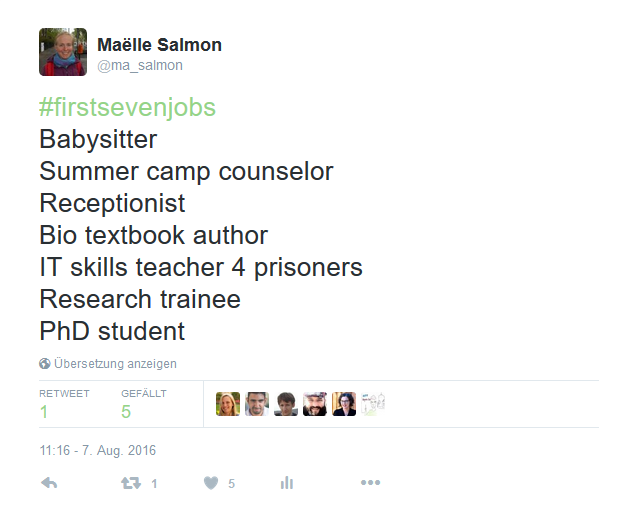
Analyzing #first7jobs tweets with Monkeylearn and R

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 commenting the hashtag, so not all the tweets contain 7 job 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 with Monkeylearn, 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 a field of work

# 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, sent between the 2016-08-10 and the 2016-08-20. This does not mean there were only that few tweets produced with the hashtags, but the Twitter API does not output *all* 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 parts. 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 jobs. 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 | rank |
| 763505013675229193 | Shopping bag | 1 |
| 763505013675229193 | Shopping assistant | 2 |
| 763505013675229193 | Housekeeper | 3 |
| 763505013675229193 | Cashier at Empik | 4 |
| 763505013675229193 | Fast food worker | 5 |
| 763505013675229193 | Microsoft's consultant | 6 |
| 763505013675229193 | Cashier at Sport shop | 7 |
| 763511170196135936 | Dish Pig | 1 |
| 763511170196135936 | Toy Packer | 2 |
| 763511170196135936 | Asian Chef | 3 |
| 763511170196135936 | Bike Fitter | 4 |
| 763511170196135936 | Bike Shop Dude | 5 |
| 763511170196135936 | Beard Grower | 6 |
| 763511170196135936 | Sports Therapist | 7 |
| 763512991731945472 | babysitter | 1 |
| 763512991731945472 | busperson | 2 |
| 763512991731945472 | camp counselor | 3 |
| 763512991731945472 | secretary/clerk | 4 |
| 763512991731945472 | graduate assistant | 5 |
| 763512991731945472 | college prof | 6 |
| 763512991731945472 | full time writer | 7 |

Rank is the rank of the jobs in the tweet, which should be the chronological rank too. For instance, for the first tweet, the first job is "shopping bag", the second "shopping assistant", etc.

