Web Scraping and Computational Methods in Correspondence Experiments

Jake Anderson

UC Love Data Week 2026 - UCLA

Are Emily and Greg More Employable Than Lakisha and Jamal?

Marianne Bertrand & Sendhil Mullainathan 2004 *AER*

- Resumes with White-sounding names received 50% more callbacks than identical resumes with African-American names.
- Racial discrimination persisted across industries, occupations, and employer types, including "Equal Opportunity Employers."
- When resume quality was improved, even greater gaps emerged.

Systemic Discrimination Among Large U.S. Employers

Kline, Rose, Walters 2022 *QJE*

Methodology

Correspondence experiment: 100,000 applications to 108 large employers, randomized across race (Black vs. White names), gender, age, etc.

Racial Gap:

Black-named applicants received 2.1 pp fewer callbacks ($\approx$ 9% drop). Some firms favor men, some women. Top 20% of firms account for nearly 50% of racial callback gap; 23 firms identified as discriminating.

Implications

Evidence supports targeted enforcement by regulators could address systemic bias.

How to Apply to Jobs?

Big Picture

1. Find eligible jobs to apply for
2. Apply for the jobs manually to understand what is necessary for the application
3. Automate the steps

Infrastructure

1. Fictitious profiles: names, addresses, resumes, phone numbers
2. Fictitious emails, phone numbers, hosted through google
3. Flask website for organizing all of the applications

Output

1. Bots applying to jobs every day, approximately 1000/week
2. (After applying), callbacks and voicemails

Browser Automation: What could go wrong?

Potential Issues

- What if the website changes?
- What if the page takes too long to load?
- How can you keep track?
- What about security?

Browser Automation

What could go wrong?

What if the website changes?

- Many times we saw small or large changes to the website
- Small changes are quick fixes, and we're back up and running
- Large changes mean an entirely new script had to be written
- Sometimes the changes are so large that automation is too hard

Browser Automation: What could go wrong?

What if the page takes too long to load?

- Something called a "decorator", basically a custom function you can apply to your code to carry out a specific task
- in this case, I wrote one to retry or reattempt if the application was not successfully submitted
- You can make your functions wait, retry one or multiple steps
- You can have a "timeout" feature to catch browser issues, or tell it to retry a maximum number of times

Browser Automation: What could go wrong?

How can you keep track?

- When tasks are prone to idiosyncratic errors, instead of running a "for loop" to go through the elements, I recommend having a separate tracker file that is updated by reading the log file you create.
  - Example: instead of "for job in job_list: apply_to_job(job_list)", apply for the job, then log a successful job with the application id or other output, and then update the tracker data.
  - This way, you can start where you left off automatically, or skip errors on the first pass and return to them later.
- Logging is so important! Log errors, status, steps, and the time each job required. These are great indicators for when something is going wrong.

Browser Automation: What could go wrong?

What about security?

- Many website owners/companies do not want bots to interact with their website.
- Websites have a dedicated robots.txt and there are conventions and etiquette for scraping and interacting with websites generally. It is up to you how to apply those principles.
- Sometimes they will encrypt things, require log in, throttle you / limit your utilization, implement CAPTCHAs, or utilize third party security software to ban your IP address.

Browser Automation: What could go wrong?

How to circumvent?

- You can plug in whatever you want into the options that the chromedriver offers: randomize screen size, device, etc. Spoof IP addresses.
- Code human-like behavior (most humans don't copy paste their name into the "name" field. They hover, click, pause, type.
- Code a cyborg–a bot that will scrape until a CAPTCHA appears, then you manually solve it, and let the bot keep working.
- Does a site require captcha / security every time you navigate to a page? Or just on the first pass? Try a lot of things.
  $\Longrightarrow$ Example: pages on USAJOBS.com

Potential Audit of Government Jobs

- What does this look like in practice?

DOL Discrimination Audit via USAJOBS.gov

Outline

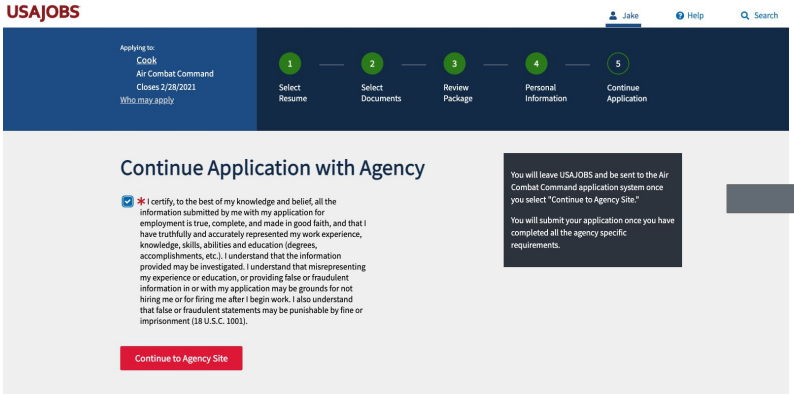
Application Process

What jobs can we audit?

How long would it take?

Application Process For USA Jobs

1. **General Application Hosted at USA Jobs**
2. **Agency Specific Application**



"USAStaffing.com"

Agency Specific Site

Example: Capitol Police



U.S. CAPITOL POLICE

Application Progress

Position Title Police Officer
Agency U.S. Capitol Police
Announcement Number USCP-21-11033194-LS
Open Period Monday, March 1, 2021 to Wednesday, March 31, 2021

Application Package Status: Application Ready to Submit

Review & Submit

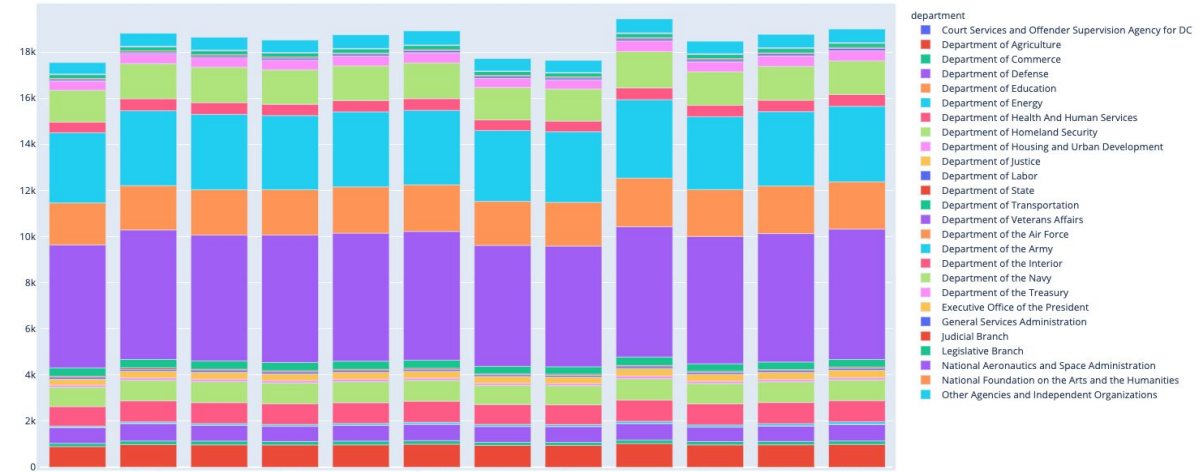
Please verify each section of your application is complete and correct. You may review and change your information before submitting.

