Scraping NFL Draft Data

Alex Romero Honrubia

6 de julio de 2018

Goal

The Goal is to download data from the NFL Draft from 2002 (last team joins the league) and do some basic statistics to explore the data and the quality of the teams. We will consider that a team that have a lot of picks in the the first 10 for several years is a bad team.

Data source

The data source is Wikipedia, that have the same URL for each draft just changing the year. The web is based on HTML5 technology.

Approach/Technology used and steps.

For downloading the data I have used the package rvest. As can be seen in the code the instruction is very easy: - A loop from 2002 to 2018 that just read the URL with *read_html* function and a *paste0* to move though the URL. - A *html_node* function that reflexes the type of the node where the data is stored, that has been identified using the *Gadget Selector*. - A *html_table* that interpret the data as a table.

The code that follows the scraping is just data management to obtain the data.frame as tidy as possible.

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
require(rvest)
```

```
## Loading required package: rvest
```

```
## Warning: package 'rvest' was built under R version 3.4.4
```

```
## Loading required package: xml2
```

```
## Warning: package 'xml2' was built under R version 3.4.4
```

```
require(knitr)
```

```
## Loading required package: knitr
```

```
## Warning: package 'knitr' was built under R version 3.4.4
```

```
urls <- character()
a = list()
fantasy = list()
for(i in 2002:2018){
  a <- read_html(paste0("https://en.wikipedia.org/wiki/",i,"_NFL_Draft"))
  rating <- a %>%
    html_nodes("div table") %>%
    html_table(fill =TRUE)
  #now the selection of the tables that we are loocking fot in the list
  #of tables in function of the wikipedia year page.
  if (i %in% c(2003,2004,2005,2006,2007,2009,2011,2012,2013,2014)){pass <- rating[[6]]}
  else if (i %in% c(2002,2008,2015,2016,2017,2018)){pass <- rating[[5]]}
  else if (i %in% c(2010)){pass <- rating[[7]]}
  #selection of the 7 relevant variables
  pass <- pass[,c(2:9)]
  #creation of the new variable of the draft year
  pass$year <- i
  pass$`Pick #` <- as.numeric(pass$`Pick #`)
  #concatenating all the tables
  if(i!=2002){
    fantasy <- bind_rows(fantasy, pass)
  }
  else {fantasy <- pass}
}

head(fantasy)
```

```
##   Rnd. Pick #           NFL team           Player Pos.        College
## 1    1      1     Houston Texans       David Carr   QB   Fresno State
## 2    1      2  Carolina Panthers Julius Peppers    DE North Carolina
## 3    1      3      Detroit Lions  Joey Harrington   QB         Oregon
## 4    1      4      Buffalo Bills    Mike Williams   OT          Texas
## 5    1      5 San Diego Chargers   Quentin Jammer   CB          Texas
## 6    1      6 Kansas City Chiefs        Ryan Sims   DT North Carolina
##    Conf.               Notes year
## 1    WAC     pre-signed[N 1] 2002
## 2    ACC                     2002
## 3 Pac-10                     2002
## 4 Big 12                     2002
## 5 Big 12                     2002
## 6    ACC from Dallas[R1 - 1] 2002
```

```
tail(fantasy)
```

```
##      Rnd. Pick #             NFL team           Player Pos.        College
## 4362   7*    251 Los Angeles Chargers   Justin Jackson   RB   Northwestern
## 4363   7*    252   Cincinnati Bengals       Rod Taylor    G       Ole Miss
## 4364   7*    253   Cincinnati Bengals       Auden Tate   WR  Florida State
## 4365   7*    254    Arizona Cardinals Korey Cunningham   OT     Cincinnati
## 4366   7*    255        Buffalo Bills    Austin Proehl   WR North Carolina
## 4367   7*    256  Washington Redskins       Trey Quinn   WR            SMU
##             Conf.                                      Notes year
```

```
## 4362      Big Ten                                                    2018
## 4363          SEC                                                    2018
## 4364          ACC                                                    2018
## 4365 The American                                                    2018
## 4366          ACC                          from Tampa Bay [R7 - 21] 2018
## 4367 The American Mr. Irrelevantfrom Atlanta via LA Rams [R7 - 22] 2018
```

As we can see, the scraping result is a data frame with seven columns and all the players selected from the draft concatenated.

Short analysis of the data

To explore the data and show how some results first of all I have done some preprocessing.

Rename teams that have changed his name along the 17 years span.

```
fantasy$`NFL team`[fantasy$`NFL team` == "San Diego Chargers"] <- "Los Angeles Chargers"
fantasy$`NFL team`[fantasy$`NFL team` == "St. Louis Rams"] <- "Los Angeles Rams"
```

Correct that some teams have 2 blanks between words instead of one.