# Monkeylearn magic: Summarizing the information by assigning a field to each job

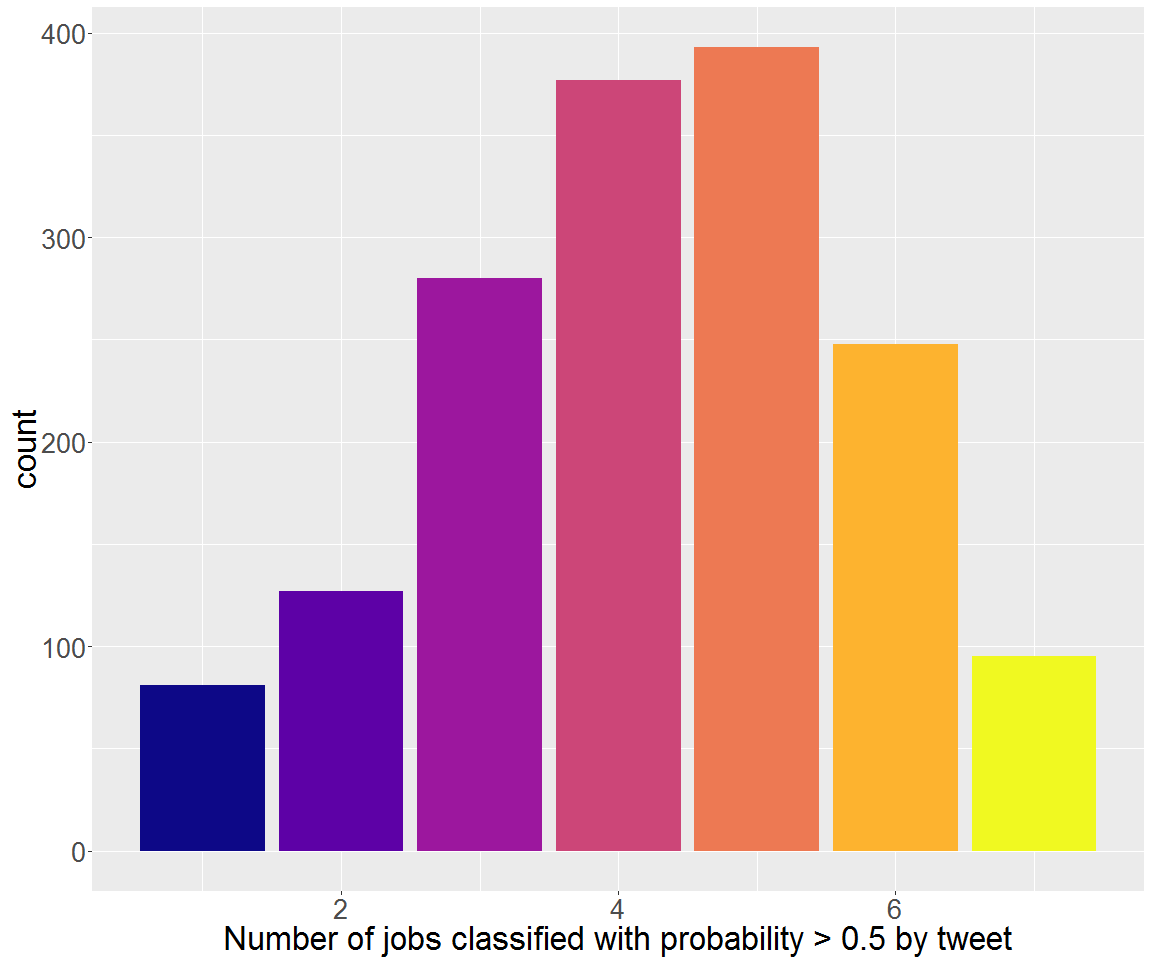
It would take a long time to read them all the tweets, although I did end up reading a lot of tweets while preparing this post. I wanted to have a general idea of what people did in their life. I turned to machine learning to help me get some information out of the tweets. I'm the creator and maintainer 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 11459 job descriptions to Monkeylearn API.

Monkeylearn's job classifier assigns a field out of 31 possible fields (called *industries*) and a probability to each job description. The algorithm uses a [supported vector machines (SVM) model](https://en.wikipedia.org/wiki/Support_vector_machine) for predicting the category of a job. It was originally developped by a client of Monkeylearn as a public module, and was then further developped by the Monkeylearn team, still as a public module -- I really like this collaborative effort. As a Monkeylearn user one could fork the classifier and play with catergories definitions, add or improve data for training the model, etc. With my package one can only use existing models, so that a possible workflow would be to develop modules outside of R and then to use them in R in production. If you wish to know more about classifiers, you can have a look at [Monkeylearn knowledge base](http://docs.monkeylearn.com/knowledge-base/) or even take a Machine learning MOOC such as [this one](https://www.coursera.org/learn/machine-learning). But I disgress, I've been using the jobs classifier as it is, and it was quite fun and above all promising.

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.

# Tweets coverage by the classifier

I then wondered how many jobs could be classified with a probability superior to 50% inside each tweet.



For each 11459 of the tweets I sent to the jobs classifier, I got a field with a probability higher to 0.5 for on average 4-5 job descriptions. We might want even more, and as I'll point it out later, we could get more if we put some effort into it and take full advantage of Monkeylearn possibilities!

# What are jobs by field?

In this work I used the classifier as it was without modifying it, but I was curious to know which jobs ended up in each category. I had a glance at descriptions by field but this can take a while given the number of jobs in some categories. Thanksfully [Federico Pascual](https://twitter.com/FedericoPascual) reminded 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 field. 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 |

For some categories, keywords seem natural to us, for some others we might be more surprised. For instance, the algorithm was trained with data wich included "'Pet Stylist', 'Dog Trainer', 'Pet Stylist (DOG GROOMER)'" for the Wellness/Beauty category, and no "Dog sitter", so that's why here dog sitting is a wellness job. But wait having a dog is good for your health so people caring for your dog help your wellness, right?

