

Data 607 Project 3: Data Science Skills

Judd Anderman

October 21, 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/klrsao",
  "https://www.linkedin.com/in/abhisvni",
  "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-cao-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))
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
userskills$skill <- iconv(userskills$skill, "latin1", "ASCII", sub = "")

knitr::kable(userskills)
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

data_scientist	title
June Andrews	Data Scientist at Pinterest
June Andrews	Data Scientist at Pinterest
June Andrews	Data Scientist at Pinterest
June Andrews	Data Scientist at Pinterest
June Andrews	Data Scientist at Pinterest
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora
Wenwen Tao	Data Scientist at Quora

```
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      438,245      followers
## 2 MBA, 1988 1987 - 1988      438,245      followers
## 3 MBA, 1988 1987 - 1988      438,245      followers
## 4 MBA, 1988 1987 - 1988      438,245      followers
## 5 MBA, 1988 1987 - 1988      438,245      followers
```



```
## 6 MBA, 1988 1987 - 1988          438,245      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

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

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

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))

knitr::kable(dice.jobs)
```

skill	job_listings	prop_listings
agile project management	24114	0.3054648
new business development	21872	0.2770642
web development	21749	0.2755061
business development	21726	0.2752147
data management	21529	0.2727192
learning management systems	21172	0.2681969
project management	21118	0.2675129
business process management	21036	0.2664741
decision management systems	20948	0.2653594
software development	20793	0.2633959
program management	20173	0.2555420
product development	19492	0.2469155
android development	19454	0.2464341
customer relationship management (crm)	19208	0.2433179
content development	19063	0.2414811
creative concept development	19003	0.2407210
leadership development	18958	0.2401510
organizational development	18795	0.2380862
community development	18785	0.2379595
career development	18779	0.2378835
land development	18770	0.2377695
it management	18695	0.2368194
customer relationship management	18537	0.2348180
sales management	18536	0.2348053

skill	job_listings	prop_listings
management consulting	18464	0.2338932
sql server integration services (ssis)	18307	0.2319044
information management	17961	0.2275215
product management	17942	0.2272808
performance management	17932	0.2271541
knowledge management	17892	0.2266474
marketing management	17774	0.2251526
team management	17605	0.2230118
risk management	17575	0.2226318
learning management	17561	0.2224545
global human resources management	17544	0.2222391
time management	17533	0.2220998
executive management	17368	0.2200096
event management	17365	0.2199716
change management	17363	0.2199463
metadata management	17350	0.2197816
asset management	17349	0.2197689
portfolio management	17346	0.2197309
vendor management	17342	0.2196803
start-ups management	17340	0.2196549
people management	17337	0.2196169
talent management	17326	0.2194776
decision management	17325	0.2194649
brand management	17321	0.2194143
personnel management	17307	0.2192369
waste management	17300	0.2191482
management	17299	0.2191356
wealth management	17299	0.2191356
stormwater management	17299	0.2191356
microsoft sql server	14613	0.1851106
sql server 2000-2008	13376	0.1694409
social network analysis	13282	0.1682501
analysis services	12898	0.1633858
project planning	12889	0.1632718
apache http server	12581	0.1593702
statistical data analysis	12522	0.1586228
data analysis	12500	0.1583441
network security	12136	0.1537331
sql tuning	11992	0.1519090
sql	11866	0.1503129
business analysis	11817	0.1496922
java	11409	0.1445238
software as a service (saas)	11171	0.1415090
network architecture	10712	0.1356946
systems engineering	10426	0.1320716
data research	10399	0.1317296
customer analysis	10344	0.1310329
multivariate testing	10218	0.1294368
system architecture	9992	0.1265739
database systems	9645	0.1221783
business consulting	9612	0.1217603
requirements analysis	9532	0.1207469

