EXTRACTING INSIGHTS FROM JOB POSTINGS

A JOURNEY INTO NATURAL LANGUAGE PROCESSING

Capstone Project

David Antzelevitch

Thinkful Data Science Bootcamp

12/1/2018

THE PROBLEM: HR NEEDS HELP ASSESSING THE JOB MARKET

- <u>HR Departments Need Information</u>
- What skills are prevalent
- Where is the talent
- What attracts talent
- How are job descriptions written

<u>Current Challenges:</u>

- Job Postings are on many sites
- They are **UNSTRUCTURED**
- Difficult to gather insights

<u>Project Goal:</u>

- Develop a toolset that enables HR Departments to gather insight from Job Posting Data

apple's strategic data solutions (sds) team leads efforts to optimize various business processes and improve the customer journey across apple. we apply data science and machine learning to drive strategic impact across multiple lines of business at apple. we are looking for an excellent analyst who is passionate about working closely with partner teams; understanding their processes and analytic needs; and developing technical solutions along the way.\nyou will sit side by side with our data scientists and analyze complex business problems, have access to internal and external data sources, and be expected to communicate meaningful insights to senior leadership and key decision makers. you should work well in team driven environments with minimal formal structure and at ease in uncertain environments with opposing priorities. you should possess extraordinary business sense, a strong quantitative technical background, natural curiosity, and the ability to effectively shift between communications styles based on the audience (technical peer review through leadership update).\n\nkey qualifications\nexperience applying analytical techniques to provide solutions to real business problems\nexperience with sql and big data systems and tools\nprogramming skills in python or similar language\nexperience eliciting business requirements and executive metrics\nexperience building business cases and project plans\ngood social skills with ability to connect and develop positive partnerships\nstrong verbal written communication skills\n(desired) background in payments or financial auditing\ncreativity to go beyond current tools to deliver the best solution to the problem\ninquisitiveness and a real passion for continued self improvement and development of new skills\nable to work independently and make key decisions on projects\nability to provide relevant insights with data\nability to operate appropriately and effectively in a dynamic, highly matrix and fast paced environment.\ndescription\n support discovery of business problems\n perform data discovery and build proof of concepts\n perform ad hoc and reoccurring statistical analyses\n work with data warehouse architects and software developers to generate precise business intelligence solutions for business partners\n present results of analysis to business units\nevery single day, people do amazing things at apple. what will you do?\n\neducation\nbachelors masters in technical field (math, statistics, engineering, computer science, analytics, or similar)\napple is an eoe that is committed to inclusion and diversity. we also take affirmative action to offer employment and advancement opportunities to all applicants, including minorities, women, protected veterans, and individuals with disabilities. apple will not discriminate or retaliate against applicants who inquire about, disclose, or discuss their compensation or that of other applicants.

THE PROJECT

- 7000 job postings were collected
- This dataset is specific to data science jobs only
- The data is unstructured.
- Entire Job Posting is in a single block of text.

The Goal

Create tools to gather insights from unstructured Job Posting Data

CLEANING THE DATA (WITH REGULAR EXPRESSIONS)

Masters and Bachelors degrees are spelled in many different ways.

Regular Expressions allow us to standardize these spellings

Here is the regular expression that replaces all variations of bachelors with the word 'bachelors':

(^|(?<=\W))(b.\a|ba|bachelor|b\.\s\.|bs|b\.s)[s']?[s']?((?=\W)|$)

Need to be careful. Replacing 'ba' to 'bachelors' can be dangerous. Any word starting with 'ba' would be converted:

basic => bachelorssic

bachelors => bachelorschelors

To prevent this, used a **lookahead** (?=\W). Which says to match only on a 'ba' that is followed by a non-word character.

Variations of Bachelors

bachelor: 1303,
bs: 1091,
bachelor's: 748,
ba: 391,
b.s: 219,
bachelors: 208,
bas': 12,
bachelors': 3,
bachelor': 2

Variations of Masters

ms: 2159
master: 1302
master's: 623
masters: 449
m.s: 251
mss: 18
masters': 7
msss: 5

