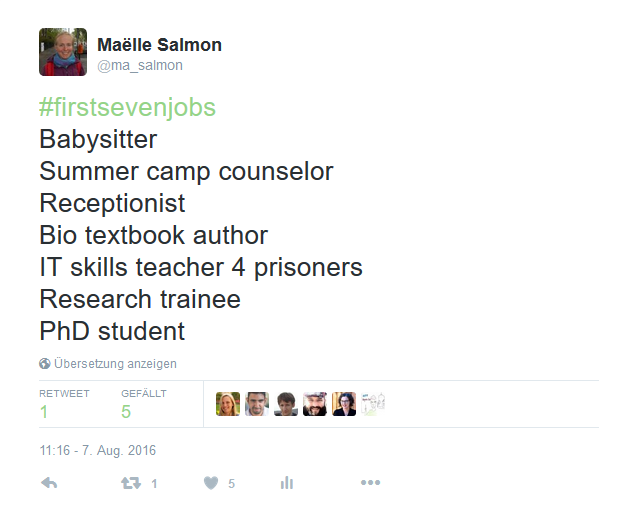
Analyzing #first7jobs tweets with Monkeylearn and R

M. Salmon

August 25, 2016

# Introduction

Have you tweeted about your "#firstsevenjobs"? I did!



My first seven jobs.

“#firstsevenjobs" and “#first7jobs" tweets initial goal was to provide a short description of the 7 first activities they were paid for. It was quite fun to read them in my timeline! Of course the hashtag was also used by spammers, for making jokes, and for advertising for analyses in R, so not all the tweets contain the 7 descriptions.

However, I am confident quite a lot of “#firstsevenjobs" and “#first7jobs" actually describe first jobs, so I decided to use them as example of text analysis in R, starting from querying Twitter API with the rtweet package, then cleaning the tweets a bit, and then using the monkeylearn package to classify the jobs in an industry.

# Getting the tweets

I used the [rtweet R package](https://github.com/mkearney/rtweet/) for getting tweets via the Twitter API, searching for both “#firstsevenjobs" and “#first7jobs" hashtags and then keeping only unique non-retweeted tweets in English. I got 4858 tweets. This does not mean there were only that few tweets produced with the hashtags, but the Twitter API does not output aaall the tweets. You'd have to pay for it. But hey that's a good number of tweets to start with, so I won't complain. Here is part of the table I got:

|  |  |
| --- | --- |
| status\_id | text |
| 765226304073404416 | What Were Your #FirstSevenJobs?: #firstsevenjobsphotography salespot washerbartenderurban plann... <https://t.co/wwGp8NCohG> #Architectbiz |
| 764629565431947264 | The unexpected joys of #FirstSevenJobs <https://t.co/wV8XeFVlv8> |
| 764104419185229824 | My piece on #firstsevenjobs <https://t.co/Il3a2Wrm0I> |
| 765643154964025344 | #first7jobs: milkshake maker, national anthem singer, babysitter, phone bank caller, tutor, camp counselor, bartender (<- badly) |
| 765407468499374080 | @BenSPLATT Oh, I thought you were posting your #firstsevenjobs! |
| 766334362447151104 | 13 Entrepreneurs and CEOs Share Their #First7Jobs #jobseekers #advice <https://t.co/k1hhSFKznH> |
| 764296241698238468 | Babysitter, educational video actor, Little League umpire, sales clerk, archery instructor, security guard, audiobook narrator. #first7jobs |
| 763868874211078144 | #firstsevenjobs 1. Landscaper 2. passenger train car attendant 3. Mail sorter at Post office 4. pt Evenings/weekends/cruiser/prod 1090 CHEC |
| 763922164655415296 | #firstsevenjobs fashion intern, retail sales, rec coordinator,receptionist, teacher, school administrator |
| 766299512205811712 | 1. Golf course kitchen 2. Whole foods 3. Dean and Deluca 4. Hyatt 5. Gotts 6. Hillstone 7. Cheesecake Factory #firstsevenjobs |

# Parsing the tweets

So you see, part of them contains actual job descriptions, others don't... I mean, even I polluted the hashtag for advertising my own analysis! Among those that do describe jobs, some use commas or new lines between descriptions, or number them, or simply use spaces... Therefore, parsing tweets for getting 7 job descriptions per tweet was a little challenge.

