Data 607 Project 3: Data Science Skills

Judd Anderman

October 23, 2016

```
library(rvest)
library(tidyr)
library(dplyr)
library(stringr)
library(jsonlite)
```

Sources:

- KDnuggets September 2016 Top 16 Active Big Data, Data Science Leaders on LinkedIn
- KDnuggets May 2016 Meet the 11 Big Data & Data Science Leaders on LinkedIn
- LinkedIn's Top 25 Data Scientist Profiles (as of 10/18/16)

```
kdlist1 <-
  read_html("http://www.kdnuggets.com/2016/05/10-big-data-data-science-leaders-linkedin.html")

kd_urls_1 <- kdlist1 %>% html_nodes("p") %>%
  html_nodes("b") %>% html_nodes("a") %>% html_attr("href")

# Drop first element in kd_urls_1 vector since Bernard Marr's profile address is in kd_urls_2
kd_urls_1 <- kd_urls_1[-1]

kdlist2 <-
  read_html("http://www.kdnuggets.com/2016/09/top-big-data-science-leaders-linkedin.html")

kd_urls_2 <- kdlist2 %>% html_nodes("p") %>%
  html_nodes("b") %>% html_nodes("a") %>% html_attr("href")

linkedin_urls <- c("https://www.linkedin.com/in/dpatil",
              "https://www.linkedin.com/in/nan-li",
              "https://www.linkedin.com/in/vincentg",
              "https://www.linkedin.com/in/halbut",
              "https://www.linkedin.com/in/ajayohri",
              "https://www.linkedin.com/in/lars-heppert-255190a7",
              "https://www.linkedin.com/in/miketamir",
              "https://www.linkedin.com/in/puneet-jain-b45b5281",
              "https://www.linkedin.com/in/vikasagrawalresearch",
              "https://www.linkedin.com/in/daviabdallah",
              "https://www.linkedin.com/in/selen-uguroglu-2a42b52a",
              "https://www.linkedin.com/in/michellewetzler",
              "https://www.linkedin.com/in/sudalairajkumar",
              "https://www.linkedin.com/in/christophe-bourgoin-ph-d-280b66",
              "https://www.linkedin.com/in/mlunardi",
              "https://www.linkedin.com/in/lillianpierson",
              "https://www.linkedin.com/in/dwaynesmurdon",
              "https://www.linkedin.com/in/mbenesty",
              "https://www.linkedin.com/in/vedprakash93",
              "https://www.linkedin.com/in/klsrao",
              "https://www.linkedin.com/in/abhisvnit",
              "https://www.linkedin.com/in/aqumar-hussain-b2057557",
              "https://www.linkedin.com/in/dmztheone",
              "https://www.linkedin.com/in/drandrews",
              "https://www.linkedin.com/in/wenwen-tao-0842aa1a"
              )

user_URLs <- c(kd_urls_1, kd_urls_2, linkedin_urls)

user_URLs <- unique(user_URLs)
```