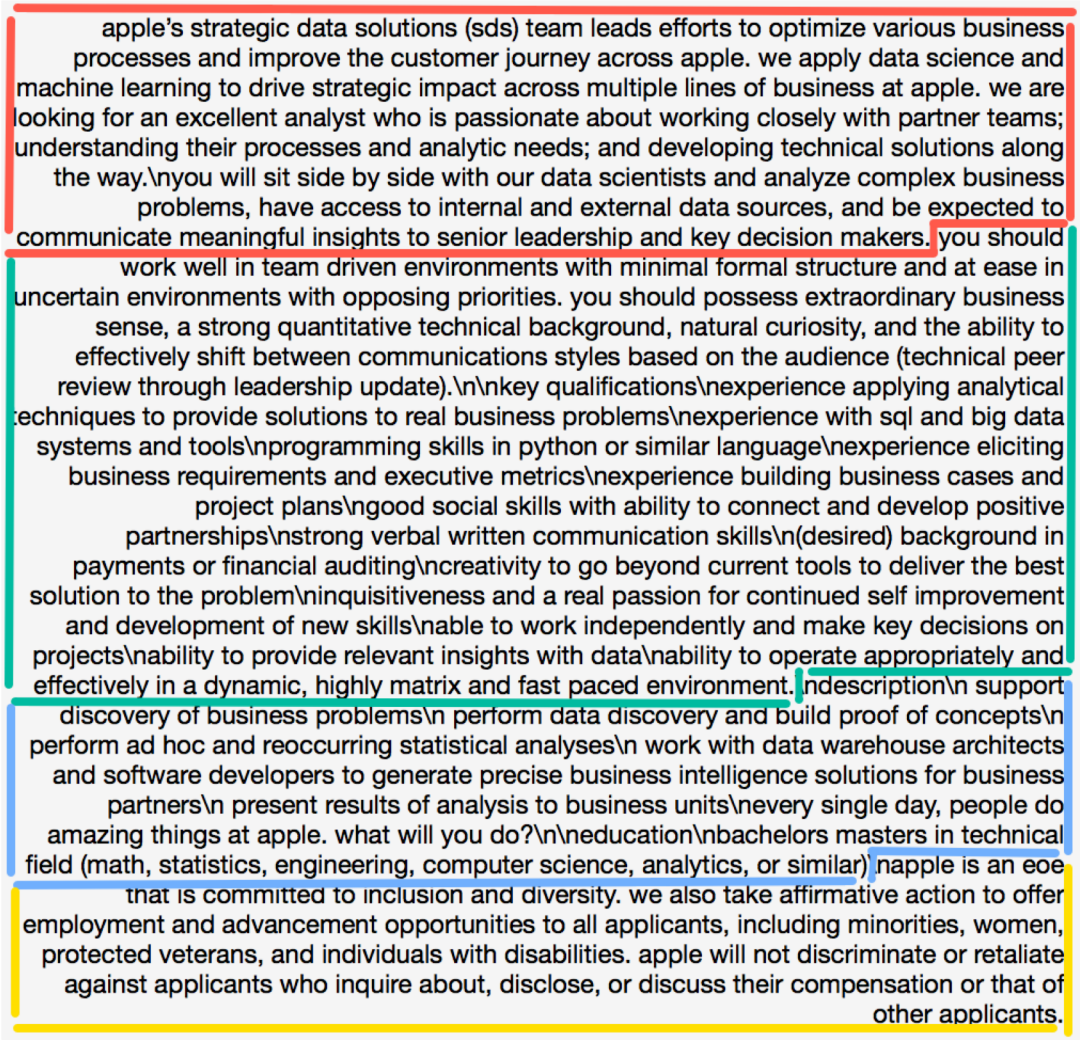
WHAT MAKES UP A JOB DESCRIPTION?

About the Company
- Industry
- Achievements
- Culture

The Ideal Candidate
- Experience
- Skills
- Education

About the Role
- Responsibilities

Legal Disclaimers (Equal Opp.)


apple's strategic data solutions (sds) team leads efforts to optimize various business processes and improve the customer journey across apple. we apply data science and machine learning to drive strategic impact across multiple lines of business at apple. we are looking for an excellent analyst who is passionate about working closely with partner teams; understanding their processes and analytic needs; and developing technical solutions along the way.\nyou will sit side by side with our data scientists and analyze complex business problems, have access to internal and external data sources, and be expected to communicate meaningful insights to senior leadership and key decision makers. you should work well in team driven environments with minimal formal structure and at ease in uncertain environments with opposing priorities. you should possess extraordinary business sense, a strong quantitative technical background, natural curiosity, and the ability to effectively shift between communications styles based on the audience (technical peer review through leadership update).\n\nkey qualifications\nexperience applying analytical techniques to provide solutions to real business problems\nexperience with sql and big data systems and tools\nprogramming skills in python or similar language\nexperience eliciting business requirements and executive metrics\nexperience building business cases and project plans\ngood social skills with ability to connect and develop positive partnerships\nstrong verbal written communication skills\n(desired) background in payments or financial auditing\ncreativity to go beyond current tools to deliver the best solution to the problem\ninquisitiveness and a real passion for continued self improvement and development of new skills\nable to work independently and make key decisions on projects\nability to provide relevant insights with data\nability to operate appropriately and effectively in a dynamic, highly matrix and fast paced environment. \ndescription\n support discovery of business problems\n perform data discovery and build proof of concepts\n perform ad hoc and reoccurring statistical analyses\n work with data warehouse architects and software developers to generate precise business intelligence solutions for business partners\n present results of analysis to business units\nevery single day, people do amazing things at apple. what will you do?\n\neducation\nbachelors masters in technical field (math, statistics, engineering, computer science, analytics, or similar) \napple is an eoe that is committed to inclusion and diversity. we also take affirmative action to offer employment and advancement opportunities to all applicants, including minorities, women, protected veterans, and individuals with disabilities. apple will not discriminate or retaliate against applicants who inquire about, disclose, or discuss their compensation or that of other applicants.

UNSUPERVISED LEARNING TO CLASSIFY COMPONENTS OF A JOB DESCRIPTION

Train the Model

- Separate training set into sentences
- Apply bag of words features to each sentence
- Reduce 1000+ word features to 2-3 features via PCA
- Iterate until the 2-3 features classify into parts of the description (ie. Job, responsibilities, qualifications)

Test the Model

Feed in an unrecognized job posting. The model is able to classify each sentence.