Status	Section	
☑	Application ▾	
	☑ Biographic Information	
	☑ Eligibilities	
☑	Assessment ▾	
	☑ Assessment 1	
☑	Documents ▾	
	Accepted Documents	**Submitted Documents**
	☑ Resume **(required)**	Resume View

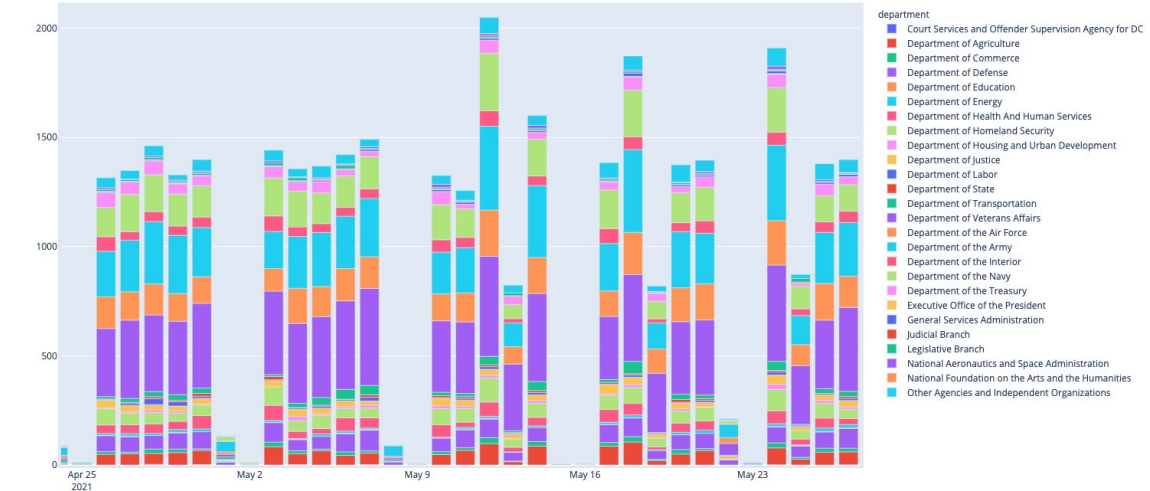
* ☐ **I acknowledge that all of the supporting documents I want to submit with this application have been assigned to an Accepted Document type on the Documents page.**

Only Resume Required ✅

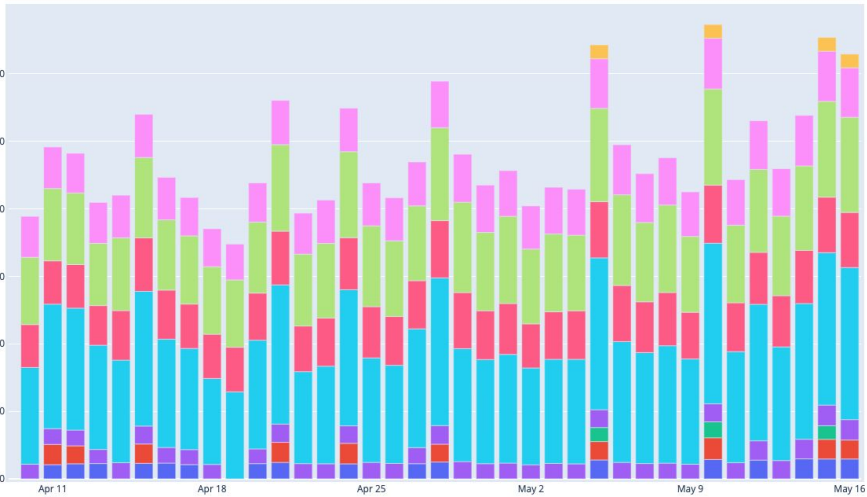
Total Jobs on USAJOBS.gov ~18,000 /day



Total Flow on USAJOBS.gov ~1500 /day



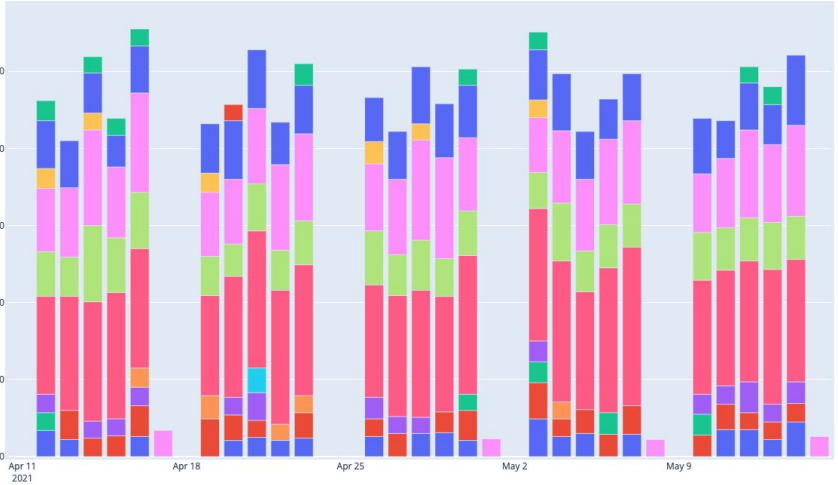
USA Jobs: Stock - Total Public Jobs Per Day



~5000 Jobs / Day

department
- Department of Agriculture
- Department of Defense
- Department of Health And Human Services
- Department of Homeland Security
- Department of Justice
- Department of Veterans Affairs
- Department of the Air Force
- Department of the Army
- Department of the Navy
- Other Agencies and Independent Organizations

USA Jobs: Flow - New Jobs / Day



~500 New Jobs / Day

department

- Department of Agriculture
- Department of Defense
- Department of Health And Human Services
- Department of Homeland Security
- Department of Justice
- Department of Transportation
- Department of Veterans Affairs
- Department of the Air Force
- Department of the Army
- Department of the Interior
- Department of the Navy
- Department of the Treasury
- Other Agencies and Independent Organizations