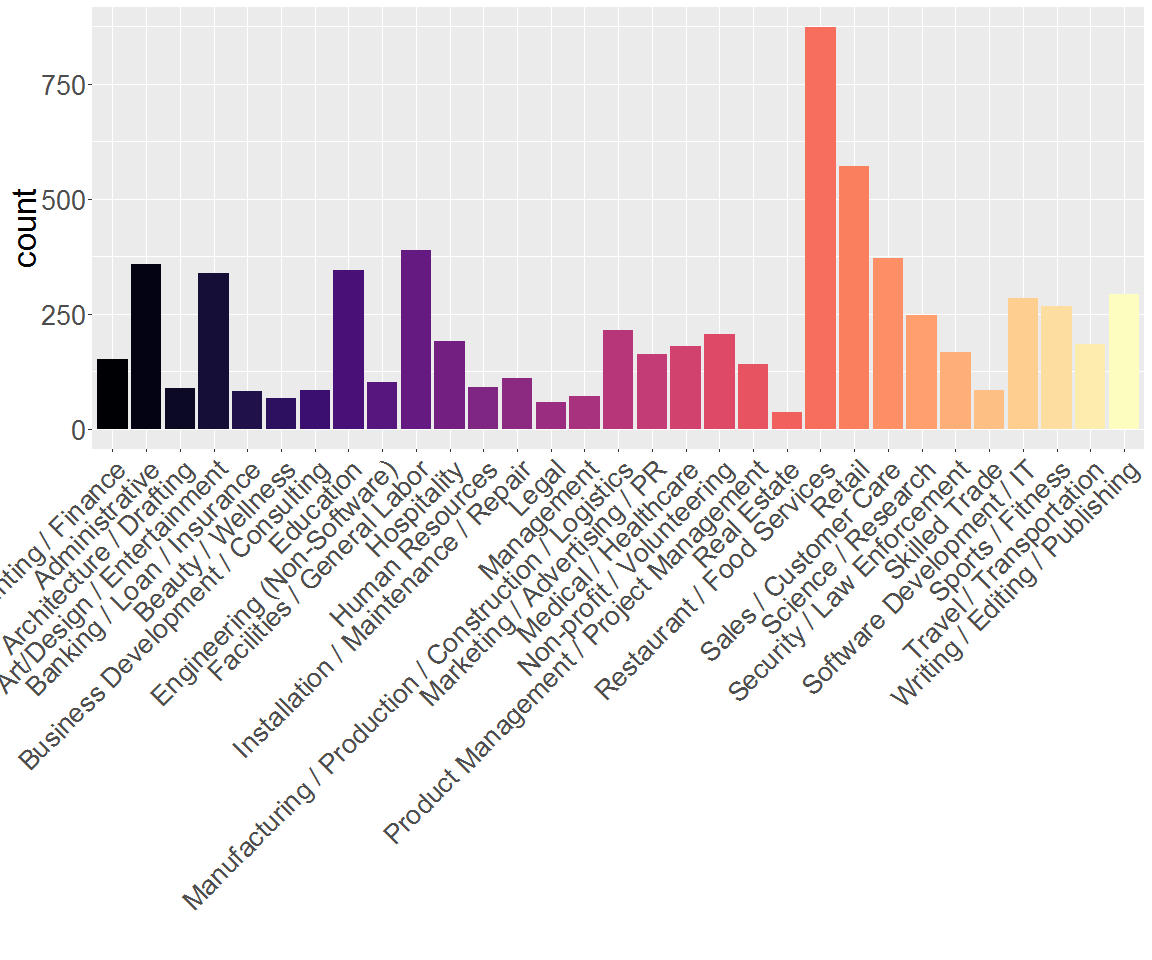


No comment, say hi to my sibling Mowgli. He's quite a beauty.

So, well, as any statistical or machine learning prediction... the data you use for training your model is quite crucial. The jobs classifier could probably use even more data for improving classification. As any Monkeylearn public module, it can be built upon and improved, so who's in for forking it? In the meanwhile, it still offers an interesting output to play with. I nearly want to add "Monkeylearn user" as an "Entertainment" job because our sample of classified job descriptions is a nice playground for looking at life trajectories.