	About the Candidate	About the Job
apple's strategic data solutions (sds) team leads efforts to optimize various business processes and improve the customer journey across apple. we apply data science and machine learning to drive strategic impact across multiple lines of business at apple. we are looking for an excellent analyst who is passionate about working closely with partner teams; understanding their processes and analytic needs; and developing technical solutions along the way.	0.090794	0.202386
you will sit side by side with our data scientists and analyze complex business problems, have access to internal and external data sources, and be expected to communicate meaningful insights to senior leadership and key decision makers. you should work well in team driven environments with minimal formal structure and at ease in uncertain environments with opposing priorities. you should possess extraordinary business sense, a strong quantitative technical background, natural curiosity, and the ability to effectively shift between communications styles based on the audience (technical peer review through leadership update).	0.078417	0.170804
key qualifications experience applying analytical techniques to provide solutions to real business problems	0.220290	-0.089559
experience with sql and big data systems and tools	0.351519	0.141982
programming skills in python or similar language	0.037130	0.006807
experience eliciting business requirements and executive metrics	0.412168	-0.170468
experience building business cases and project plans	0.341119	-0.146551

FROM SENTENCES CLASSIFIED AS "ABOUT THE CANDIDATE", WE CAN GET N-GRAMS DESCRIBING PREFERRED EXPERIENCE

	count
machine learning	3521
computer science	2634
natural language	989
problem solving	894
language processing	832
programming language	824
natural language processing	798
software development	785
data science	727
deep learning	677
large scale	602
data analysis	595
fast paced	589
big data	586
data scientist	520
molecular biology	502
project management	481
r python	468
c c++	449
cross functional	448
software engineering	422
python r	413
science engineering	413
operation research	412
statistical analysis	412
electrical engineering	381
new york	362
language python	357
data set	352
genetic information	330
cell culture	317
u. s.	316
scripting language	307
data mining	304
artificial intelligence	304
product development	302
computer vision	293
large data	286
learning algorithm	286
open source	283
cutting edge	280
computer science engineering	278
characteristic protected	276
san francisco	275
mathematics statistic	273
microsoft office	272
life science	272
drug discovery	271
science statistic	269
clinical trial	266
neural network	264
public health	263
science mathematics	262
engineering computer	259
dod clearance	256
machine learning algorithm	255
relational database	251
python java	250

Can we dig deeper into programming languages?

TO DIG DEEPER INTO THE DATA, WE BUILD A VOCABULARY USING WORD2VEC

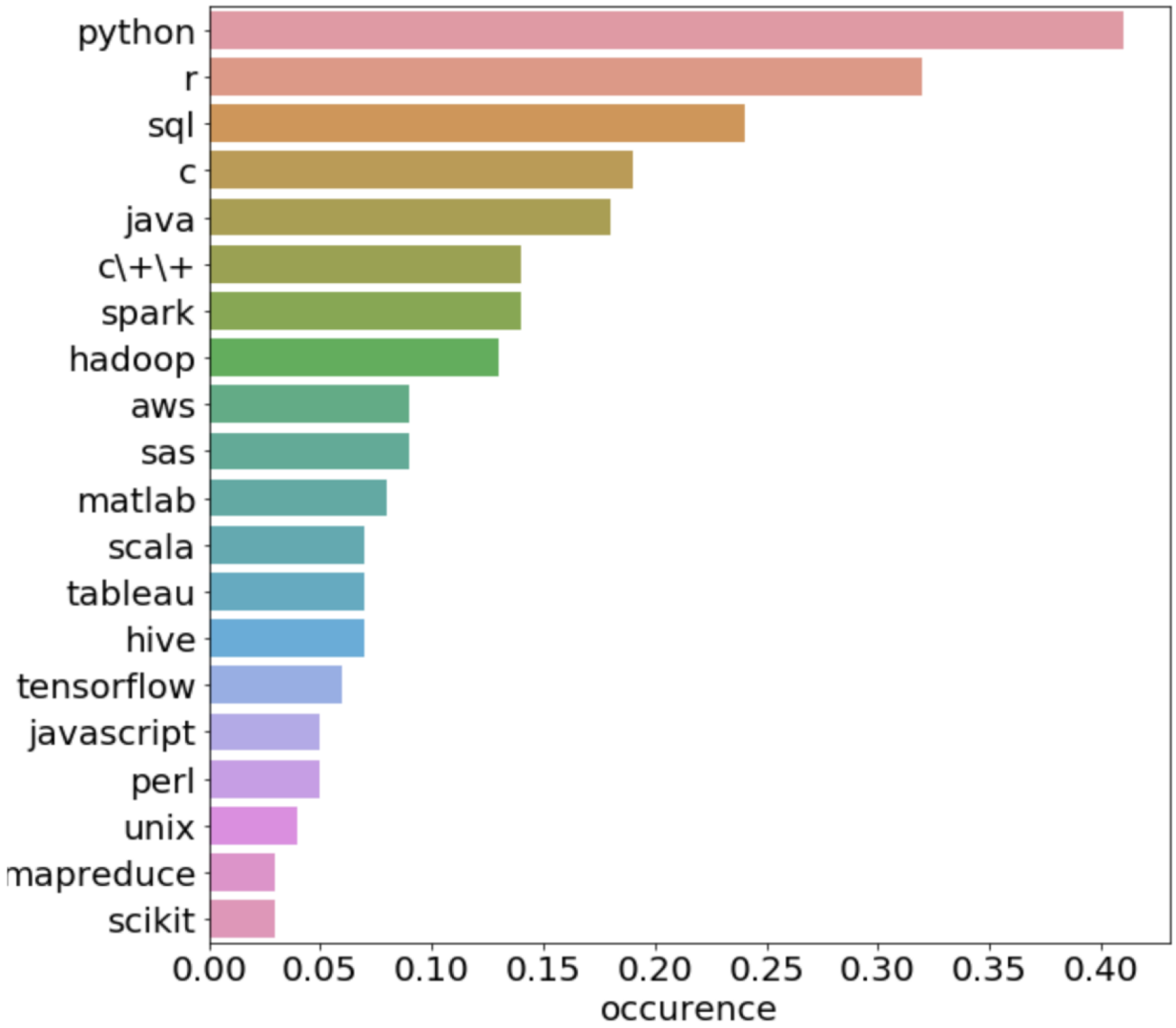
- The word2vec algorithm scans all 6000 job descriptions
- It assigns a vector to each word
  - Ie. Python = [0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0]
- Each vector explains how similar the word is to all other words in the corpus.
- With this vocabulary, we can now dig in deeper.
  - Ie. Show me all words like Python and R

```
np.array(model.wv.most_similar(positive=['python','r'],
                               topn=100))[:,0]
```

```
array(['scripting', 'programming', 'jmp', 'sas', 'matlab', 'bash', 'spss',
       'ruby', 'labview', 'java', 'vba', 'c', 'julia', 'packages', 'php',
       'perl', 'scala', 'jupyter', 'fluent', 'stata', 'coding', 'pandas',
       'c++', '.net', 'shiny', 'toolkits', 'libraries', 'sql',
       'languages', 'd3', 'fluency', 'javascript', 'similar', 'querying',
       'cuda', 'numpy', 'tableau', 'shell', 'django', 'preferable',
       'proficiency', 'flask', 'hive', 'scipy', 'elasticsearch', 'looker',
       'xml', 'prism', 'scikit', 'frameworks', 'probability',
       'computation', 'git', 'unix', 'v', 'apache', 'html5', 'spark',
       'tensorflow', 'kafka', 'docker', 'query', 'spring', 'programmer',
       'postgresql', 'version', 'tools', 'h2o', 'statistical', 'hadoop',
       'svm', 'angular', 'postgres', 'intermediate', 'node', 'numerical',
       'mastery', 'proficient', 'caffe', 'english', 'relational', 'torch',
       'cassandra', 'pig', 'visualization', 'github', 'mapreduce',
       'coursework', 'stats', 'redshift', 'helpful', 'mathematics',
       'json', 'macros', 'manipulation', 'either', 'rdbms', 'react', 'js',
       'cvpr'], dtype='<U19')
```