R code for `scrape_linkedin()` adapted from Dean Attali's GitHub Gist:

```
scrape_linkedin <- function(user_url) {
  linkedin_url <- "http://linkedin.com/"
  pgsession <- html_session(linkedin_url)
  pgform <- html_form(pgsession)[[1]]
  filled_form <- set_values(pgform,
                            session_key = username,
                            session_password = password)

  submit_form(pgsession, filled_form)

  pgsession <- jump_to(pgsession, user_url)
  page_html <- read_html(pgsession)

  name <-
    page_html %>% html_nodes("#name") %>% html_text()

  title <-
    page_html %>% html_nodes("p.title") %>% html_text()

  location <-
    page_html %>% html_nodes("#location .locality") %>% html_text()

  edu_school <-
    page_html %>% html_node("div.education") %>% html_nodes("h4") %>%
    html_text()

  degree <- page_html %>% html_node("div.education") %>%
    html_node("span.degree") %>% html_text()

  major <- page_html %>% html_node("div.education") %>%
    html_node("span.major") %>% html_text()

  edu_dates <- page_html %>% html_node("div.education") %>%
    html_node("span.education-date") %>% html_nodes("time") %>% html_text()

  num_connections <-
    page_html %>% html_nodes(".member-connections strong") %>% html_text()
  type_connections <-
    str_extract(page_html %>% html_nodes(".member-connections") %>% html_text(),
                "[:alpha:]+")

  skills_nodes <-
    page_html %>% html_nodes("#profile-skills")

  skills <-
     lapply(skills_nodes, function(node) {
       num <- node %>% html_nodes(".num-endorsements") %>%
          html_attr("data-count")
       name1 <- node %>% html_nodes("li.has-endorsements") %>%
          html_attr("data-endorsed-item-name")
       name2 <- node %>% html_nodes("li.no-endorsements") %>%
          html_attr("data-endorsed-item-name")
       data.frame(name = c(name1, name2), num = num)
      })


  skills <- do.call(rbind, skills)


  list(
    name = name,
    title = title,
    location = location,
    edu_school = edu_school,
    degree = ifelse(!is.na(major), str_c(degree, major),
                    ifelse(!is.na(degree), degree, "NA")),
    edu_dates = ifelse(length(str_c(edu_dates, collapse = "")) > 0,
                       str_c(edu_dates, collapse = ""), "NA"),
    num_connections = num_connections,
    type_connections = type_connections,
    skills = skills
  )
}
```

```
data.skills <- lapply(user_URLs, scrape_linkedin)
```

```
list_el_to_df <- function(x){
  data.frame(data_scientist = rep(data.skills[[x]]$name, times =
                                    nrow(data.skills[[x]]$skills)),
             title = rep(data.skills[[x]]$title, times =
                           nrow(data.skills[[x]]$skills)),
             location = rep(data.skills[[x]]$location, times =
                              nrow(data.skills[[x]]$skills)),
             edu_school = rep(data.skills[[x]]$edu_school, times =
                                nrow(data.skills[[x]]$skills)),
             degree = rep(data.skills[[x]]$degree, times =
                            nrow(data.skills[[x]]$skills)),
             edu_dates = rep(data.skills[[x]]$edu_dates, times =
                               nrow(data.skills[[x]]$skills)),
             num_connections = rep(data.skills[[x]]$num_connections, times =
                                     nrow(data.skills[[x]]$skills)),
             type_connections = rep(data.skills[[x]]$type_connections, times =
                                      nrow(data.skills[[x]]$skills)),
             skill = data.skills[[x]]$skills[, 1],
             endorsements =  data.skills[[x]]$skills[, 2],
             stringsAsFactors = FALSE)
}

userskills <- bind_rows(lapply(seq_along(data.skills), list_el_to_df))

# Character encoding conversion
userskills$data_scientist <-
  iconv(userskills$data_scientist, "latin1", "ASCII", sub = "")
userskills$title <-
  iconv(userskills$title, "latin1", "ASCII", sub = "")
userskills$edu_dates <-
  str_replace_all(iconv(userskills$edu_dates, "latin1", "ASCII", sub = "-"),
                  "---", "-")

#DT::datatable(userskills, options = list(scrollX = TRUE))
head(userskills)
```

```
##   data_scientist                                     title
## 1    Josh Bersin Principal and Founder, Bersin by Deloitte
## 2    Josh Bersin Principal and Founder, Bersin by Deloitte
## 3    Josh Bersin Principal and Founder, Bersin by Deloitte
## 4    Josh Bersin Principal and Founder, Bersin by Deloitte
## 5    Josh Bersin Principal and Founder, Bersin by Deloitte
## 6    Josh Bersin Principal and Founder, Bersin by Deloitte
##              location
## 1 Oakland, California
## 2 Oakland, California
## 3 Oakland, California
## 4 Oakland, California
## 5 Oakland, California
## 6 Oakland, California
##                                                                edu_school
## 1 University of California, Berkeley - Walter A. Haas School of Business
## 2 University of California, Berkeley - Walter A. Haas School of Business
## 3 University of California, Berkeley - Walter A. Haas School of Business
## 4 University of California, Berkeley - Walter A. Haas School of Business
## 5 University of California, Berkeley - Walter A. Haas School of Business
## 6 University of California, Berkeley - Walter A. Haas School of Business
##      degree   edu_dates num_connections type_connections
## 1 MBA, 1988 1987 - 1988         438,275        followers
## 2 MBA, 1988 1987 - 1988         438,275        followers
## 3 MBA, 1988 1987 - 1988         438,275        followers
## 4 MBA, 1988 1987 - 1988         438,275        followers
## 5 MBA, 1988 1987 - 1988         438,275        followers
## 6 MBA, 1988 1987 - 1988         438,275        followers
##                    skill endorsements
## 1      Talent Management          699
## 2 Leadership Development          517
## 3             Leadership          476
## 4  Management Consulting          350
## 5        Human Resources          284
## 6             Consulting          263
```

```r
write.csv(userskills, "linkedin-profiles-skills.csv")
```

```r
skills <- userskills %>% distinct(skill = tolower(skill))
```

```r
base_url <- "http://service.dice.com/api/rest/jobsearch/v1/simple.json?skill="

dice.jobs <- data.frame(skills, job_listings = integer(nrow(skills)))

for (i in 1:nrow(dice.jobs)) {
  dice.jobs$job_listings[i] <-
    fromJSON(paste0(base_url, URLencode(skills$skill[i], reserved = TRUE)))$count
}

total_jobs <- fromJSON("http://service.dice.com/api/rest/jobsearch/v1/simple.json")$count

dice.jobs <- dice.jobs %>%
  mutate(prop_listings = job_listings/total_jobs) %>%
  arrange(desc(job_listings))

#DT::datatable(dice.jobs)
head(dice.jobs)
```

```
##                          skill job_listings prop_listings
## 1    agile project management        24234     0.3070199
## 2    new business development        21987     0.2785527
## 3             web development        21879     0.2771844
## 4        business development        21842     0.2767157
## 5             data management        21614     0.2738272
## 6 learning management systems        21306     0.2699251
```

```r
write.csv(dice.jobs, "dice-listings-skills.csv")
```