

Adatállományok gyűjtése

Kehl Dániel

PTE-KTK Közgazdaságtan és Ökonometria Intézet



PÉCSI KÖZGÁZ
aholman a karrier indul

Adatállományok gyűjtése

Bevezetés

Cél: máshogy gondolni az elérhető, elemzendő adatok körére, mint hagyományosan

- ▶ elsősorban R csomagokon, illetve példákon keresztül, de más szoftverek is rendelkezésre állnak
- ▶ reproducible research népszerűsítése
- ▶ nem cél egyetlen kutatási eredmény részletezése
- ▶ lehetőségek felvillantás
- ▶ kódokat igény esetén szívesen megosztok

Adatállományok gyűjtése

Adatok közvetlen elérése

Adatbázisok

- ▶ relációs adatbázisok
 - ▶ RMySQL, RPostgreSQL, ROracle, RSQLite stb.
- ▶ nem-relációs adatbázisok

Hivatalos statisztikai adatbázisok keresése és közvetlen elérése

- ▶ restatapi: Eurostat
- ▶ WDI: World Development Indicators (World Bank)
- ▶ OECD: OECD
- ▶ PxWebApiData: Statistics Norway, Statistics Sweden and Statistics Finland
- ▶ tidycensus: USA népszámlálási és geo adatok
- ▶ essurvey: European Social Survey

Adatállományok gyűjtése

Adatok közvetlen elérése – eurostat

```
# GDP and main components (output, expenditure and income)

dt1 <- get_eurostat_data("nama_10_gdp")

# Production of cow's milk on farms by NUTS 2 regions

dt2 <- get_eurostat_data("agr_r_milkpr",
                        date_filter = 2008,
                        keep_flags = TRUE)

# Employment by A*10 industry breakdowns

dt3 <- get_eurostat_data("nama_10_a10_e",
                        filters = c("Annual", "EU28", "Belgium",
                                   "AT", "Total", "EMP_DC", "person"),
                        date_filter = c(2002, 2008, 2013:2018))
```

Adatállományok gyűjtése

Adatok, szolgáltatások API-n keresztül

Social listening

- ▶ rtweet: Twitter kapcsolatok, követők, tweet szövegek, nyelvi, területi szűrésben, posztolás
- ▶ gtrendsR: Google Trends
- ▶ RoogLeVision: képek automatikus felismerése
 - ▶ márkanév keresés
 - ▶ látnivaló felismerés
 - ▶ kép tartalmának meghatározása tömegesen
- ▶ instaR, Rlinkedin, Rfacebook

Tőzsdei és pénzügyi adatok

- ▶ tidyquant
- ▶ aemo: Australian Energy Market Operator, real-time adatok

Adatállományok gyűjtése

<https://www.r-bloggers.com/google-vision-api-in-r-rooglevision/>



```
dog_mountain_label = getGoogleVisionResponse('dog_mountain.jpg',  
                                              feature = 'LABEL_DETECTION')  
  
head(dog_mountain_label)
```

##	mid	description	score
## 1	/m/09d_r	mountain	0.9188690
## 2	/g/11jxkqbpp	mountainous landforms	0.9009549
## 3	/m/023bbt	wilderness	0.8733696
## 4	/m/0kpmf	dog breed	0.8398435
## 5	/m/0d4djn	dog hiking	0.8352048

Adatállományok gyűjtése

Web scraping

- ▶ rvest: magas szintű csomag statikus és dinamikus weboldalak adatainak begyűjtésére
 - ▶ egy wikipedia oldal
 - ▶ futballisták piaci értéke
 - ▶ holtankoljak.hu

Adatállományok gyűjtése

Web scraping – wikipedia

https://en.wikipedia.org/wiki/2018_FIFA_World_Cup_squads

- 11 References
- 12 External links

Group A [edit]

Egypt [edit]

Coach:  Héctor Cúper

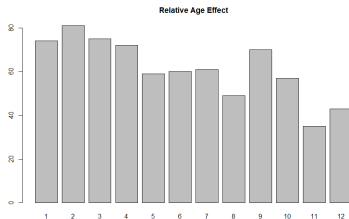
Egypt's 29-man preliminary squad was announced on 13 May 2018.^[6] The final squad was announced on 4 June.^[7]