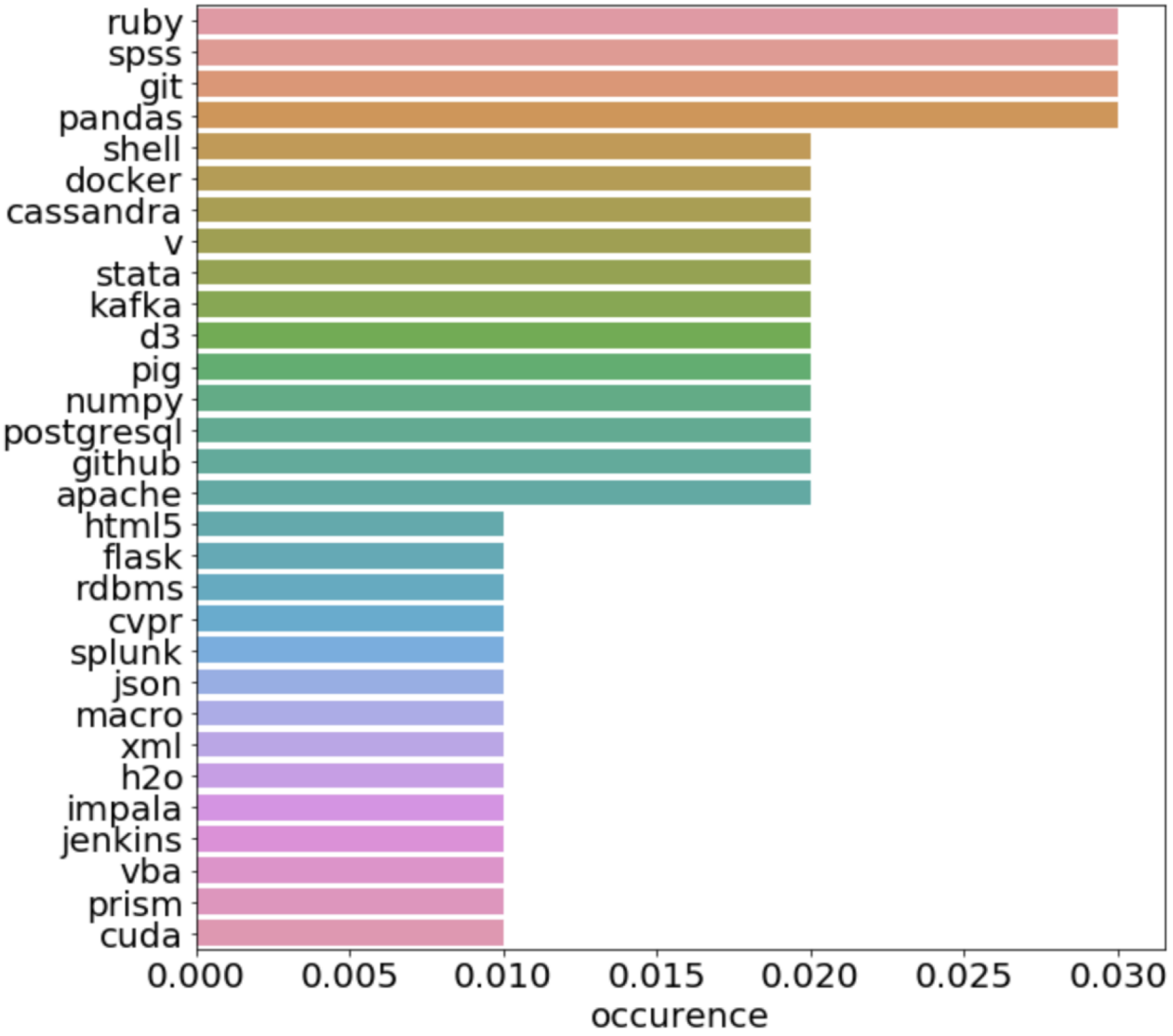
TOP DATA SCIENCE TECHNOLOGIES

- Now that we have a list of all technologies listed in the corpus
- We can see how often each technology appears in a job description
- 40% of all data science job postings contain the word python