Departments & Agencies

Department of Homeland Security
- Transportation Security Administration

Department of the Army
- Army Installation Management Command
- U.S. Army Reserve Command
- U.S. Army Medical Command
- U.S. Military Entrance Processing Command

Department of Agriculture
- Farm Service Agency
- Food Safety and Inspection Service
- Natural Resources Conservation Service

Department of Veterans Affairs
- Veterans Affairs, Veterans Health Administration
- Veterans Health Administration

Legislative Branch
- House of Representatives

Department of the Air Force
- Air Combat Command
- U.S. Air Forces, Europe
- Air Mobility Command
- Air Education and Training Command
- Air Force Global Strike Command

Department of Justice
- Justice, Bureau of Prisons/Federal Prison System

Most frequent jobs by department

Transportation Security Administration - Transportation Security Officer (TSO)
Army Installation Management Command - Child and Youth Program Assistant, Recreation Assistant, Custodial Worker, Cook, Food and Beverage Attendant NA-03, Lead Child and Youth Program Assistant , Food Service Worker NA-02, Recreation Aid, Laborer
U.S. Army Reserve Command - Heavy Mobile Equipment Repairer
Farm Service Agency - County Program Technician
Veterans Affairs, Veterans Health Administration - Medical Technologist, Nursing Assistant, Medical Support Assistant, Licensed Practical Nurse, Social Worker, Psychologist, Physician, Registered Nurse
Veterans Health Administration - Advanced Medical Support Assistant, Nursing Assistant, Medical Technologist, Medical Support Assistant, Social Worker, Physician (Primary Care), Food Service Worker, Housekeeping Aid, Licensed Practical Nurse, Practical Nurse, Psychologist
Food Safety and Inspection Service - Consumer Safety Inspector, Food Inspector
Natural Resources Conservation Service - Soil Conservationist, Soil Conservation Technician
U.S. Army Medical Command - Clinical Psychologist, Practical Nurse/Vocational Nurse, Practical Nurse
House of Representatives - Gold Star Family Fellow
Air Combat Command - Cook, Child and Youth Program Assistant, Laborer
U.S. Military Entrance Processing Command - Test Administrator
U.S. Air Forces, Europe - Child and Youth Program Assistant, Food Service Worker
Air Mobility Command - Child and Youth Program Assistant, Laborer, Recreation Aid
Air Education and Training Command - Child and Youth Program Assistant, Recreation Aid
Air Force Global Strike Command - Cook
Justice, Bureau of Prisons/Federal Prison System - Cook Supervisor (Cook Foreman)

Job Title Analysis

- Okay, you apply to all the eligible jobs, now what?
- Need to clean up the job titles, and match to job characteristics

Job Titles Cleaning

Problems and solutions

Items to remove	Job Title
FT, Hours	"Community Branch Banker I (FT 40 hrs) Clanton"
City	"CDL - A Truck Driver - Rochester"
City, PT	"LEAD SALES ASSOCIATE-PT in TREVOR, WI"
Institution	"Sous Chef - Leisure - Manhattan College"
Full-Time	"Full-Time Retail Sales Associate"
not benefit eligible	"Part Time Teller 18 hours Opitz Crossing 20% Premium Pay (*not benefit eligible)"
Store number	"Retail Staff (Store 6022)"
Full time	"Beauty Brand Expert, Full Time: First Colony"
Empire Outlet	"Brand Associate - Empire Outlet" (Empire Outlet is the mall name)
Street	“Pick 'n Save Retail Clerk (Caledonia/Douglas Ave.)"
number/job code	"Sales Operations Analyst#R16904"
Location	"SALES ASSOCIATE in ORANGE BEACH, AL"
Nights, pay	"Driver - Tractor Trailer Driver RTD DOT - Part Time - NIGHTS - $19.80"
Days, pay	"Warehouse Package Handler Non Dot Part Time DAYS $13.53"
Weird symbol	'Part-Time (25 hours) Associate Banker √¢¬Ä¬ì Willits Branch √¢¬Ä¬ì Willits, CA'
Symbol	"Customer Service Associate Representative ‚Äì Indiana ‚Äì Express Scripts"

Job Titles Cleaning

Most of the work is done through regex (regular expressions)

Job Titles Cleaning

Useful functions!

```
In [718]: # Catches "Part Time" and "P
          df = remove_and_flag(r"(\(?p
          # catches 3 cases of 'pt'
          df = remove_and_flag(r"( +pt
          # catches PT
          df = remove_and_flag(r"(\(?P
          # catches "full Time"
          df = remove_and_flag(r"(\(?F
          # catches FT
          df = remove_and_flag(r"(\(?P
          # catches "xx hours" or "xxh
          hours_re = r"(\(?[0-9]+ ?[Hh
          df = remove_and_save(hours_
          # removes city and state fro
          city_state_re = r"([A-Z][a-z
          df = remove_and_save(city_st

          Entries corrected regarding
          Entries corrected regarding
          Entries corrected regarding
          Entries corrected regarding
          Entries corrected regarding
          Entries corrected regarding hours:
          Entries corrected regarding locati
```

jobtitle	link	firm	part_time	full_time	hours	location
Delivery Driver - Pharmacy Services (full-time	https://jobs.cardinalhealth.com/search/jobdeta...	Cardinal Health	False	False	40 hours)	NaN
Delivery Driver - Pharmacy Services -	https://jobs.cardinalhealth.com/search/jobdeta...	Cardinal Health	True	False	20 hrs)	NaN
Teller -	https://usbank.taleo.net/careersection/10000/j...	U.S. Bancorp	False	False	(25hrs)	Monticello, MN
Client Relationship Consultant 1 NMLS Overlan...	https://usbank.taleo.net/careersection/10000/j...	U.S. Bancorp	False	False	(40hrs)	Park, KS

This allows for easy regressing / filtering later. Suggestions / Improvements?