I counted the number of possible separators for finding which one I should probably use to cut the tweet into 7 part. This yielded tweets cut in several parts... sometimes less than 7, sometimes more. I could not parse tweets whose descriptions were separated only by spaces because words inside a description are separated by spaces too so I could not make the difference. Besides, some people have tweeted about less or more than 7 packages. For instance one tweet says I have not had seven jobs yet but so far...- Accounts Assistant- Executive PA- Social Media Lead,yoga instructor?#FirstSevenJobs". I did my best to remove tweet parts that were something like "Here are my #firstsevenjobs", in order to keep only the job descriptions. At the end I kept only the tweets that had exactly 7 parts.

Out of 4858 I got 1637 tweets, that is 11459 job descriptions. That is *a lot*. Here is an excerpt of the table:

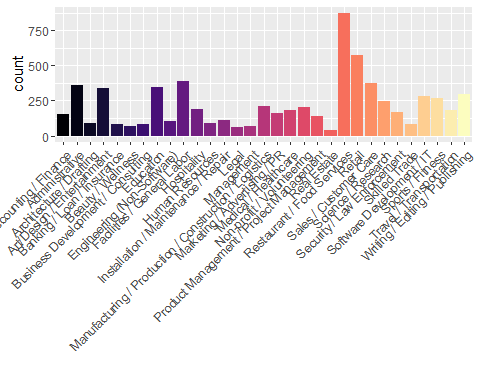
|  |  |
| --- | --- |
| status\_id | wordsgroup |
| 763505013675229193 | Shopping bag |
| 763505013675229193 | Shopping assistant |
| 763505013675229193 | Housekeeper |
| 763505013675229193 | Cashier at Empik |
| 763505013675229193 | Fast food worker |
| 763505013675229193 | Microsoft's consultant |
| 763505013675229193 | Cashier at Sport shop |
| 763511170196135936 | Dish Pig |
| 763511170196135936 | Toy Packer |
| 763511170196135936 | Asian Chef |
| 763511170196135936 | Bike Fitter |
| 763511170196135936 | Bike Shop Dude |
| 763511170196135936 | Beard Grower |
| 763511170196135936 | Sports Therapist |
| 763512991731945472 | babysitter |
| 763512991731945472 | busperson |
| 763512991731945472 | camp counselor |
| 763512991731945472 | secretary/clerk |
| 763512991731945472 | graduate assistant |
| 763512991731945472 | college prof |
| 763512991731945472 | full time writer |

# Summarizing the information by assigning an industry to each job

It would take a long time to read them all the tweets, although I did end up reading a lot of tweets while preparing this post. I wanted to have a general idea of what people did in their life. I turned to machine learning to help me get some information out of the tweets. I'm the creator and maintainer of an [R package called monkeylearn](https://github.com/ropenscilabs/monkeylearn), which is part of the [rOpenSci project](http://ropensci.org/), that allows to use existing Monkeylearn classifiers and extractors, so I knew that Monkeylearn had a [cool job classifier](https://app.monkeylearn.com/main/classifiers/cl_i7vMzUB7/). I sent all the nrow(first7packages\_parsed) job descriptions to Monkeylearn API.

Monkeylearn's job classifier assigns an industry out of 31 possible industries and a probability to each job description. ASK FEDERICO FOR MORE DETAILS ABOUT THE CLASSIFIER, TRAINING DATA, PROBABILITY MEANINING, ETC.