AND THE LESS COMMON TECHNOLOGIES

- Yes, there is a programming platform called pig
- And v
- And h2o



GETTING OTHER INFORMATION USING WORD2VEC VOCABULARY

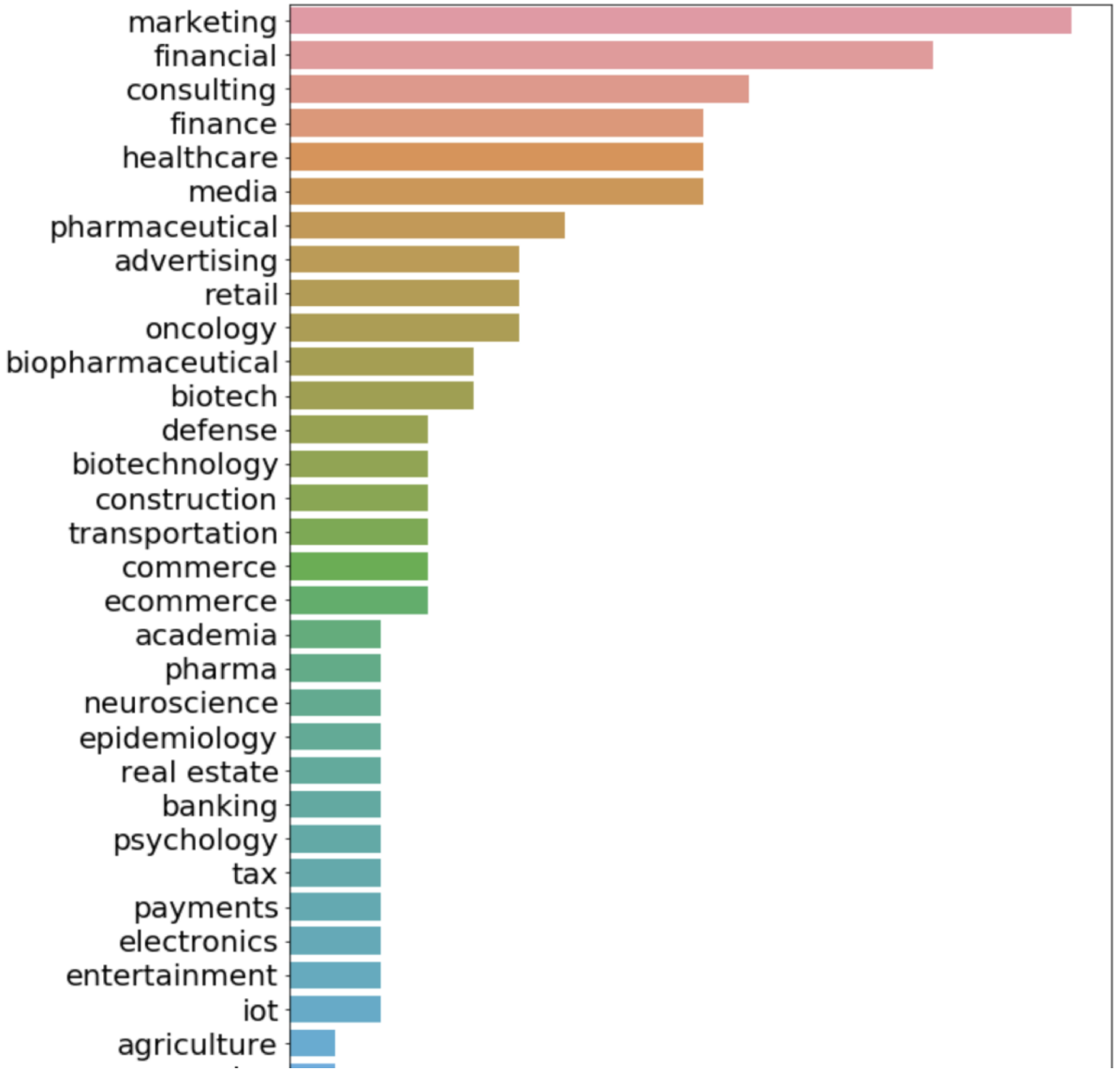
- In which INDUSTRIES are the data science jobs?
- Search the word2vec vocabulary for terms similar to automotive and banking

```
np.array(model.wv.most_similar(positive=['automotive','banking'],
                               topn=100))[:,0]
```

```
array(['retail', 'cpg', 'telecom', 'securities', 'lending', 'sector',
       'loans', 'markets', 'credit', 'consumer', 'industries',
       'investment', 'finance', 'manufacturers', 'commercial', 'cdk',
       'bank', 'gas', 'b2b', 'chase', 'entertainment', 'wealth', 'tech',
       'fraud', 'financial', 'vertical', 'retailers', 'sectors',
       'specialty', 'advisory', 'buying', 'services', 'market',
       'verticals', 'agriculture', 'valuation', 'derivatives',
       'utilities', 'healthcare', 'asset', 'portfolios', 'diversified',
       'consulting', 'income', 'global', 'pricing', 'aviation',
       'industrial', 'investors', 'medicare', 'assets', 'capital',
       'corporate', 'investing', 'venture', 'accounting', 'liability',
       'returns', 'dollar', 'payments', 'pharmaceutical', 'ecommerce',
       'selling', 'major', 'marketplace', 'p', 'provider',
       'manufacturing', 'technology', 'categories', 'companies',
       'private', 'advertising', 'threat', 'trillion', 'merchandising',
       'leading', 'transaction', 'saas', 'logistics', 'ge', 'fidelity',
       'positioning', 'llc', 'investments', 'radio', 'food', 'contracts',
       'ratings', 'wireless', 'brands', 'supply', 'pharma',
       'pharmaceuticals', 'websites', 'equity', 'cybersecurity',
       'defense', 'corporations', 'card'], dtype='<U19')
```

INDUSTRIES NEEDING DATA SCIENTISTS

- Marketing wins the top spot
- Healthcare / Pharmaceutical / Medical is popular
- Also some govt. jobs such as defense, tax.