Job Titles Cleaning

```
Entries corrected regarding part_time          :        980
Entries corrected regarding part_time          :          3
Entries corrected regarding part_time          :        237
Entries corrected regarding full_time          :        320
Entries corrected regarding full_time          :        265
Entries corrected regarding hours              :        378
Entries corrected regarding title_city_state   :       1524
Entries corrected regarding seasonal           :        443
Entries corrected regarding city               :        823
Entries corrected regarding zipcode            :          7
Entries corrected regarding state              :        153
Entries corrected regarding title_wage         :        207
Entries corrected regarding bonus              :         52
Entries corrected regarding benefits           :         35
Entries corrected regarding 'Äì'               :         42
Entries corrected regarding '&amp'             :         39
Entries corrected regarding ':'                :        282
Entries corrected regarding '-'                :       4086
Entries corrected regarding '\.'               :        117
Entries corrected regarding flexible           :         19
Entries corrected regarding non_dot            :         18
Entries corrected regarding dot                :         44
Entries corrected regarding requisition        :        112
Entries corrected regarding # numbers          :        183
Entries corrected regarding times              :         61
Entries corrected regarding sign-on            :         24
Entries corrected regarding store numbers      :        121
```

~60% of the 11,506 job titles are cleaned

Job Titles Cleaning

1. Cleaned specific Job titles or locations that caused downstream issues
2. Cleaned up ordinal and other numbers
3. Days and shift times
4. Replaced certain jobs with titles that reflect the duties more accurately

```
# Very Specific Changes
df.loc[df.job_id == 10689,'city'] = "St. George"
df.loc[10755,"jobtitle"] = "General Warehouse Associate"
df.loc[df.firm == "Starbucks",'jobtitle'] = "Barista"
# Should be 125 entries
print(f"Entries corrected regarding {'Baristas':<20}: {df.loc[df.firm == 'Starbucks'].shape[0] : >10}")
# Removing # and everything after
pound_re = r"(#.+)"
df = remove_from_title(df, pound_re, "# sign")
num_re1 = r"([0-9]nd)"
df = remove_from_title(df, num_re1, "Ordinal Numbers -nd")
num_re2 = r"([0-9]th)"
df = remove_from_title(df, num_re2, "Ordinal Numbers -th")
num_re3 = r"([0-9]st)"
df = remove_from_title(df, num_re3, "Ordinal Numbers -st")
# getting rid of all numbers
num_re = r"([0-9]+)"
df = remove_from_title(df, num_re, "[0-9]")
# Getting rid of remaining "hours"
hour_re2 = r"((an)?(per)? ?/?[Hh]our(ly)?)"
df = remove_from_title(df, hour_re2, "an hour/per hour")
# Other words
other_re = r"([Nn]ew)"
df = remove_from_title(df, other_re, "new")
# Too many spaces
df = replace_in_title(df, r"(  +)", " ", "spaces")
df.loc[:,'jobtitle'] = df.jobtitle.str.strip()
# nights
nights_re = r"(Nights?)"
df = remove_from_title(df, nights_re, "nights")
# shift
df = remove_from_title(df, r"([Ss]hift)", "shift")
days = ['Mon','Tues','Wed','Thurs','Fri','Sat','Sun','[Ss]tart','Time']
for day in days:
    df = remove_from_title(df, day, f"'{day}'")
```

```
title_replacements = {
"guestcoordinator": "Guest Coordinator",
"TSS Stores": "Security Guard",
"Target Security Specialist":"Security Guard",
"HHA/CNA":"Home Health Aide / Certified Nursing Aide",
}
```

Job Titles Cleaning - 90% complete

```
jobtitles_main = ['Sales Associate','Teller','Bartender',
                  'Laborer','Customer Service Representative',
                  'Medical Assistant','Driver','Dishwasher',
                  'Barista','Merchandiser','Dispatcher','store job',
                  'Specimen Processing Specialist','Sales & Support Generalist',
                  'Outside Sales Representative','Server','Commission Sales Associate',
                  'Breakfast Attendant','Merchandiser','Retail Merchandiser', 'Sous Chef',
                 'Retail Sales Specialist','Installation Technician','Sales Consultant',
                 'Prep Cook','Cashier','Certified Nursing Assistant','Housekeeper','Beauty Advisor',
                 'Associate','Cook','Maintenance','Housekeeping','Sales Representative','Claims Adjuster',
                 'cashier','CASHIER','Mortgage Loan Consultant','Security Specialist','Shift Lead',
                 'Service Advisor','General Labor','Team Leader','Branch Administrator',
                 'Material Handler','Forklift Operator','Loss Prevention Officer','Sanitation Operator'
                 'Service Coordinator','Service Coordinator','Warehouse Selector','SALES FLOOR ASSOCIATE',
                 'Assembler','Automotive Technician','Cargo Handler','Office Administrator','Outside Sales Rep',
                 'SALES ASSOCIATE','Banker','Machine Operator','store driver','Medical Courier','Retail Staff',
                 'Sorter','Property Adjuster','Service Technician','HVAC Pipefitter','Administrative Assistant',
                 'Information Security Compliance Analyst','Custodial Services','Customer Service Specialist',
                 'On Call Shuttler','Any position','KFC Team Member','Patient Care Technician','Guest Coordinator',
                 'Porter Service','Busser','TM Receiver','Team Member','Field InHome Solutions Technician',
                 'Inside Sales Rep','COOK','Sales Advisor','Transporter','Luxury Retail Beauty Sales',
                 'Oil Tire Changer','Utility Worker','Claims Specialist','Receiver','Claim Specialist',
                 'Reliability Technician','TSS Stores','driver','custodian','Technician','Clerk','WAREHOUSE SPECIALIST
                 'In Home Sales Expert','Design Consultant','Disney Store Key Holder Cast Member',
                 'Client Relationship Consultant','Helper','Guest Coordinator','Security Guard','Host','Receiving',
                 'Loader','Order Selector','Sales Agent','Certified Nursing Aide','Asset Protection Agent',
                 'Oil & Tire Changer','Retail Sales Professional','HOUSEKEEPER','Service AC Repair','Customer Service A
                 'Customer Experience Representative','Yard Worker','Car Cleaner','Scalehouse Attendant',
                 'Administrative Support','ASSET PROTECTION SPECIALIST','Appliance Sales Specialist',
                 'Pizza Artist','sales associate', 'Guest Service Aide','sales representative','Auto Damage Adjuster',
                 'Room Attendant','Production Worker', 'Oil and Tire Changer','Dockworker','Liability Adjuster',
                 'Selling and Support Captain','Outside Sales']

jobtitles_secondary = ['Stocker','Warehouse','eCommerce','InHome Sales','Merchandise','Processor','Freight',
                       'Unloader','Home Health Aide']
```

~150 Job Titles cover
~10,000 / 11,506 Job Titles

Job Titles Cleaning

Examples	
Selling and Support Captain	Not a captain
Customer Experience Representative	Retail worker
Target Security Specialist	Greeter
Subway Sandwich Artist	Not an Artist

O*Net

"The O*NET survey began in 1998 and is updated periodically."

Background
15 questions about a worker's amount of experience and background.

Education and Training
6 questions about the education and experience requirements for a job.

Knowledge
66 questions about work-related areas of knowledge. Knowledge areas are sets of facts and principles needed to address problems and issues that are part of a job.

Work Activities
84 questions about work activities. A work activity is a set of similar actions that are performed together in many different jobs.

Work Context
57 questions about working conditions. These questions are about a work setting and its possible hazards, the pace of work, and dealings with other people.

Work Styles
16 questions about work styles. A work style is a personal characteristic that can affect how well someone does a job.