No. ↕	Pos. ↕	Player ↕	Date of birth (age) ↕	Caps ↕	Goals ↕	Club ↕
1	GK	Essam El Hadary (<i>captain</i>)	15 January 1973 (aged 45)	158	0	 Al-Taawoun
2	DF	Ali Gabr	1 January 1989 (aged 29)	21	1	 West Bromwich Albion
3	DF	Ahmed Elmohamady	9 September 1987 (aged 30)	78	2	 Aston Villa
4	MF	Omar Gaber	30 January 1992 (aged 26)	24	0	 Los Angeles FC
5	MF	Sam Morsy	10 September 1991 (aged 26)	5	0	 Wigan Athletic
6	DF	Ahmed Hegazi	25 January 1991 (aged 27)	45	1	 West Bromwich Albion
7	DF	Ahmed Fathy	10 November 1984 (aged 33)	126	3	 Al Ahly
8	MF	Tarek Hamed	24 October 1988 (aged 29)	21	0	 Zamalek
9	FW	Marwan Mohsen	26 February 1989 (aged 29)	24	4	 Al Ahly
10	FW	Mohamed Salah	15 June 1992 (aged 25)	57	33	 Liverpool
11	FW	Kahraba	13 April 1994 (aged 24)	19	3	 Al-Ittihad
12	DF	Ayman Ashraf	9 April 1991 (aged 27)	4	0	 Al Ahly
13	DF	Mohamed Abdel Shafy	1 July 1985 (aged 32)	51	1	 Al-Fateh
14	FW	Ramadan Sobhi	23 January 1997 (aged 21)	23	1	 Stoke City
15	DF	Mahmoud Hamdy	1 June 1995 (aged 23)	0	0	 Zamalek
16	GK	Sherif Ekramy	10 July 1983 (aged 34)	22	0	 Al Ahly
17	MF	Mohamed Elneny	11 July 1992 (aged 25)	61	5	 Arsenal

Adatállományok gyűjtése

Web scraping – wikipedia

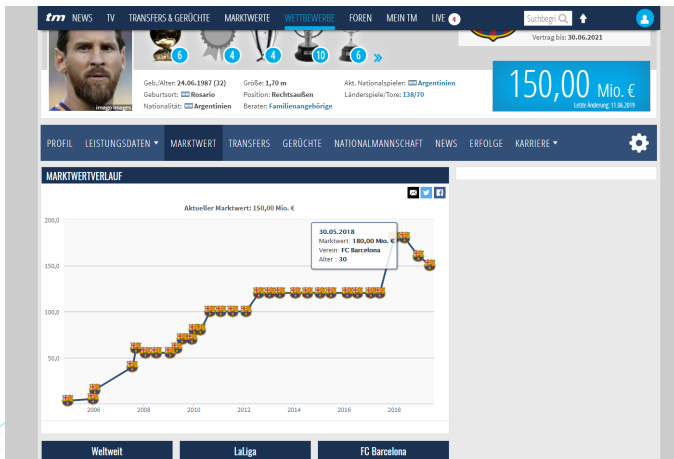
```
mytable <- read_html("https://en.wikipedia.org/wiki/2018_FIFA_World_Cup_squads") %>%  
  html_table() %>%  
  magrittr::extract(1:32) %>%  
  bind_rows() %>%  
  mutate(  
    mob = substr('Date of birth (age)', 7, 8) %>% as.numeric()  
  )  
  
barplot(table(mytable$mob), main = "Relative Age Effect")
```