JOB DESCRIPTIONS

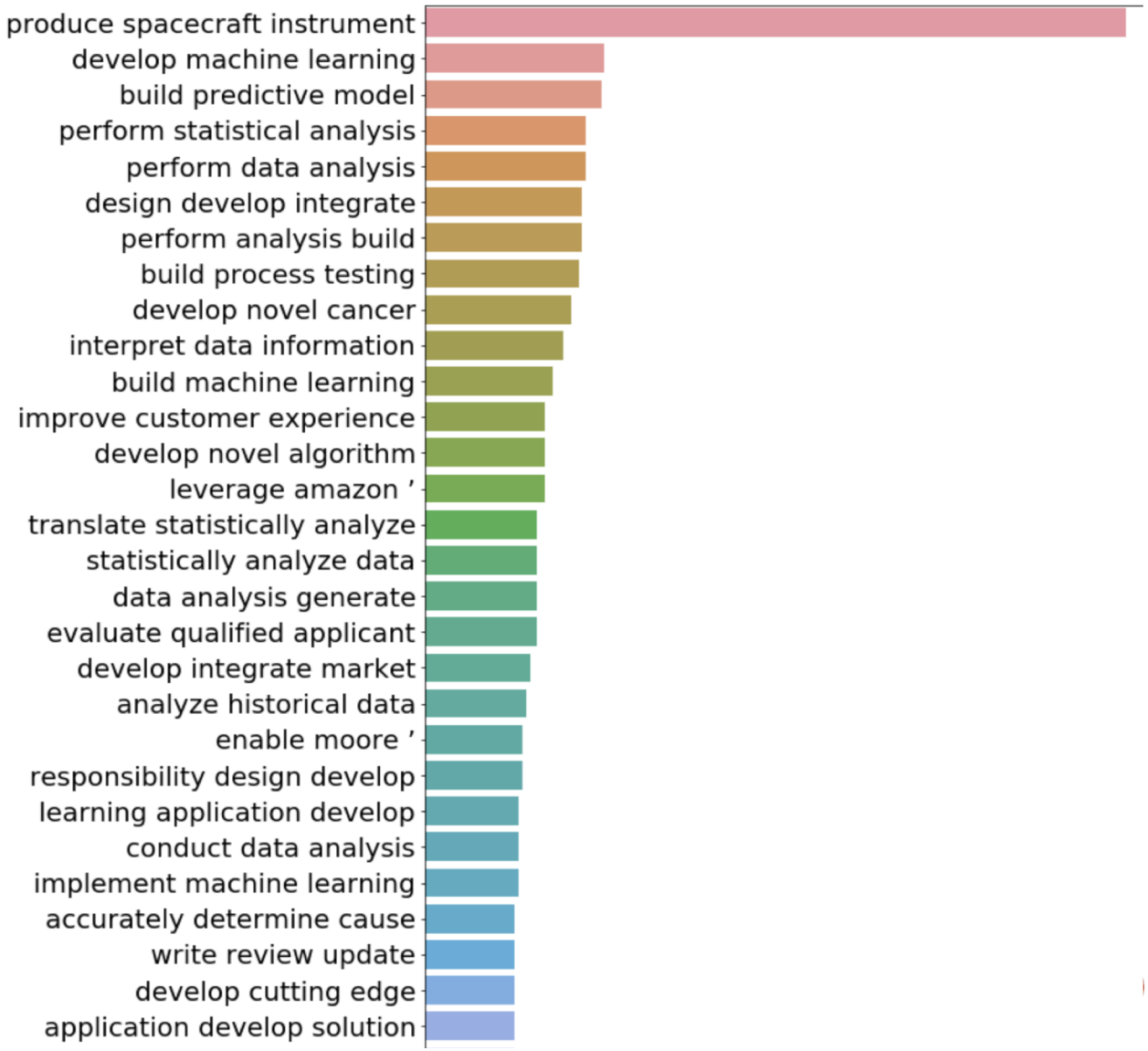
- Single words are not very useful to gain insights into job descriptions
- Instead, we need n-grams
- First, using word2vec we get a list of verbs in our corpus that describe a job activity, such as “perform”, “build”, “create”

```
np.array(model.wv.most_similar(positive=['perform', 'build', 'create'],
                               topn=100))[:,0]
```