O*Net - How Does Dave Deming Do It?

1. **Convert average scores by occupation to a 0-10 scale.**
    a. *"I follow Autor, Levy, and Murnane (2003) and convert average scores by occupation on O*NET questions to a 0–10 scale that reflects their weighted percentile rank in the 1980 distribution of task inputs."*
2. **Define "math" task intensity by subset of questions.**
    a. *"Second, I closely follow Autor, Levy, and Murnane (2003) and define nonroutine analytical (math) task intensity as the average of three O*NET variables that capture an occupation's mathematical reasoning requirements."*
3. **Define "social skills" task intensity by subset of questions.**
    a. *'Third, I define an occupation's social skill intensity as the average of the four items in the O*NET module on "social skills": (i) coordination, (ii) negotiation, (iii) persuasion, and (iv) social perceptiveness'*

Onet Job Matching - Top 20 Jobs

Deck for DOL

→ **Onet job Matching**

rr_job	onet_title	onet_soc_code	method	match_score
Barista	Baristas	35-3023.01	ratio	93.0
Cashier	Cashiers	41-2011.00	ratio	93.0
Customer Service Associate	Customer Service Representatives	43-4051.00	ratio	72.0
Retail Sales Consultant	Retail Salespersons	41-2031.00	ratio	71.0
Outside Sales Representative	Patient Representatives	29-2099.08	ratio	71.0
Cashier	Gambling Change Persons and Booth Cashiers	41-2012.00	partial_ratio	100.0
Barista	Baristas	35-3023.01	partial_ratio	100.0
Cashier	Cashiers	41-2011.00	partial_ratio	100.0
Branch Office Administrator	Web Administrators	15-1299.01	partial_ratio	88.0
Field InHome Solutions Technician	Avionics Technicians	49-2091.00	partial_ratio	81.0
Barista	Baristas	35-3023.01	token_sort_ratio	93.0
Cashier	Cashiers	41-2011.00	token_sort_ratio	93.0
Outside Sales Representative	Patient Representatives	29-2099.08	token_sort_ratio	75.0
Retail Sales Associate	Real Estate Sales Agents	41-9022.00	token_sort_ratio	74.0
Retail Sales Specialist	Retail Salespersons	41-2031.00	token_sort_ratio	71.0

Updates

Deck for DOL

→ **Onet job Matching**

It worked! ✅
Matched the following ratings:

- 'Achievement/Effort'
- 'Persistence'
- 'Initiative'
- 'Leadership'
- 'Cooperation'
- 'Concern for Others'
- 'Social Orientation'
- 'Self Control'
- 'Stress Tolerance'
- 'Adaptability/Flexibility'
- Dependability'
- 'Attention to Detail'
- 'Integrity'
- 'Independence'
- 'Innovation'
- 'Analytical Thinking'

randres_job_title	onet_job_title	onet_soc_code	match_score	Achievement/Effort	Persistence	Initiative	Leadership	Cooperation	Concern for Others
Customer Service Representative	Customer Service Representatives	43-4051.00	98.0	3.83	3.58	3.45	3.10	4.51	4.12
Medical Assistant	Medical Assistants	31-9092.00	97.0	4.08	3.97	4.12	3.88	4.54	4.42
Security Guard	Security Guards	33-9032.00	97.0	2.91	3.17	3.19	3.43	3.64	3.38
Dishwasher	Dishwashers	35-9021.00	95.0	2.87	2.85	3.11	2.31	3.39	2.94
Bartender	Bartenders	35-3011.00	95.0	3.70	3.63	3.68	3.31	4.13	3.95
Barista	Baristas	35-3023.01	93.0	3.06	2.71	3.29	3.10	4.32	3.79

Updates

Deck for DOL

→ **Onet job Matching:**

There are still some corrections to be made to the worst matches

randres_job_title	onet_job_title	
SALES FLOOR ASSOCIATE	Real Estate Sales Agents	👎
Host	Hosts and Hostesses, Restaurant, Lounge, and C...	👍
WAREHOUSE SPECIALIST	Data Warehousing Specialists	👎
ASSET PROTECTION SPECIALIST	Project Management Specialists	👎
HOUSEKEEPER	Choreographers	👎

Voicemail Data

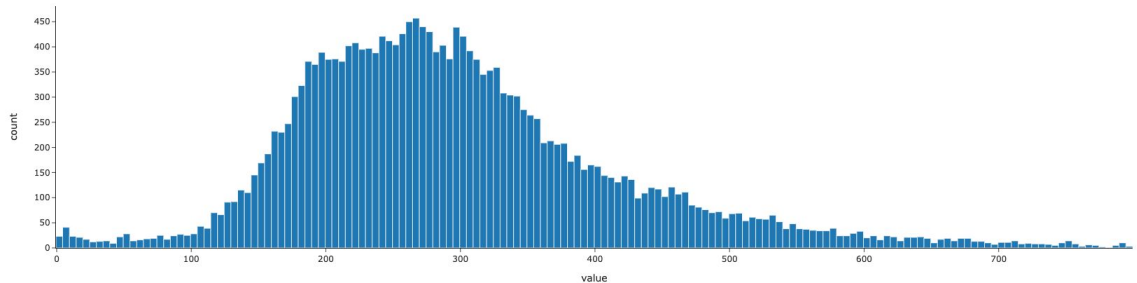
- The employers called you back, but what did they say?

00 Overview

~31k voicemails

- ~1/3rd have no text, leaving ~21k

Length of Voicemails (Word Count)



01 Example Voicemails

- Action [redacted]his My name is John and Michael on the deli manager of the public heavens I saw you put a hapless cation in with this I was curious if you're available for an interview on Wednesday afternoon about 3 pm if you can give me a call at the store 7 of [redacted] thank you.
- Hey Misty this is cow to harvest and store Publix I was just calling to see if you'd be interested in coming and setting up an interview with my store manager if you are get a call back to the store our phone number is [redacted]Have a great want.
- Hi Jason this is your gauge from various management human resources and I'm calling regarding a job application with us for the position of assistant diesel mechanic on the job location. Is sent to Paul Minnesota I would request you to please call me back at [redacted] between 9 am to 5 30 pm c.s.t. so that we can complete your interview process you can also leave me a waste mail stating that what would be a good time to call you back and I will give you a call back [redacted]hank you and have a great day.

03 “This is...”

~9,000 matches

Captured Name

Could be a name in original recording

False Positive

So Laurie the suspect are calling from Waste Management and this is regarding your application for the health purpose Isha I wanted to check with you if you were sitting Kristin that we can schedule a phone interview with you simply to precede our G ive us a call back [redacted] thank you.