```
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 3.4.4
```

```
fantasy$`NFL team` <- gsub("\\s+", " ", str_trim(fantasy$`NFL team`))
```

Basic stastistics.

Top ten teams with more top ten draft picks.

```
fantasy %>%
  filter(`Pick #` %in% c(1:10)) %>%
  group_by(`NFL team`) %>%
  summarise(n=n()) %>%
  arrange(desc(n)) %>%
  top_n(10) %>%
  kable()
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
## Selecting by n
```

NFL team	n
Jacksonville Jaguars	13
Cleveland Browns	10
Detroit Lions	10
New York Jets	8
Oakland Raiders	8
Arizona Cardinals	7
Buffalo Bills	7
Los Angeles Rams	7
San Francisco 49ers	7
Tennessee Titans	7

Top colleges with more drafted players.

```
fantasy %>%
  group_by(College, Conf.) %>%
  summarise(n=n()) %>%
  arrange(desc(n)) %>%
  head(10)%>%
  kable()
```

College	Conf.	n
Ohio State	Big Ten	109
LSU	SEC	103
Alabama	SEC	102
Florida	SEC	96
Georgia	SEC	91
Florida State	ACC	90
Oklahoma	Big 12	84
USC	Pac-10	69
Clemson	ACC	68
Miami (FL)	ACC	67

Top colleges with more drafted players in the top ten picks.

```
fantasy %>%
  filter(`Pick #` %in% c(1:10)) %>%
```

```
  group_by(College) %>%
  summarise(n=n()) %>%
  arrange(desc(n)) %>%
  head(10) %>%
  kable()
```

College	n
USC	11
Alabama	9
LSU	9
Ohio State	8
Oklahoma	7
Georgia	6
Miami (FL)	6
Texas	6
Texas A&M	6
Auburn	5

Drafteds Heisman Trophies winners (Heisman Trophie is the college award for the best player of the year).

```
fantasy %>%
  group_by(year) %>%
  filter(str_detect(Notes, 'Heisman')) %>%
  arrange(`Pick #`) %>%
  select(`Pick #`, `NFL team`, Player, Pos., College, year) %>%
  kable()
```

Pick #	NFL team	Player	Pos.	College	year
1	Cincinnati Bengals	Carson Palmer	QB	USC	2003
1	Carolina Panthers	Cam Newton	QB	Auburn	2011
1	Tampa Bay Buccaneers	Jameis Winston	QB	Florida State	2015
1	Cleveland Browns	Baker Mayfield	QB	Oklahoma	2018
2	Washington Redskins	Robert Griffin III	QB	Baylor	2012
2	Tennessee Titans	Marcus Mariota	QB	Oregon	2015
10	Arizona Cardinals	Matt Leinart	QB	USC	2006
22	Cleveland Browns	Johnny Manziel	QB	Texas A&M	2014
28	New Orleans Saints	Mark Ingram Jr.	RB	Alabama	2011
32	Baltimore Ravens	Lamar Jackson	QB	Louisville	2018
45	Tennessee Titans	Derrick Henry	RB	Alabama	2016
95	Los Angeles Rams	Eric Crouch	WR	Nebraska	2002

Teams that have drafted more Heisman Trophie winners.

```
fantasy %>%
  group_by(`NFL team`) %>%
  filter(str_detect(Notes, 'Heisman')) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  kable()
```

NFL team	n
Cleveland Browns	2
Tennessee Titans	2
Arizona Cardinals	1
Baltimore Ravens	1
Carolina Panthers	1
Cincinnati Bengals	1
Los Angeles Rams	1
New Orleans Saints	1
Tampa Bay Buccaneers	1
Washington Redskins	1

Teams that have drafted more Pro-bowlers. (Pro-bowlers are something like the best players in the season)

The difficult here is to select the players that have the called *dagger* symbol in unicode, that means that the player is a Pro-Bowler.

```
fantasy %>%
  group_by(`NFL team`) %>%
  filter(str_detect(Player, '[^\001-\177]')) %>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  head(10) %>%
  kable()
```

NFL team	n
Dallas Cowboys	25
Los Angeles Chargers	22
Green Bay Packers	20
Kansas City Chiefs	20
New Orleans Saints	19
Carolina Panthers	18
Houston Texans	17
Minnesota Vikings	17
Pittsburgh Steelers	17
San Francisco 49ers	17

Limitations found

The limitations that I have found is that the table of the wikipedia page where the data is stored is not the same every year, so it's different for every page. I have lost some time to understand the problem and found the solution, which has been focus in every year and find the exactly table where the data is stored, for data through 50 years it would have been impossible or a big lost of time.