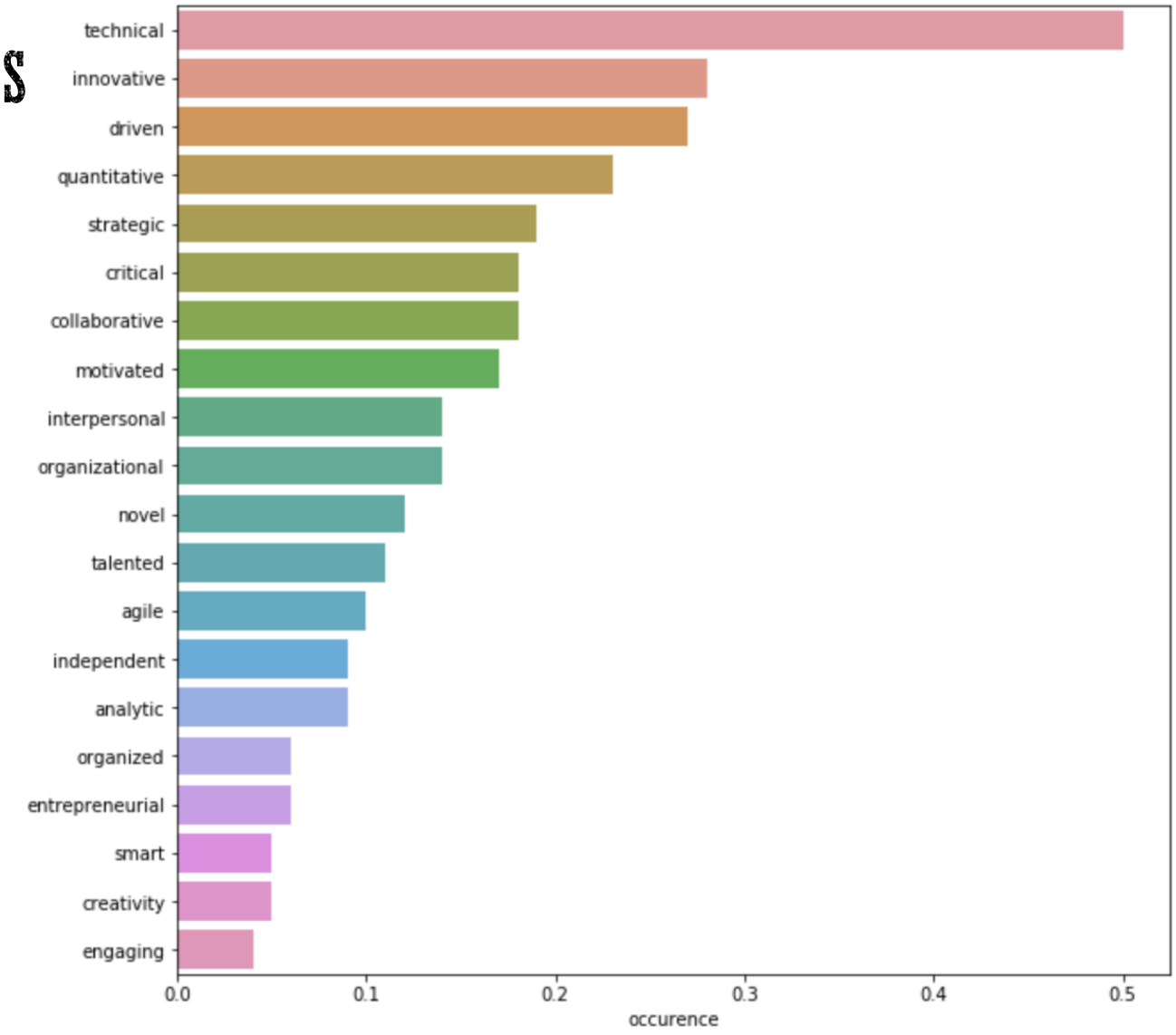
```
array(['develop', 'generate', 'deliver', 'implement', 'craft', 'define',
       'produce', 'provide', 'leverage', 'establish', 'construct', 'run',
       'write', 'enable', 'deploy', 'drive', 'conduct', 'formulate',
       'contribute', 'creating', 'utilize', 'develops', 'generates',
       'enhance', 'integrate', 'evaluate', 'devise', 'validate', 'find',
       'builds', 'acquire', 'building', 'execute', 'manage', 'prepare',
       'investigate', 'allow', 'bring', 'streamline', 'expand', 'add',
       'creates', 'configure', 'adapt', 'automate', 'lead', 'uses',
       'analyze', 'provides', 'derive', 'achieve', 'facilitate',
       'understand', 'uncover', 'optimize', 'improve', 'conceptualize',
       'developing', 'incorporate', 'complete', 'operate', 'capture',
       'organize', 'providing', 'performs', 'gain', 'ship', 'guide',
       'extract', 'predict', 'connect', 'navigate', 'delivering',
       'engage', 'empower', 'inform', 'debug', 'prototype', 'propose',
       'adopt', 'identify', 'educate', 'determine', 'leverages',
       'oversee', 'teach', 'seek', 'delight', 'make', 'transform',
       'architect', 'relate', 'evolve', 'refine', 'use', 'collaborate',
       'document', 'describe', 'fulfill', 'train'], dtype='<U19')
```

JOB DESCRIPTIONS

- Using that list of verbs, we then use nltk ngram utility to find 3 word n-grams containing those words