Well overshoot this is Phil are calling from with management decision regarding your application for the whole purpose Isha and I wanted to check with you if you were sitting interested in that position so that we can schedule a phone interview wit h you sublist if you still are please give us a call back at [redacted] thank you hope you have a great day.

Hi This message is for Jeremy [redacted] this is calling from Cardinal Health if he could to which I might call and I will by 6 to [redacted] thank you.

Good afternoon this is Levi from Dollar General I'm looking for him Meredith [redacted]. We have your application here and I was going to call the see if you're interested in coming in for interview if so give me a call back and the number is [redacted] hearings do you have a wonderful day for by.

Just and this is a dangerous game stop in Hot Springs I'm just calling to set up a time for an interview on May 27th. 215 in the afternoon [redacted] or 545 if you can just give me a call back with your response Thank you my.

03 “My name is...”

~5,400 matches

Captured Name

Could be a name in original recording

False Positive

Hi This call was from Meredith My name is to Angela [redacted] and I was giving you a call from Pink into a box I was actually called a c.v. we're still interested in a position in peek. So we can set up our interview so in when you g et its message please give us a call back at [redacted] again my name is the Angela [redacted] talk to you so my bos s.

Hi Emily this serene from Victoria Secret We received replication I'm just calling to follow up with a couple questi ons give a 2nd to call me back at the store today the number is [redacted] extension 0 again my name is serene Tha nk you.

I carry my name is Cheryl [redacted] I'm a recruiter with spectrum I am calling regarding your application to our fiel d technician jobs in Marysville Can you give me a call when you get a chance at [redacted]. I'm also going to send you a text 2 if it's easier to call that number thank you.

Hi I'm calling from Kenya My name is Kristen from Victoria's Secret We were calling about the application you submit ted if you could give us a call back at [redacted] that would be great thank you.

My name is Tony I'm calling from the Rite Aid in Bern Mar looking for Kerry to put an application in Please Give me a call by numbers [redacted] thank you by.

03 Proper Noun Tag

Via Natural Language Tool Kit (NLTK)

Example

```
In [88]: sentence = "Michael Jackson likes to eat at McDonalds"
         tagged_sent = pos_tag(sentence.split())
         tagged_sent

Out[88]: [('Michael', 'NNP'),
          ('Jackson', 'NNP'),
          ('likes', 'VBZ'),
          ('to', 'TO'),
          ('eat', 'VB'),
          ('at', 'IN'),
          ('McDonalds', 'NNP')]
```

“NNP” indicates proper nouns

```
In [96]: sentence.split()

Out[96]: ['Michael', 'Jackson', 'likes', 'to', 'eat', 'at', 'McDonalds']

In [89]: propernouns = [word for word,pos in tagged_sent if pos == 'NNP']

In [90]: propernouns

Out[90]: ['Michael', 'Jackson', 'McDonalds']
```

Final Result

03 Proper Noun Tag

~20,000 voicemails had some kind of proper noun

Captured Name

Could be a name in original recording

False Positive

```
['Hi', 'Jacob', 'David', [redacted], "I'm"]
['Laurie', 'Waste', 'Management', 'Isha', 'Kristin', 'Give']
['Phil', 'Isha']
['Hi', 'Lori', 'Patty', 'Victoria', 'Secret', 'Parkway', 'Plaza', 'Mall']
['Freeport']
['Hello', 'John', 'Freeport', 'Emery's"]
["I'll", 'H', "I'm", "I'm", 'Congress', 'Box', 'Denver', 'Colorado', 'Thank']
["I'm", "Marquis's", "I'll", "I'll", 'Friday']
['Well']
["Let's", 'Kerry', [redacted], 'Cardinal']
['Hi', 'Jeremy', [redacted], 'Cardinal', 'Health']
['Never', 'Carton', 'House']
['Levi', 'Dollar', 'General', "I'm", 'Meredith', [redacted]']
['Hot', 'Springs', "I'm", 'May', 'Thank']
['Lutetia', 'Lateesha', 'My', 'Sarah', [redacted], "I'm", "Dillard's", 'Department', 'Eastwood', 'Mall', "I'm", 'Sarah', [redacted]]
[[redacted], 'My', 'Sara', [redacted], [redacted], "Dillard's", 'Department', 'Eastwood', 'Mall', "I'm", "I'd", 'Sarah', [redacted], 'Mall', 'Treo']
['Brett', 'Brett', 'My', [redacted], "Dillard's", 'Department', 'Eastwood', 'Mall', "Dillard's", 'Department', 'Store', 'Sarah', [redacted], 'Treo']
['Brad', 'Brad', 'My', 'Sara', [redacted], "I'm", "Dillard's", 'Department', "I'm", 'Please', 'My', 'Sarah', [redacted], "Dillard's", 'Eastwood', 'Mall', 'Thank']
[[redacted], 'Athens', "I'm"]
['Hi', 'My', 'Chris', [redacted].', 'St', 'Paul', 'Minnesota', 'Open', 'Mobile', 'Thank']
['Hi', 'Dollar', 'General', "I'm"]
['Hyla', 'Keesha', 'Hi', 'Melrose', 'Park', 'Please']
['Hi', 'Ms', [redacted], 'My', 'Marie', "I'm", 'Avis', 'Budget', 'Marie']
['Anton', 'My', 'Marie', "I'm", 'Avis', 'Budget', "I'd", 'Marie']
['Hi', 'Marie', 'Marie']
['Hi', 'Emily', "I'm", 'Ok', 'Thank']
['Lateesha', 'Jane', 'My', 'Sarah', [redacted], "I'm", "Dillard's", 'Department', 'Sarah', [redacted], 'Treo']
['Don', 'Don', 'My', 'Sarah', [redacted], "Dillard's", 'Department', 'Eastwood', 'Mall', "I'm", 'March', "you're", 'Sarah', [redacted], "Dillard's", 'East']
['Brett', 'Brett', 'My', 'Sarah', [redacted], "I'm", "Dillard's", 'Department', "I'm", "I'm", 'Sarah', [redacted], "I'm", "Dillard's", 'Department', 'East']
['Brad', 'Brad', 'My', 'Sarah', [redacted], "Dillard's", 'Department', 'Store', 'Sarah', [redacted]]
```

High number of false positives

- We can use the phone number and applicant name to remove one of the proper nouns

03 Proper Noun Tag



Different samples of 50 yield wildly different proportions of names vs. nonsense.

Names:

- Lori, Jacob, Laurie, William, Maxx

Nonsense:

- Nickelback, Could, Ok

04 Summary & Next Steps

Summary

1.	"My Name is…"	**~5,400**	~25%	Low number of false positives
	a. Time required to tag:	low, ~8-10 hours		*Based on sample of 50*
2.	"This is…"	~9,000	~40%	Moderate number of false positives
	a. Adjusted Estimate:	**~4,500**	~20%	*Based on sample of 50*
	b. Time required to tag:	low, ~8-10 hours		
3.	Proper noun tagging			Over captures proper nouns
	a. Adjusted Estimate:			
	b. Time required to tag:	high, may require manually listening to the recordings		

Overall Estimate 12/20: Jake can tag roughly **10,000** proper nouns in the next 2-3 weeks.