# What sorts of jobs did people describe in their tweets?

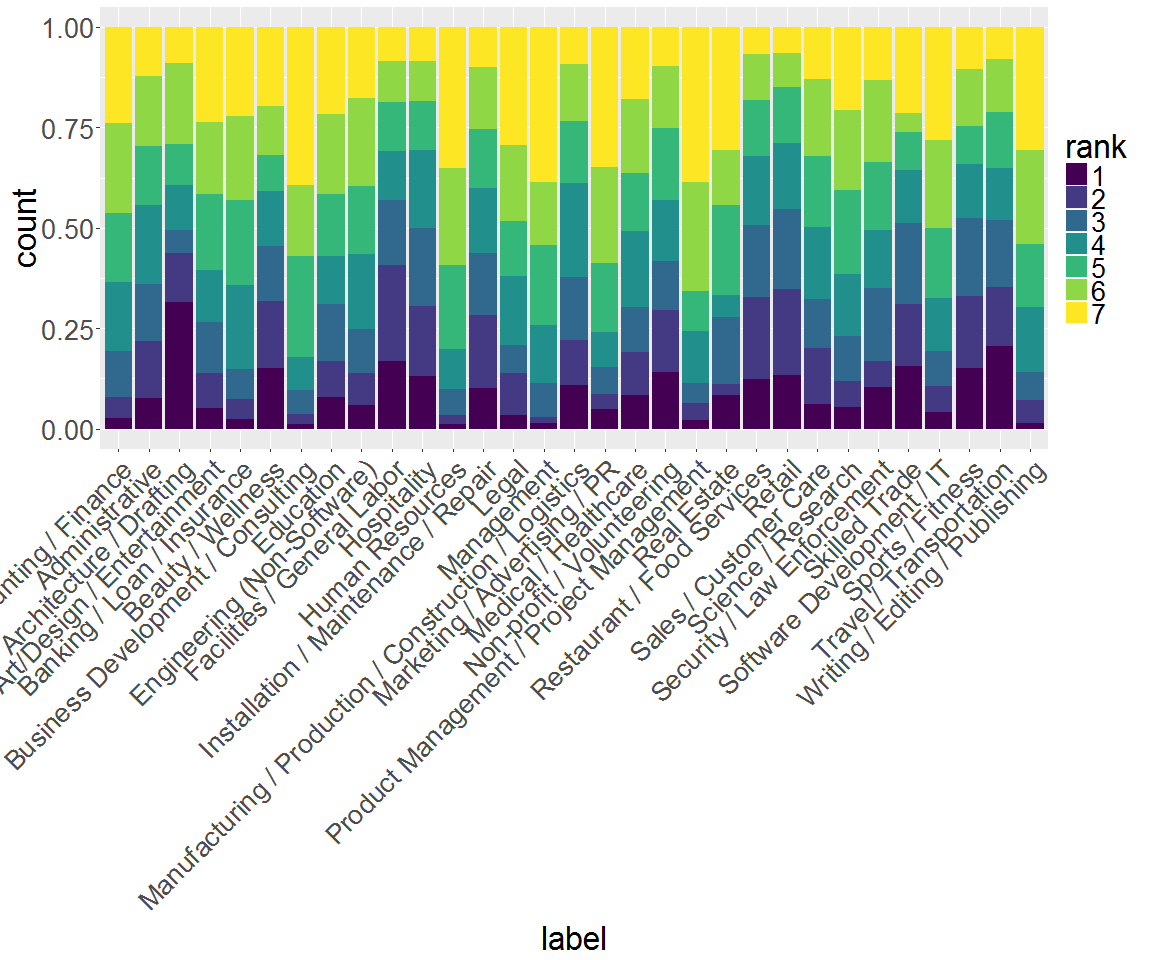
The 6801 jobs for which we predicted a category with a probability higher than 0.5 are divided as follows among industries:



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

# Juniorness of the jobs in each field

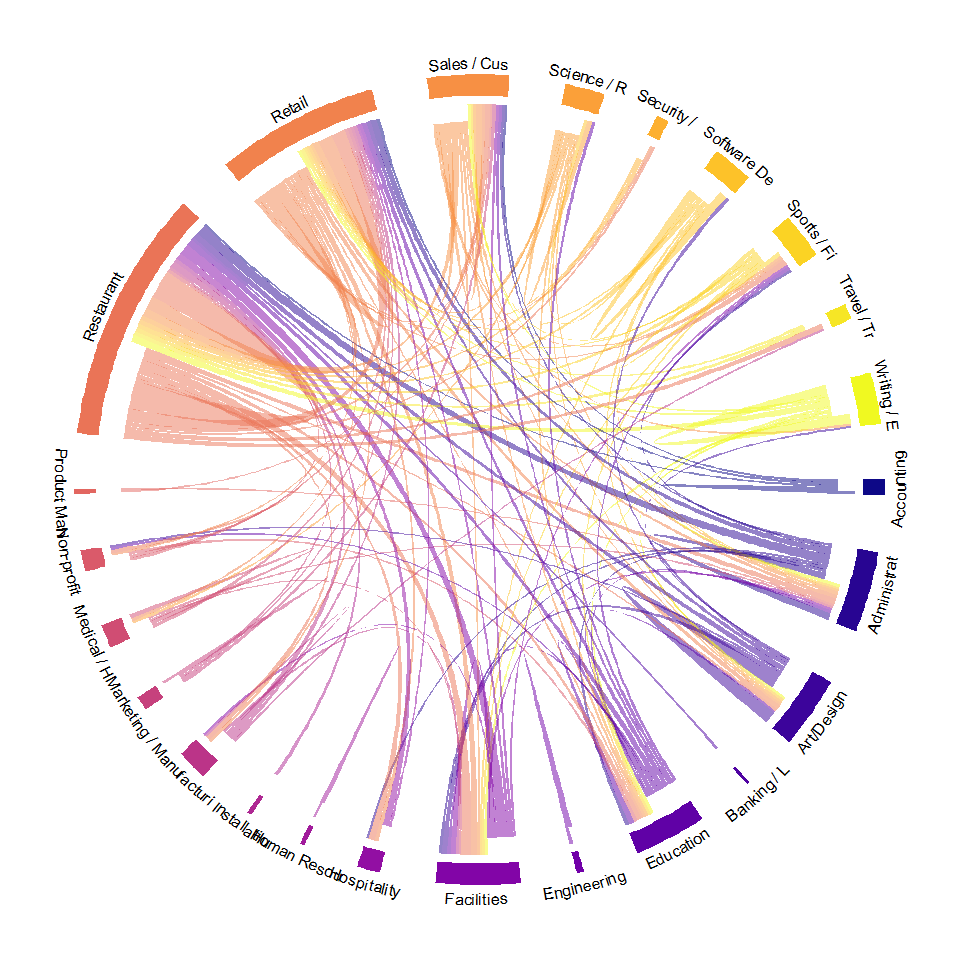
Since we know for each job whether it was the first, third or seventh job of the tweeter, we can explore whether some categories are rather first first jobs than late first jobs. For this, inside each field we can look if the field was mostly a label for first first jobs or for seventh first jobs. See it for yourself:



I'd tend to say that some industries such as Business Development / Consulting are not first-entry jobs (more yellow/green i.e. later jobs), while Non-Profit / Volunteering have a higher proportion of brand-new workers (more blue). 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 a field was predicted with a probability higher than 0.5. I kept only possible transitions where there were present more than 10 times in the data, otherwise we'll end up looking at a [hairball](https://twitter.com/drob/status/768485328244056065).



On this circle you see different industries, and the transition between them. The length of the circle occupied by each field depends on the number of jobs belonging to this category, so again the Food and Restauration category is the biggest one. One can see that people taking a position in the Hospitality field, below the circle, often come from the Restauration or the Retail field When they leave this field, they'll often go work in the Restauration field [David Robinson](http://varianceexplained.org/) suggested I find 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!

As a conclusion, I'm quite excited by the possibilities offered by Monkeylearn for text mining. I might be a grumpy and skeptical statistician so I'll tend to look at all the shortcomings of predictions, but really I think that if ones takes the time to train a module well, they can then get pretty cool information from text written by humans. Now if you tweet about this article, I might go and look at [Monkeylearn's sentiment analysis for tweets module](https://app.monkeylearn.com/main/classifiers/cl_qkjxv9Ly/) instead of reading them.

# 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.