skill	job_listings	prop_listings
amazon web services (aws)	9361	0.1185807
windows server	9338	0.1182894
software architecture	9290	0.1176813
software consulting	9224	0.1168453
computer security	9203	0.1165793
mathematical modelling and simulation	9043	0.1145525
business process design	9008	0.1141091
market opportunity analysis	8890	0.1126143
financial analysis	8861	0.1122470
information security	8828	0.1118289
market analysis	8809	0.1115883
interactive data visualization design	8745	0.1107775
time series analysis	8629	0.1093081
security clearance	8535	0.1081174
advanced statistical analysis	8515	0.1078640
internet security	8445	0.1069773
cluster analysis	8432	0.1068126
image analysis	8417	0.1066226
securities regulation	8382	0.1061792
mathematical analysis	8380	0.1061539
cyber security	8371	0.1060399
national security	8360	0.1059005
security	8351	0.1057865
homeland security	8351	0.1057865
regression analysis	8340	0.1056472
quantitative analysis	8335	0.1055838
decision analysis	8303	0.1051785
fraud analysis	8298	0.1051151
spatial analysis	8289	0.1050011
competitive analysis	8287	0.1049758
analysis	8275	0.1048238
numerical analysis	8275	0.1048238
software engineering	8275	0.1048238
jboss application server	8092	0.1025056
web services	7921	0.1003395
database design	7663	0.0970713
data quality	7594	0.0961972
html + css	7352	0.0931317
software design	7239	0.0917002
javascript	7217	0.0914215
data modeling	7039	0.0891667
data processing	6748	0.0854805
data integration	6725	0.0851891
linux	6599	0.0835930
social networking sites	6496	0.0822883
environmental engineering	6426	0.0814015
microsoft excel	6376	0.0807682
solution architecture	6359	0.0805528
social networking	6353	0.0804768
artificial neural networks	6330	0.0801855
deep neural networks	6326	0.0801348
business objects	6320	0.0800588

skill	job_listings	prop_listings
agile methodologies	6283	0.0795901
bayesian networks	6282	0.0795774
big data analytics	6278	0.0795267
neural networks	6277	0.0795141
networking	6277	0.0795141
html	6276	0.0795014
business planning	6270	0.0794254
data science	6261	0.0793114
reporting systems	6244	0.0790960
enterprise architecture	6241	0.0790580
data analytics	6214	0.0787160
biomedical engineering	6198	0.0785133
engineering mathematics	6174	0.0782093
business process improvement	6155	0.0779686
engineering	6153	0.0779433
business process	6140	0.0777786
agile methoden	6139	0.0777660
data migration	6105	0.0773353
sas programming	6030	0.0763852
data warehousing	5995	0.0759418
data governance	5989	0.0758658
data structures	5976	0.0757011
big data	5974	0.0756758
embedded systems	5951	0.0753845
quantitative data	5951	0.0753845
business writing	5940	0.0752451
data profiling	5928	0.0750931
data visualization	5925	0.0750551
data mining	5915	0.0749284
business analytics	5853	0.0741430
r programming	5753	0.0728763
multi agent systems	5701	0.0722176
business requirements	5693	0.0721162
system deployment	5584	0.0707355
gis programer	5584	0.0707355
business strategy	5553	0.0703428
mathematical programming	5553	0.0703428
market research	5534	0.0701021
marketing research	5534	0.0701021
new web technologies	5518	0.0698994
business intelligence	5498	0.0696461
architecture	5479	0.0694054
architectures	5479	0.0694054
programming	5468	0.0692660
business statistics	5468	0.0692660
social business	5467	0.0692534
expert systems	5466	0.0692407
business rules	5445	0.0689747
business transformation	5436	0.0688607
business alliances	5401	0.0684173
customer service	5395	0.0683413
distributed systems	5379	0.0681386