03 “My name is...”

~5,400 matches

Captured Name

Could be a name in original recording

False Positive

Hi This call was from Meredith My name is to Angela [redacted] and I was giving you a call from Pink into a box I was actually called a c.v. we're still interested in a position in peek. So we can set up our interview so in when you g et its message please give us a call back at [redacted] again my name is the Angela [redacted] talk to you so my bos s.

Hi Emily this serene from Victoria Secret We received replication I'm just calling to follow up with a couple questi ons give a 2nd to call me back at the store today the number is [redacted] extension 0 again my name is serene Tha nk you.

I carry my name is Cheryl [redacted] I'm a recruiter with spectrum I am calling regarding your application to our fiel d technician jobs in Marysville Can you give me a call when you get a chance at [redacted]. I'm also going to send you a text 2 if it's easier to call that number thank you.

Hi I'm calling from Kenya My name is Kristen from Victoria's Secret We were calling about the application you submit ted if you could give us a call back at [redacted] that would be great thank you.

My name is Tony I'm calling from the Rite Aid in Bern Mar looking for Kerry to put an application in Please Give me a call by numbers [redacted] thank you by.

03 “My name is...”

4,455 matches

Captured Name

Could be a name in original recording

False Positive

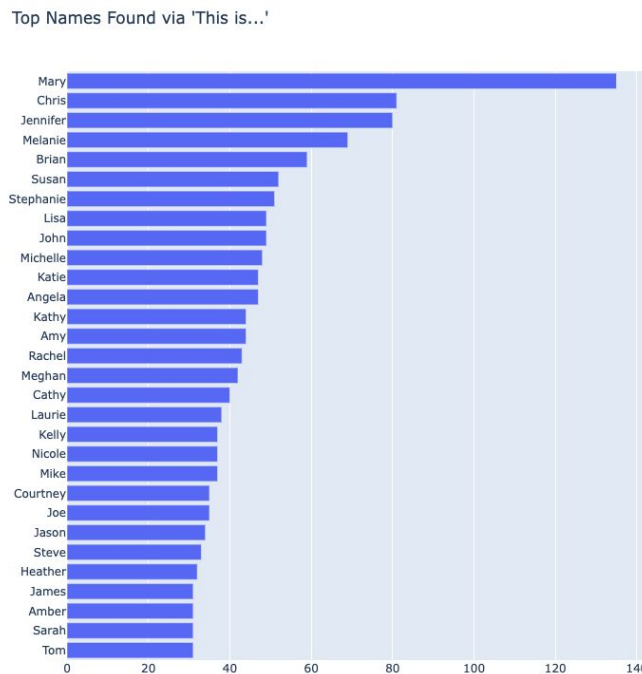
	recording_sid	text	name
31093	REde90defe597095d1169446f4d9fa19f0	Hello my name is Kevin I am calling from the P...	Kevin
31096	REbb314f2a2e94396d752309f09c9f7bb4	Had to though I'm trying to get in contact wit...	Kevin
31098	RE41aef51182710be110964256f33e84b7	Hello my name is Kevin I'm calling from the Pa...	Kevin
31099	RE396b50369cdbb0851129e33aca35269e	Hello this is Kevin calling from the Pacific b...	Kevin
31103	RE3316c60fcdfe77d76d3af77fc70af403	Hello Jermaine This is Kevin calling from the ...	Kevin
31105	REfec4f45f3d5456796210fe3c9f7b4b37	Hi Lori This is Kevin calling from the Pacific...	Kevin
31131	REdec5aa7449fe287465da81a28ea7e619	This message is former Keith[redacted] My name is ...	dusty
31134	REaf44e82bbbb1bdc117c1f8cc644eccbf	Hi this is Mary calling from McLean we were lo...	Mary
31137	RE8f97a1f1edef19402226ef29a16c5c5d	Hi This message is friendly hospital or. My na...	Craig
31151	REd4d942c84f9675cf37a235e4757e8eae	Hi this message I'm sorry this message is for ...	Beverly
31157	REcb13822eb53c13edcf9a9077b49fcf8f	Hi This message is for tomorrow my name is Dan...	Dan
31158	RE71bb13dd3693e2f281b1fff969c1a372	Hi this message just for out of my name is Don...	Don
31169	RE22f54699f35c35441c6f6c1dd577952f	Hi This message is for Kerry My name is Sarah ...	Sarah
31180	RE9e62dbb95f1d5808028c9e1a96ab010e	Hi dispenses for Bradley[redacted] Hi Bradley ...	look
31182	RE338aa3e3e11ae89aa69dcbfdca8e6c6b	Hi This message is for Alison My name is alley...	alley
31183	RE00ade98ee7d28c34d9548f45883b4ce9	Hi This message is for Kareem My name is Allie...	Allie
31184	RE148882d4f51449db3b25b672bafe0adb	I just messages for Justin My name is Alan I'm...	Alan
31185	RE512cbc92c68adbb0debab33e2f2885fd	Hi This message is for Latoya My name is alley...	alley
31186	RE7dc5150d6dc12749a976d9a42a3669eb	Hi This message is for Rashid My name is Allie...	Allie
31198	RE40c3d29c8d05bcc4b471abc878cc6295	Moses 1st Sarah Sarah My name is Robbie snap a...	Robbie



03 “This is...”

5,266 matches





04 Summary & Next Steps

Summary

1.	“My Name is…”	**4500** Captured
2.	“This is…”	**5266** Captured
3.	Proper Nouns	**[In Progress]**

01 Name vs. Establishment

Some Weirdness:

One to many relationship between Phone Number and firm:

- Three phone numbers account for 243 calls from different firms, times, and names
- These are probably “system numbers” or called through the platform

phone_from
+17378742833
+19739374802
+266696687

01 Name vs. Establishment

One to many relationship between phone and name:

One to many relationship between phone and name:

"Unique Names" per phone number	Count
1	2081
2	514
3	144
4	51
5	15
7	7
6	7
8	2
18	1
12	1
10	1

Example: Kroger called 143 times from the same phone number but 18 different names

01 Name vs. Establishment

Are They The Same Name?