I decided to keep only job descriptions for which the probability given by the classifier was higher than 50%. This corresponds to 6801 job descriptions out of the initial 11459 job descriptions. Jobs for which we predicted a category with a probability higher than 0.5 are divided as follows among industries:



The most important categories are Restaurant/Food services and Retail. Usual first jobs?

# What are jobs by industry?

In this work I did not try to change the classifier so I trusted it, but I was curious to know which jobs ended up in each category. I had a glance at descriptions by industry but this can take a while given the number of jobs in some categories. Thanksfully [Federico Pascual](https://twitter.com/FedericoPascual) told me I could use [Monkeylearn's keyword extractor](https://app.monkeylearn.com/main/extractors/ex_y7BPYzNG/) on all job descriptions of each category to find dominant patterns. Such a nice idea, and something my package supports. I chose to get 5 keywords by industry. Here is the result:

|  |  |
| --- | --- |
| label | keyword |
| Accounting / Finance | Accounting clerk, financial analyst, account manager, Bookkeeper, Accountant |
| Administrative | office manager, front desk, office assistant, receptionist, assistant |
| Architecture / Drafting | Land surveyor, surveyor, Job, applications, Landscaper |
| Art/Design / Entertainment | House painter, sandwich artist, web designer, Graphic Designer, designer |
| Banking / Loan / Insurance | Private tutor, University, insurance, bank teller, teller |
| Beauty / Wellness | hot dog vendor, Physical Therapy Aide, dog sitter, Dog walker, Dog |
| Business Development / Consulting | Business Owner, Mgmt consultant, strategist, analyst, consultant |
| Education | high school teacher, substitute teacher, library assistant, Math tutor, teacher |
| Engineering (Non-Software) | Audio Engineer, Engineer intern, network engineer, sales engineer, engineer |
| Facilities / General Labor | factory worker, Grocery Bagger, bagger, Janitor, Warehouse |
| Hospitality | Gas station attendant, Gas Station, Kitchen porter, Stock boy, Hostess |
| Human Resources | event coordinator, Recruitment Consultant, Manager, Recruiter, coordinator |
| Installation / Maintenance / Repair | golf course maintenance, ice cream shop, shop assistant, maintenance, shop |
| Legal | Law Office Runner, corporate filth monkey, Law clerk, Paralegal, Law firm |
| Management | retail assistant manager, assistant manager, staff, manager, Director |
| Manufacturing / Production / Construction / Logistics | park ride operator, Assembly line worker, construction laborer, construction worker, assembly |
| Marketing / Advertising / PR | Market researcher, Social Media, Marketing Intern, intern, Marketing |
| Medical / Healthcare | ice cream scooper, Paperboy, Waiter, Waitress, Babysitter |
| Non-profit / Volunteering | student assistant, Orientation Leader, social worker, camp counselor, Camp |
| Product Management / Project Management | Program manager, Programming Intern, Production Manager, project manager, manager |
| Real Estate | Real Estate Broker, mortgage broker, Actor Commercials, trainee, real estate |
| Restaurant / Food Services | fast food, Barista, Bartender, Dishwasher, clerk |
| Retail | grocery clerk, grocery store, retail sales, Grocery, cashier |
| Sales / Customer Care | Customer Service Rep, Sales assistant, customer service, Sales Associate, sales |
| Science / Research | tech support, lab tech, research assistant, Tech, research |
| Security / Law Enforcement | office temp, Office admin, Security guard, Security, office |
| Skilled Trade | Computer Repair Tech, Manufacturer, repair, Carpenter, summer |
| Software Development / IT | Software Engineer, Web Developer, data entry, Programmer, Developer |
| Sports / Fitness | Gymnastics coach, Soccer Referee, Swim instructor, Lifeguard, instructor |
| Travel / Transportation | delivery driver, Bus Boy, Paper route, Newspaper delivery, Pizza delivery |
| Writing / Editing / Publishing | Freelance Writer, Copywriter, reporter, writer, intern |

Some keywords look logical to our human understanding, some others don't, for instance dog-sitting as a beauty/wellness job... But wait having a dog is good for your health so people caring for your dog help your wellness, right?