skill	job_listings	prop_listings
oracle rac	5375	0.0680880
oracle	5367	0.0679866
systems biology	5334	0.0675686
intelligent speech-enabled systems	5314	0.0673152
recommender systems	5311	0.0672772
knowledge-based systems	5311	0.0672772
professional services	5288	0.0669859
enterprise software	5278	0.0668592
advanced excel	5092	0.0645031
statistical consulting	5033	0.0637557
legal research	5000	0.0633376
educational research	4987	0.0631730
research	4974	0.0630083
web analytics	4945	0.0626409
tableau software	4917	0.0622862
start-up consulting	4870	0.0616909
strategic consulting	4866	0.0616402
spatial databases	4859	0.0615515
databases	4847	0.0613995
web mining	4828	0.0611588
consulting	4812	0.0609561
hr software	4800	0.0608041
software deployment	4785	0.0606141
software documentation	4748	0.0601454
user experience	4688	0.0593854
adaptive software	4515	0.0571939
web 2.0	4489	0.0568645
customer support	4469	0.0566112
semantic web	4454	0.0564212
python	4453	0.0564085
sap erp	4346	0.0550531
sap implementation	4273	0.0541283
mobile applications	4263	0.0540017
object oriented design	4123	0.0522282
c#	4100	0.0519369
unix	3786	0.0479593
css	3719	0.0471105
sap	3710	0.0469965
microsoft office	3368	0.0426642
water distribution design	3209	0.0406501
organizational design	3141	0.0397887
design patterns	3136	0.0397254
pcb design	3136	0.0397254
lift station design	3121	0.0395354
wastewater treatment design	3110	0.0393960
customer analytics	3076	0.0389653
peoplesoft crm	3032	0.0384079
enterprise it strategy	2996	0.0379519
financial modeling	2890	0.0366092
dynamic mathematical modeling	2865	0.0362925
spring framework	2623	0.0332269
customer loyalty	2537	0.0321375

skill	job_listings	prop_listings
customer satisfaction	2532	0.0320742
cross-cultural communication skills	2502	0.0316942
genetic algorithms	2486	0.0314915
c++	2455	0.0310988
mathematical modeling	2447	0.0309974
advertising sales	2434	0.0308328
statistical modeling	2433	0.0308201
xml	2430	0.0307821
it operations	2422	0.0306808
sales advisor	2411	0.0305414
siebel crm	2337	0.0296040
analytical skills	2330	0.0295153
predictive modeling	2298	0.0291100
decision modeling	2293	0.0290466
internet of things	2288	0.0289833
working with children	2285	0.0289453
modeling	2264	0.0286793
decision support	2242	0.0284006
social crm	2235	0.0283119
scrum	2191	0.0277546
it strategy	2176	0.0275645
crm	2158	0.0273365
information assurance	2146	0.0271845
technology integration	2090	0.0264751
teaching english as a foreign language	2031	0.0257277
technical writing	2021	0.0256011
microsoft office suite	2013	0.0254997
perl	1974	0.0250057
water quality	1969	0.0249424
technical training	1922	0.0243470
technical recruiting	1881	0.0238276
human computer interfaces	1841	0.0233209
technical presentations	1822	0.0230802
cloud computing	1817	0.0230169
mobile marketing	1806	0.0228776
mysql	1794	0.0227255
ibm db2	1766	0.0223709
sdlc	1752	0.0221935
dashboard metrics	1717	0.0217501
strategic financial planning	1710	0.0216615
php	1603	0.0203060
hadoop	1546	0.0195840
accounting	1532	0.0194067
financial controlling	1515	0.0191913
git	1495	0.0189380
c	1485	0.0188113
digital image processing	1481	0.0187606
apache spark	1454	0.0184186
product marketing	1446	0.0183172
algorithms	1414	0.0179119
digital signal processing	1408	0.0178359
mobile internet	1360	0.0172278

