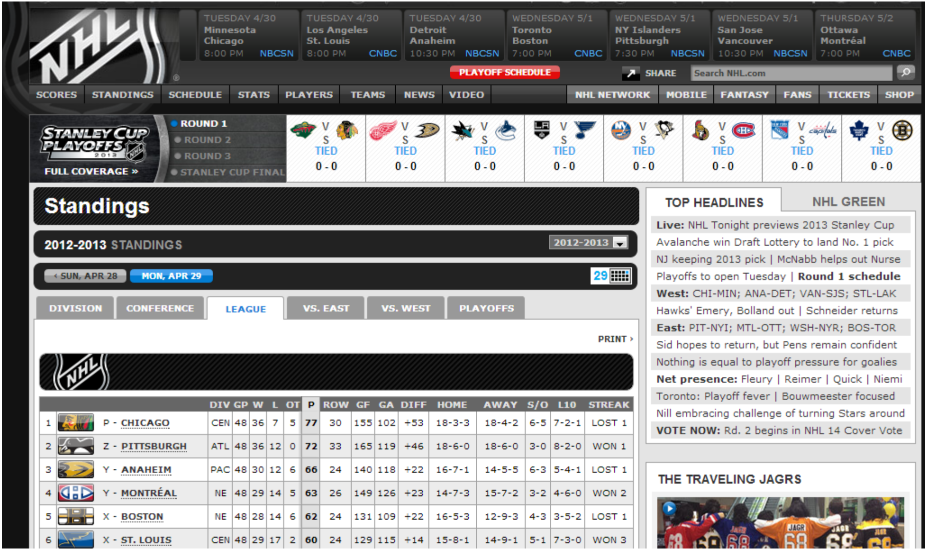
2012-2013 Regular Season

As of Apr 28, 2013

Rank	Team	Fight Totals
1	Toronto Maple Leafs	44
2	Columbus Blue Jackets	34
3	Philadelphia Flyers	33
4	Boston Bruins	32
5	Tampa Bay Lightning	31
6	Vancouver Canucks	28
7	Dallas Stars	26
8	Ottawa Senators	26
9	Winnipeg Jets	26
10	Buffalo Sabres	25
11	Montreal Canadiens	25
12	Minnesota Wild	24
13	Anaheim Ducks	23
14	Carolina Hurricanes	23

NHL
2012-2013 Regular Season:
Fight Log
Player FM Leaders

Example: Cherry Hypothesis



TUESDAY 4/30 Minnesota Chicago 8:00 PM NBCSN | TUESDAY 4/30 Los Angeles St. Louis 8:00 PM CNBC | TUESDAY 4/30 Detroit Anaheim 10:30 PM NBCSN | WEDNESDAY 5/1 Toronto Boston 7:00 PM CNBC | WEDNESDAY 5/1 NY Islanders Pittsburgh 7:30 PM NBCSN | WEDNESDAY 5/1 San Jose Vancouver 10:30 PM NBCSN | THURSDAY 5/2 Ottawa Montréal 7:00 PM CNBC

VS TIED 0 - 0 | VS TIED 0 - 0 | VS TIED 0 - 0 | VS TIED 0 - 0 | VS TIED 0 - 0 | VS TIED 0 - 0 | VS TIED 0 - 0 | VS TIED 0 - 0

Standings

2012-2013 STANDINGS

			DIV	GP	W	L	OT	P	ROW	GF	GA	DIFF	HOME	AWAY	S/O	L10	STREAK
1		P - CHICAGO	CEN	48	36	7	5	77	30	155	102	+53	18-3-3	18-4-2	6-5	7-2-1	LOST 1
2		Z - PITTSBURGH	ATL	48	36	12	0	72	33	165	119	+46	18-6-0	18-6-0	3-0	8-2-0	WON 1
3		Y - ANAHEIM	PAC	48	30	12	6	66	24	140	118	+22	16-7-1	14-5-5	6-3	5-4-1	LOST 1
4		Y - MONTRÉAL	NE	48	29	14	5	63	26	149	126	+23	14-7-3	15-7-2	3-2	4-6-0	WON 2
5		X - BOSTON	NE	48	28	14	6	62	24	131	109	+22	16-5-3	12-9-3	4-3	3-5-2	LOST 1
6		X - ST. LOUIS	CEN	48	29	17	2	60	24	129	115	+14	15-8-1	14-9-1	5-1	7-3-0	WON 3