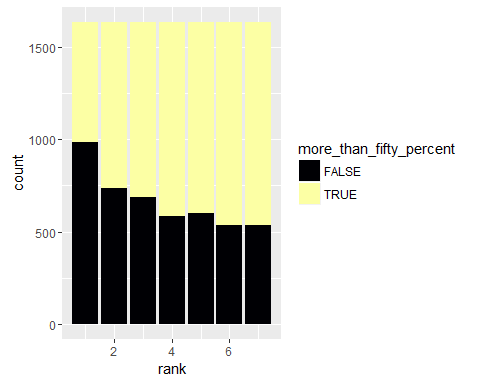
No comment, say hi to my sibling Mowgli. He's quite a beauty.

FEDERICO -> YOU MIGHT WANNA ADD A FEW LINES ABOUT HOW TRAINING A CLASSIFIER CAN IMPROVE IT ETC.

# Rank and industry prediction

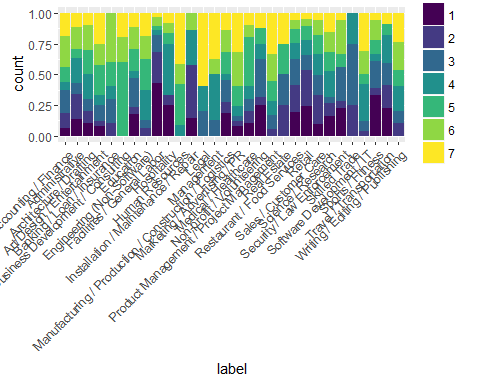
Now let's focus on exhaustive tweets only, id est tweets for which we could predict an industry for all 7 jobs with a probability higher than 50%.

This corresponds to 95. Remember that we started out with 4858 from Twitter API, of which we could parse 11459. We lost a lot along the way, but remember that computers still do not read as well as humans, and I also did not choose to update the classifier, which one could do for real life applications. For each tweet the order of jobs gave their chronological order (well I hope so). Were job descriptions harder to classify depending on their rank?

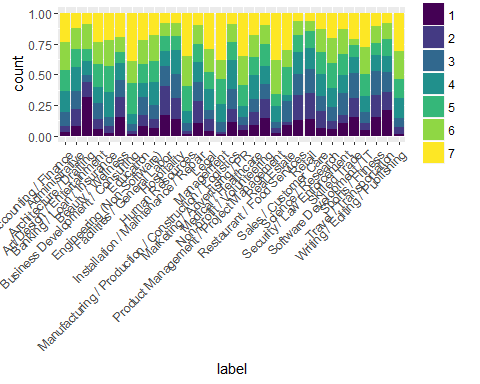


It seems to me that later job descriptions are easier to classify. Maybe because first jobs can be something like "Daddy's knitting helper" or "Serial Lego builder" while later jobs are adult jobs?

Now let's go back to tweets for which all industries could be predicted. As biased as it is, our sample of 95 tweets is still a nice playground for looking at life trajectories. For instance, are some categories rather first first jobs than late first jobs?