skill	job_listings	prop_listings
computer graphics	1353	0.0171392
search engine technology	1346	0.0170505
high performance computing	1321	0.0167338
natural language processing	1293	0.0163791
rest	1277	0.0161764
html5	1271	0.0161004
play framework	1260	0.0159611
blended learning solutions	1250	0.0158344
marketing communications	1237	0.0156697
apache pig	1221	0.0154671
intelligent natural user interfaces	1214	0.0153784
intelligence community	1213	0.0153657
ibm watson	1209	0.0153150
competitive intelligence	1204	0.0152517
computational intelligence	1201	0.0152137
artificial intelligence	1187	0.0150364
intelligence	1187	0.0150364
financial markets	1179	0.0149350
human computer interaction	1177	0.0149097
apache storm	1149	0.0145550
computer science	1142	0.0144663
image processing	1141	0.0144536
visual basic	1132	0.0143396
semantic technologies	1128	0.0142890
science communication	1124	0.0142383
disruptive technologies	1117	0.0141496
nosql	1114	0.0141116
strategic planning	1097	0.0138963
cissp	1094	0.0138583
lead optimization	1084	0.0137316
executive reporting	1083	0.0137189
etl	1082	0.0137063
signal processing	1071	0.0135669
succession planning	1068	0.0135289
apache	1067	0.0135163
business-intelligence	1066	0.0135036
automation	1064	0.0134782
pl/sql	1055	0.0133642
controlling budgets	1050	0.0133009
process improvement	1049	0.0132882
visual studio	1045	0.0132376
stochastic processes	1031	0.0130602
crystal reports	1029	0.0130349
computer vision	1020	0.0129209
performance tuning	1012	0.0128195
performance improvement	1009	0.0127815
integration	1008	0.0127689
key performance indicators	985	0.0124775
big 4	966	0.0122368
enterprise 2.0	947	0.0119961
performance measurement	942	0.0119328
enterprise collaboration	941	0.0119201

skill	job_listings	prop_listings
product evangelism	926	0.0117301
peoplesoft	898	0.0113754
information retrieval	837	0.0106027
erp	811	0.0102734
knowledge discovery	800	0.0101340
marketing strategy	749	0.0094880
corporate communications	747	0.0094626
social media marketing	718	0.0090953
access	716	0.0090700
db2	704	0.0089179
communication	692	0.0087659
direct marketing	690	0.0087406
sas	686	0.0086899
mongodb	670	0.0084872
integrated marketing	660	0.0083606
kpi implementation	657	0.0083226
financial economics	655	0.0082972
digital media strategy	652	0.0082592
streaming analytics	651	0.0082466
proposal writing	651	0.0082466
online marketing	647	0.0081959
legal writing	646	0.0081832
book writing	633	0.0080185
ssis	630	0.0079805
cross-functional team leadership	621	0.0078665
quantitative analytics	620	0.0078539
writing	619	0.0078412
team leadership	608	0.0077019
text analytics	598	0.0075752
open source	596	0.0075498
soa	594	0.0075245
predictive analytics	591	0.0074865
grassroots marketing	581	0.0073598
marketing	581	0.0073598
simulation	571	0.0072332
simulations	571	0.0072332
analytics	563	0.0071318
prescriptive analytics	563	0.0071318
scientific computing	544	0.0068911
digital strategy	542	0.0068658
value based selling	541	0.0068531
telecommunications	541	0.0068531
ssrs	537	0.0068025
science education	535	0.0067771
statistical modelling	525	0.0066505
physical sciences	517	0.0065491
hr strategy	491	0.0062198
matlab	490	0.0062071
digital media	475	0.0060171
science	457	0.0057891
e-commerce	450	0.0057004
objective-c	448	0.0056751

skill	job_listings	prop_listings
troubleshooting	447	0.0056624
physics	446	0.0056497
trading strategies	445	0.0056370
spark	434	0.0054977
quantitative finance	430	0.0054470
saas	422	0.0053457
text mining	422	0.0053457
corporate finance	417	0.0052824
tableau	414	0.0052444
scala	407	0.0051557
t-sql	380	0.0048137
finance	362	0.0045856
environmental awareness	362	0.0045856
hr transformation	354	0.0044843
uml	346	0.0043830
tomcat	343	0.0043450
hive	340	0.0043070
cognos	338	0.0042816
new media	336	0.0042563
organizational learning	321	0.0040663
machine learning	319	0.0040409
online training	318	0.0040283
stored procedures	317	0.0040156
deep learning	313	0.0039649
internet recruiting	311	0.0039396
federal government	310	0.0039269
hidden markov models	307	0.0038889
probabilistic modelling	305	0.0038636
probabilistic models	305	0.0038636
chinese language teaching	304	0.0038509
r	303	0.0038383
derivatives trading	293	0.0037116
collaborative learning	290	0.0036736
blended learning	283	0.0035849
vulnerability assessment	281	0.0035596
banking	280	0.0035469
tfs	279	0.0035342
imo (international mathematical olympiads)	275	0.0034836
thought leadership	270	0.0034202
proprietary trading	261	0.0033062
corporate strategy formulation	258	0.0032682
creative problem solving	257	0.0032556
inspiring leadership	256	0.0032429
leadership	249	0.0031542
government	248	0.0031415
advanced linear algebra	247	0.0031289
high availability	237	0.0030022
training	234	0.0029642
statistical inference	233	0.0029515
statistical arbitrage	230	0.0029135
international policy	228	0.0028882
electronics	227	0.0028755