Adatállományok gyűjtése

Web scraping – Transfermarkt.de

<https://www.transfermarkt.de/lionel-messi/marktwertverlauf/spieler/28003>



Adatállományok gyűjtése

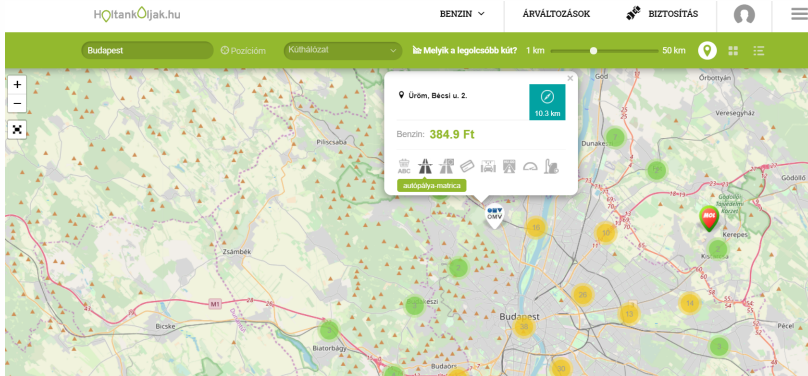
Web scraping – Transfermarkt.de

```
mydata <- read_html(  
  "https://www.transfermarkt.de/lionel-messi/marktwertverlauf/spieler/28003") %>%  
  html_nodes('script') %>%  
  html_text() %>%  
  magrittr::extract(45) %>%  
  str_extract("\\[\\{ 'type': 'line', 'name': 'Marktwert', 'data': .*") %>%  
  str_replace("\\[\\{ 'type': 'line', 'name': 'Marktwert', 'data': ", " ") %>%  
  str_replace(pattern = "\\}\\}\\]", 'credits': \\{ 'enabled': false\\}, 'legend'.*, " ") %>%  
  str_replace_all(pattern = "'", replacement = '\\\"') %>%  
  encodeString() %>%  
  fromJSON()
```

Adatállományok gyűjtése

Web scraping – holtankoljak.hu

<https://holtankoljak.hu/uzemanyagok#a=map&view=map&fuelTypeID=1&city=Budapest&searchLat=47.498880&searchLng=19.043634&range=20>



Adatállományok gyűjtése

Web scraping – holtankoljak.hu

```
get_stations <- function(min_lat = 46.2, max_lat = 48.5,
                           min_lon = 16, max_lon = 23,
                           step_lat = 10, step_lon = 10,
                           my_fuel = 1,
                           one_df = TRUE){

  res <- list()
  for (i in 0:step_lon) {
    for (j in 0:step_lat) {
      my_lat <- min_lat + (max_lat - min_lat)/step_lat * j
      my_lon <- min_lon + (max_lon - min_lon)/step_lon * i
      my_JSON <- fromJSON(
        paste0(
          "https://holtankoljak.hu/ajax.php?a=search&view=map&fuelTypeID=",
          my_fuel,
          "&city=&lat=",
          my_lat,
          "&lng=",
          my_lon,
          "&range=50&subpage_path=uzemanyagok"),
        flatten=TRUE)
      my_df <- my_JSON %>% magrittr::extract2("markers")
      my_fuel_type <- my_JSON %>% magrittr::extract2("fuelTypeName") %>%
        make.names()
```

Adatállományok gyűjtése

Web scraping – holtankoljak.hu

```
if (is.data.frame(my_df)) {  
  res[[paste(my_lat, "-", my_lon)]] <- my_df %>%  
    select(-c(marker, logo, distance, image, pricesHTML)) %>%  
    mutate(review_num = my_df %>% extract("review_num") %>%  
      unlist %>% as.numeric(),  
           price = price %>% as.numeric(),  
           price = replace(price, price == 0, NA),  
           url = str_replace(url, pattern = "\\?.*",  
                             replacement = "")) %>%  
    rename(!my_fuel_type := price)  
}  
}  
}  
if (one_df == TRUE) {  
  res <- res %>% bind_rows() %>% distinct(url, .keep_all = TRUE)  
}  
return(res)  
Encoding(res$address) <- "UTF-8"  
}
```

Adatállományok gyűjtése

Köszönetnyilvánítás

A kutatás az Emberi Erőforrás Fejlesztési Operatív Program, EFOP-3.6.2-16-2017-00003: „Sport- Rekreációs- és Egészséggazdasági Kooperációs Kutatóhálózat létrehozása” című projektjének támogatásával készült.