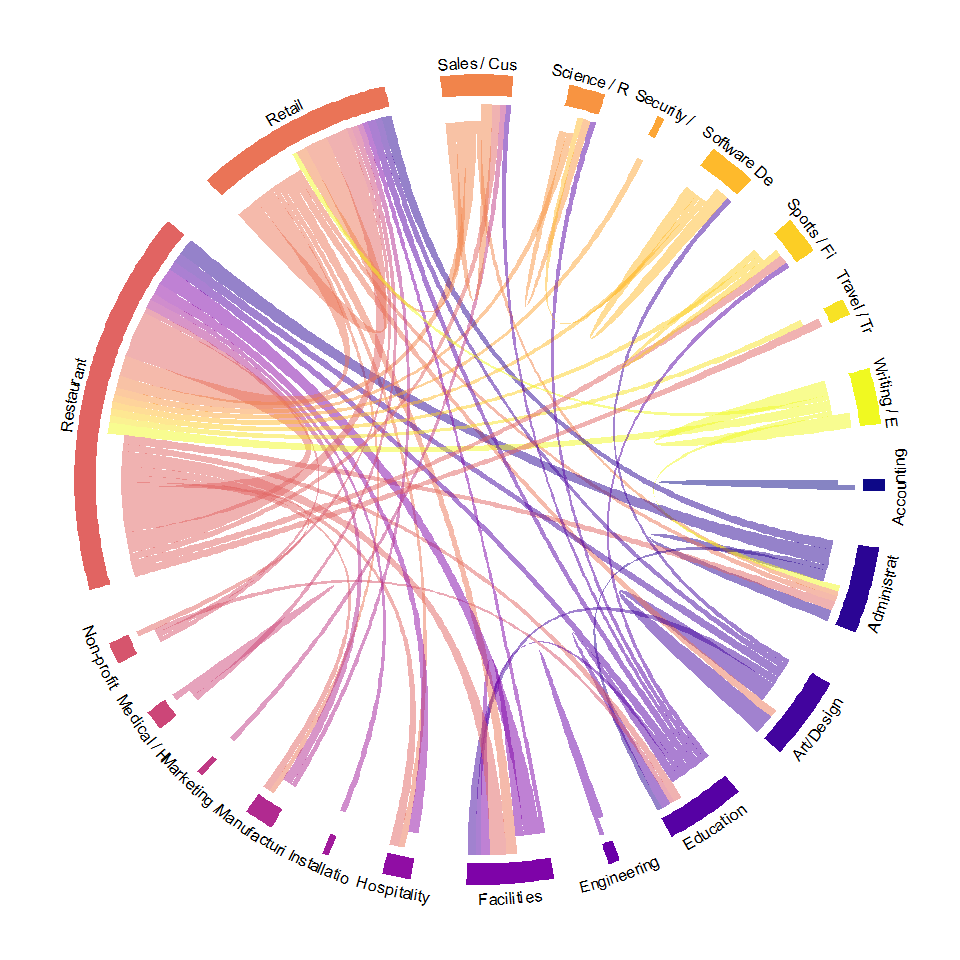
We can do the same graph for all job descriptions, even the ones in incompletely predicted tweets:



In both cases I'd tend to say that some industries such as Business Development / Consulting are not first-entry jobs, while Non-Profit / Volunteering are. Not a real surprise I guess?

# Transitions between industries

I've said I wanted to look at life trajectories. This dataset won't give me any information about the level of the job of course, e.g. whether you start as a clerk and end up leading your company, but I can look at how people move from one category to another. [My husband](http://dacornu.github.io/) gave me a great idea of a circle graph he had seen in a newspaper. For this I used only job descriptions for which an industry was predicted with a probability higher than 0.5. I kept only possible transitions where there were present more than 15 times in the data, otherwise we'll end up looking at a [hairball](https://twitter.com/drob/status/768485328244056065). I chose not to keep only complete tweets, in order to observe enough transitions.



ADD INTERPRETATION.

On this circle you see different industries, and the transition between them. [David Robinson](http://varianceexplained.org/) suggested I found the most common transitions and showed them in directed graphs but I'll keep this idea for later, since this post is quite long already, ah!

ADD CONCLUDING WORDS.

# Acknowledgements

Note that my whole code is [in this Github repo](https://github.com/masalmon/first_7_jobs). I used those R packages: rtweet, dplyr, tidyr, ggplot2, stringr, circlize and of course monkeylearn. Thanks a lot to their authors, and obviously thanks to people whose tweets I used... I might be a *little bit* more grateful to people who used separators and only posted 7 descriptions in their tweet. If you want to read another "#first7" analysis in R, I highly recommend [David Robinson's post](http://varianceexplained.org/r/seven-fav-packages/) about the "7FavPackages" hashtag.