skill	job_listings	prop_listings
metadata standards	225	0.0028502
pattern recognition	218	0.0027615
virtual personal assistants	218	0.0027615
siebel	214	0.0027109
parallel computing	211	0.0026728
ejb	210	0.0026602
human resources	206	0.0026095
datastage	206	0.0026095
industrial internet	205	0.0025968
human capital	204	0.0025842
context-aware computing	198	0.0025082
recruiting	196	0.0024828
go-to-market strategy	194	0.0024575
hindi	193	0.0024448
strategy	192	0.0024322
eclipse	185	0.0023435
ssas	184	0.0023308
social media measurement	180	0.0022802
salesforce.com	179	0.0022675
social media	174	0.0022041
higher education	170	0.0021535
programmatic media buying	166	0.0021028
erwin	164	0.0020775
cluster	161	0.0020395
django	157	0.0019888
positioning	153	0.0019381
executive coaching	151	0.0019128
github	148	0.0018748
governance	142	0.0017988
olap	141	0.0017861
arcgis online	141	0.0017861
search	139	0.0017608
credit derivatives	137	0.0017355
logistic regression	136	0.0017228
optimization	135	0.0017101
public speaking	133	0.0016848
gis	131	0.0016594
general awesomeness	127	0.0016088
dashboards	126	0.0015961
dashboard	126	0.0015961
online publishing	122	0.0015454
mapreduce	122	0.0015454
robotics	122	0.0015454
opengl	118	0.0014948
game theory	117	0.0014821
online advertising	117	0.0014821
venture capital	113	0.0014314
jpa	113	0.0014314
six sigma	107	0.0013554
air force	107	0.0013554
presentations	105	0.0013301
evIEWS	102	0.0012921

skill	job_listings	prop_listings
elasticsearch	98	0.0012414
multivariate statistics	96	0.0012161
real-time bidding	94	0.0011907
statistics	93	0.0011781
linear regression	93	0.0011781
hdfs	92	0.0011654
balanced scorecard	90	0.0011401
seo	90	0.0011401
consumer behavior	90	0.0011401
pre-sales	89	0.0011274
mergers & acquisitions	89	0.0011274
multithreading	89	0.0011274
applied mathematics	88	0.0011147
human-computer interaction	87	0.0011021
cto	87	0.0011021
dod	85	0.0010767
consumer behaviour	84	0.0010641
fortran	83	0.0010514
corporate university	80	0.0010134
coaching	78	0.0009881
anomaly detection	77	0.0009754
strategic thinking	75	0.0009501
pki	75	0.0009501
quantitative investing	75	0.0009501
star schema	75	0.0009501
gov 2.0	73	0.0009247
regression	72	0.0009121
iot	69	0.0008741
investments	68	0.0008614
strategic partnerships	61	0.0007727
tax law	61	0.0007727
inspiring people	60	0.0007601
equities	58	0.0007347
arcgis	57	0.0007220
command	53	0.0006714
mentoring	53	0.0006714
published author	52	0.0006587
acl	52	0.0006587
employee engagement	51	0.0006460
activity context representation	51	0.0006460
current affairs	50	0.0006334
forecasting	49	0.0006207
organizational effectiveness	47	0.0005954
community engagement	46	0.0005827
decision trees	45	0.0005700
akka	45	0.0005700
nginx	45	0.0005700
omniture	43	0.0005447
graph theory	43	0.0005447
future trends	42	0.0005320
target identification	41	0.0005194
collaborative filtering	39	0.0004940