TOP HEADLINES | NHL GREEN

- **Live:** NHL Tonight previews 2013 Stanley Cup
- Avalanche win Draft Lottery to land No. 1 pick
- NJ keeping 2013 pick | McNabb helps out Nurse
- Playoffs to open Tuesday | **Round 1 schedule**
- **West:** CHI-MIN; ANA-DET; VAN-SJS; STL-LAK
- Hawks' Emery, Bolland out | Schneider returns
- **East:** PIT-NYI; MTL-OTT; WSH-NYR; BOS-TOR
- Sid hopes to return, but Pens remain confident
- Nothing is equal to playoff pressure for goalies
- **Net presence:** Fleury | Reimer | Quick | Niemi
- Toronto: Playoff fever | Bouwmeester focused
- Nill embracing challenge of turning Stars around
- **VOTE NOW:** Rd. 2 begins in NHL 14 Cover Vote

THE TRAVELING JAGRS

Example: Cherry Hypothesis

```
# install and load scrapeR package
install.packages("scrapeR", dependencies=TRUE)
library(scrapeR)
```

Example: Cherry Hypothesis

```
# install and load scrapeR package
install.packages("scrapeR", dependencies=TRUE)
library(scrapeR)

# scrape data
URLs <- c("http://www.nhl.com/ice/standings.htm?type=lea",
          "http://www.hockeyfights.com/leaders/teams/")
pageSource <- scrape(url=URLs, headers=FALSE, parse=TRUE)
```

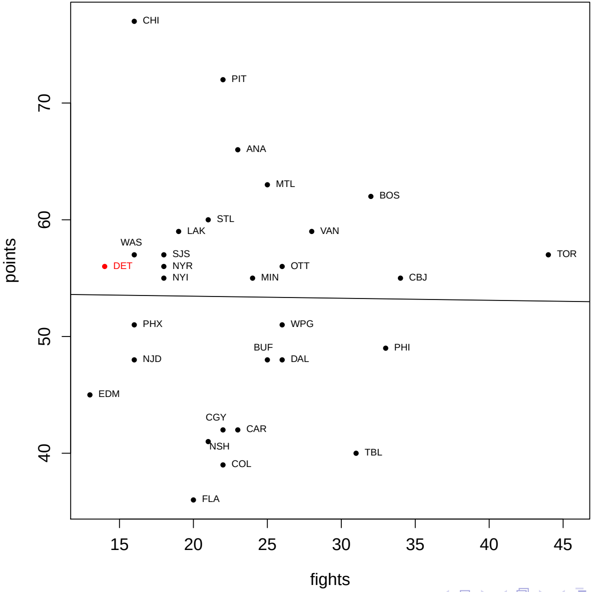
Example: Cherry Hypothesis

```
# install and load scrapeR package
install.packages("scrapeR", dependencies=TRUE)
library(scrapeR)

# scrape data
URLs <- c("http://www.nhl.com/ice/standings.htm?type=lea",
          "http://www.hockeyfights.com/leaders/teams/")
pageSource <- scrape(url=URLs, headers=FALSE, parse=TRUE)

# get two tables with info we need
NHL.tables <- readHTMLTable(pageSource[[1]])
  team.records <- NHL.tables[[3]]
HockeyFights.tables <- readHTMLTable(pageSource[[2]])
  team.fights <- HockeyFights.tables[[1]]
```

Example: Cherry Hypothesis



Questions?

Email: j.thomas.walsh@gmail.com
GitHub: jtwalsh0