Take the Levenshtein distance between names:

L-Distance: 1 2 3 4

name_list	name_list	name_list	name_list
[Trish, Trisha]	[Sam, Stan]	[Jamie, Jeannie]	[Chris, Caleb]
[Holly, Polly]	[Norm, Nora, No]	[Steve, Eve]	[Skype, Scott]
[Corey, Cory]	[Danielle, Daniel]	[Joe, Jill]	[Tina, Amy]
[Jen, Jan]	[Laura, Lauren]	[Jessica, Jesse]	[Mae, Miguel]
[Demie, Demi]	[David, Dave]	[Rick, Ok, Mark]	[Renee, Kevin]

Yellow may be equivalent by “L-vocalization”, Red perhaps not equivalent.

00 Proper Noun Tag

~20,000 voicemails had some kind of proper noun

Captured Name

Could be a name in original recording

False Positive

High number of false positives

- We can use the phone number and applicant name to remove one of the proper nouns
- Avg of ~10 or more “proper nouns”

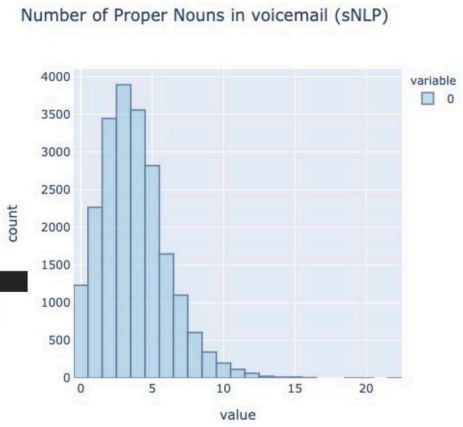
['Hi', 'Jacob', 'David', [redacted], "I'm"]
['Laurie', 'Waste', 'Management', 'Isha', 'Kristin', 'Give']
['Phil', 'Isha']
['Hi', 'Lori', 'Patty', 'Victoria', 'Secret', 'Parkway', 'Plaza', 'Mall']
['Freeport']
['Hello', 'John', 'Freeport', 'Emery's"]
["I'll", 'H', "I'm", "I'm", 'Congress', 'Box', 'Denver', 'Colorado', 'Thank']
["I'm", "Marquis's", "I'll", "I'll", 'Friday']
['Well']
["Let's", 'Kerry', [redacted], 'Cardinal']
['Hi', 'Jeremy', [redacted], 'Cardinal', 'Health']
['Never', 'Carton', 'House']
['Levi', 'Dollar', 'General', "I'm", 'Meredith', [redacted]']
['Hot', 'Springs', "I'm", 'May', 'Thank']
['Lutetia', 'Lateesha', 'My', 'Sarah', [redacted], "I'm", "Dillard's", 'Department', 'Eastwood', 'Mall', "I'm", 'Sarah', [redacted]]
[[redacted], 'My', 'Sara', [redacted], "Dillard's", 'Department', 'Eastwood', 'Mall', "I'm", "I'd", 'Sarah', [redacted], 'Mall', 'Treo']
['Brett', 'Brett', 'My', [redacted], "Dillard's", 'Department', 'Eastwood', 'Mall', "Dillard's", 'Department', 'Store', 'Sarah', [redacted], 'Treo']
['Brad', 'Brad', 'My', 'Sara', [redacted], "I'm", "Dillard's", 'Department', "I'm", 'Please', 'My', 'Sarah', [redacted], "Dillard's", 'Eastwood', 'Mall', 'Thank']
[[redacted], 'Athens', "I'm"]
['Hi', 'My', 'Chris', [redacted].', 'St', 'Paul', 'Minnesota', 'Open', 'Mobile', 'Thank']
['Hi', 'Dollar', 'General', "I'm"]
['Hyla', 'Keesha', 'Hi', 'Melrose', 'Park', 'Please']
['Hi', 'Ms', [redacted], 'My', 'Marie', "I'm", 'Avis', 'Budget', 'Marie']
['Anton', 'My', 'Marie', "I'm", 'Avis', 'Budget', "I'd", 'Marie']
['Hi', 'Marie', 'Marie']
['Hi', 'Emily', "I'm", 'Ok', 'Thank']
['Lateesha', 'Jane', 'My', 'Sarah', [redacted], "I'm", "Dillard's", 'Department', 'Sarah', [redacted], 'Treo']
['Don', 'Don', 'My', 'Sarah', [redacted]ard's', 'Department', 'Eastwood', 'Mall', "I'm", 'March', "you're", 'Sarah', [redacted], "Dillard's", 'East']
['Brett', 'Brett', 'My', 'Sarah', [redacted], "I'm", "Dillard's", 'Department', "I'm", "I'm", 'Sarah', [redacted], "I'm", "Dillard's", 'Department', 'East']
['Brad', 'Brad', 'My', 'Sarah', [redacted], "Dillard's", 'Department', 'Store', 'Sarah', [redacted]]

01 Stanford NLP

Lower Number of Proper Nouns

- Avg of ~3 proper nouns
- Usually:
  a. employee name,
  b. location,
  c. company name,
  d. manager name

```
['Jacob', 'David', [redacted]']
['Laurie', 'Waste', 'Management', 'Isha', 'Kristin']
['Phil', 'Isha']
['Lori', 'Patty', 'Victoria', 'Secret', 'Parkway', 'Plaza', 'Mall']
['Freeport']
['John', 'Freeport', 'Emery']
['Congress', 'Box', 'Denver', 'Colorado']
['Marquis', 'Friday']
['j.l.']
['Kerry', '[redacted]', 'Cardinal', 'hell']
['Jeremy', '[redacted]', 'Cardinal', 'Health']
['Carton', 'House']
['Levi', 'Dollar', 'General', 'Meredith', '[redacted]']
['Hot', 'Springs', 'May']
['Lutetia', 'Lateesha', 'Sarah', '[redacted]', 'Dillard', 'Department', 'Eastwood', 'Mall',
['[redacted]', 'Sara', '[redacted]', 'Dillard', 'Department', 'Eastwood', 'Mall', 'Sarah', '[redacted]
['Brett', 'Brett', 'Sarah', '[redacted]', 'Dillard', 'Department', 'Eastwood', 'Mall', 'Dil
['Brad', 'Brad', 'Sara', '[redacted]', 'Dillard', 'Department', 'Sarah', '[redacted]', 'Dillarc
['[redacted]', 'Athens']
['Chris', 'j.b.', '[redacted]', 'St', 'Paul', 'Minnesota', 'Open', 'Mobile']
['Dollar', 'General']
['Hyla', 'Keesha', 'Melrose', 'Park']
['Ms', '[redacted]', 'Marie', 'Avis', 'Budget', 'Marie']
['Anton', 'Marie', 'Avis', 'Budget', 'Marie']
['Marie', 'Marie']
```



Thank you!

Jake Anderson
jakeanderson@g.ucla.edu
UC Data Love Week 2026 – February 10th, 2026