skill	job_listings	prop_listings
nlp	39	0.0004940
d3.js	39	0.0004940
spss	37	0.0004687
publishing	36	0.0004560
editing	36	0.0004560
computational biology	36	0.0004560
gwt	34	0.0004307
mathematics	33	0.0004180
sem	33	0.0004180
mathematica	33	0.0004180
lucene	32	0.0004054
brand loyalty	30	0.0003800
english	30	0.0003800
kpi	29	0.0003674
community outreach	28	0.0003547
bioinformatics	27	0.0003420
analytik	25	0.0003167
semiconductors	24	0.0003040
fx	24	0.0003040
marketo	24	0.0003040
fraud	23	0.0002914
conference speaking	23	0.0002914
mandarin	23	0.0002914
scalability	22	0.0002787
opencv	22	0.0002787
technological innovation	21	0.0002660
dax	21	0.0002660
microcontrollers	20	0.0002534
innovation	19	0.0002407
mpp	19	0.0002407
editor	18	0.0002280
automata theory	18	0.0002280
e-learning	16	0.0002027
segmentation	16	0.0002027
navy	16	0.0002027
economics	16	0.0002027
blockchain	15	0.0001900
eda	15	0.0001900
neo4j	15	0.0001900
dts	15	0.0001900
motivational speaking	14	0.0001773
gnu octave	13	0.0001647
military	12	0.0001520
teamwork	12	0.0001520
generalists	12	0.0001520
teaching	10	0.0001267
ppc	10	0.0001267
paper craft	10	0.0001267
due diligence	10	0.0001267
french	10	0.0001267
hydraulics	9	0.0001140
jmp	8	0.0001013

skill	job_listings	prop_listings
keen io	8	0.0001013
wastewater treatment	8	0.0001013
vim	8	0.0001013
army	7	0.0000887
hospitality	7	0.0000887
arcpy	5	0.0000633
julia	5	0.0000633
scikit-learn	5	0.0000633
start-ups	4	0.0000507
blogging	4	0.0000507
context	4	0.0000507
ontology	4	0.0000507
sewer	4	0.0000507
benchmarking	3	0.0000380
svm	3	0.0000380
monetization	2	0.0000253
weka	2	0.0000253
econometrics	2	0.0000253
astronomy	2	0.0000253
knime	2	0.0000253
minitab	2	0.0000253
music	1	0.0000127
poetry	1	0.0000127
backtesting	1	0.0000127
entrepreneurship	0	0.0000000
visionary	0	0.0000000
latex	0	0.0000000
endorsements	0	0.0000000
astrophysics	0	0.0000000
cosmology	0	0.0000000
people-oriented	0	0.0000000
invention	0	0.0000000
gamification	0	0.0000000
watercolor	0	0.0000000
maschinelles lernen	0	0.0000000
softwareentwicklung	0	0.0000000
knstliche intelligenz	0	0.0000000
produktentwicklung	0	0.0000000
automobilindustrie	0	0.0000000
datenbanken	0	0.0000000
prozessverbesserung	0	0.0000000
teamentwicklung	0	0.0000000
unternehmensfhrung	0	0.0000000
simulationen	0	0.0000000
algorithmen	0	0.0000000
computerwissenschaft	0	0.0000000
vernderungsmanagement	0	0.0000000
kontinuierliche verbesserung	0	0.0000000
macroeconomics	0	0.0000000

```
head(dice.jobs)
```

```
##               skill job_listings prop_listings
## 1 agile project management      24114      0.3054648
## 2 new business development      21872      0.2770642
## 3 web development              21749      0.2755061
## 4 business development         21726      0.2752147
## 5 data management              21529      0.2727192
## 6 learning management systems   21172      0.2681969
```

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