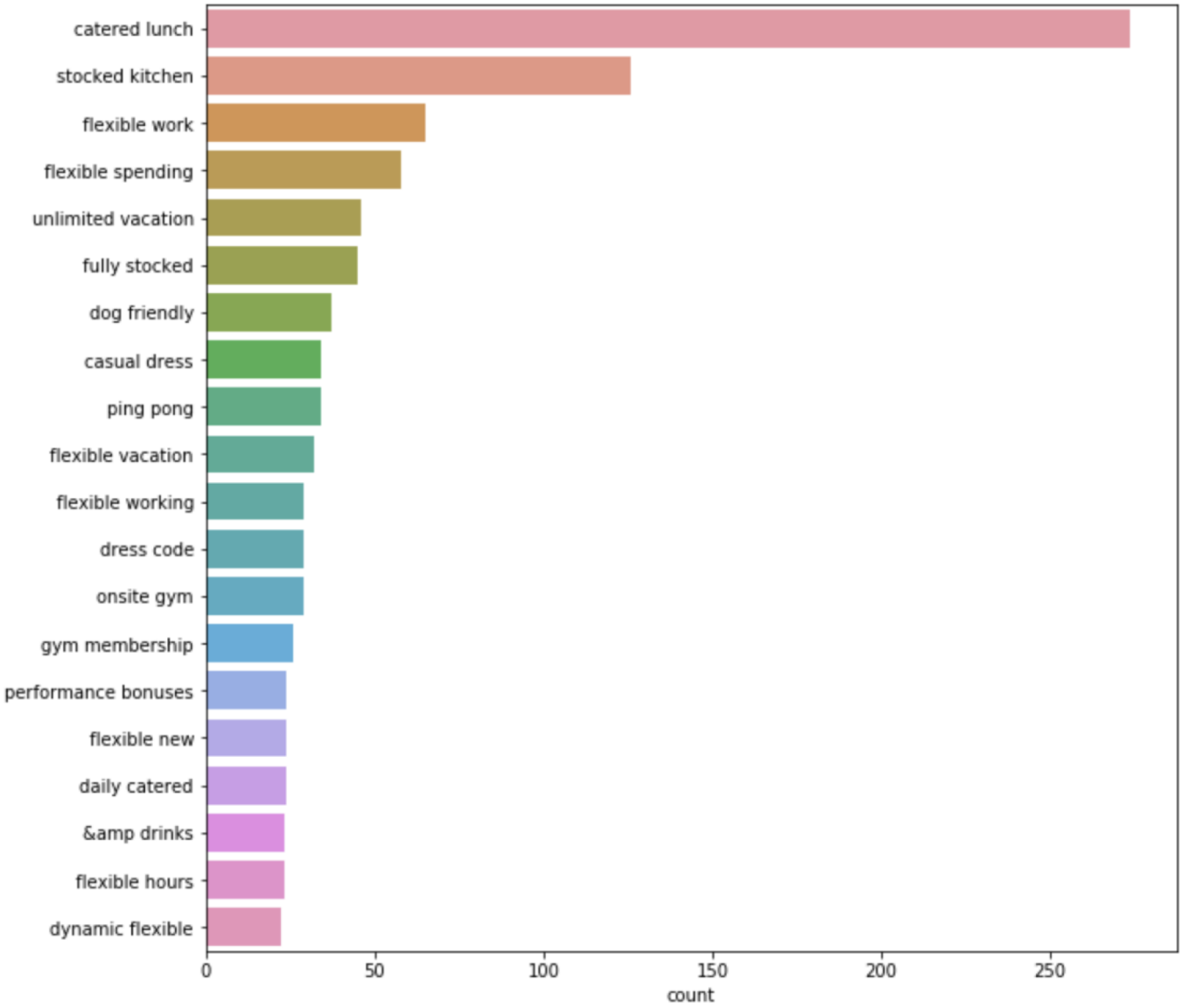


PERSONALITY ATTRIBUTES OF DATA SCIENTISTS

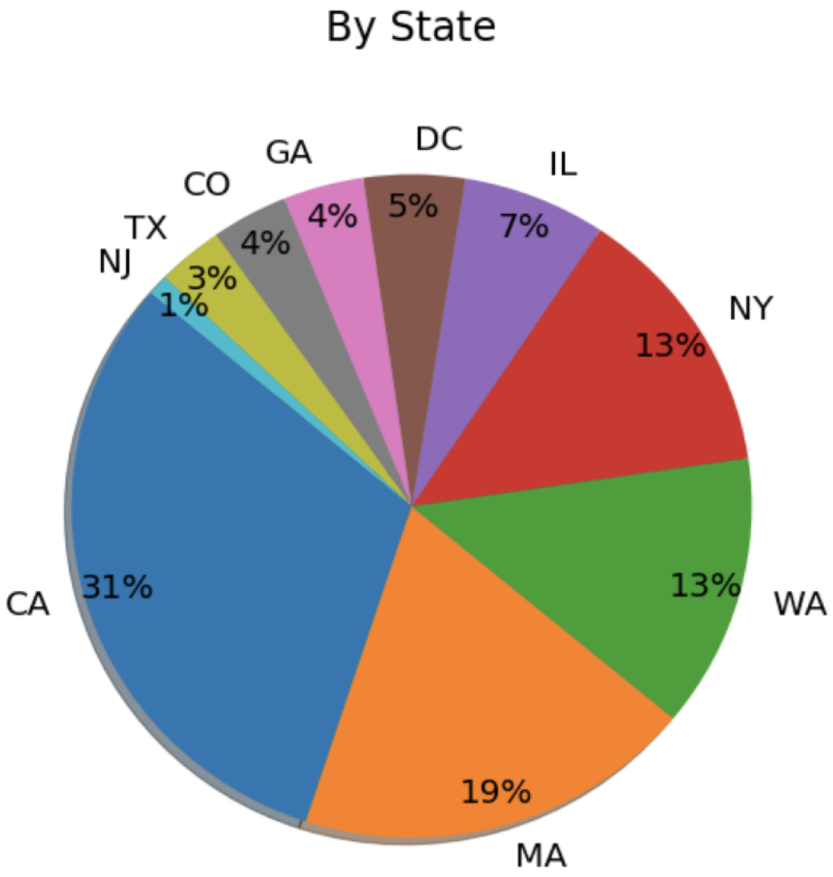


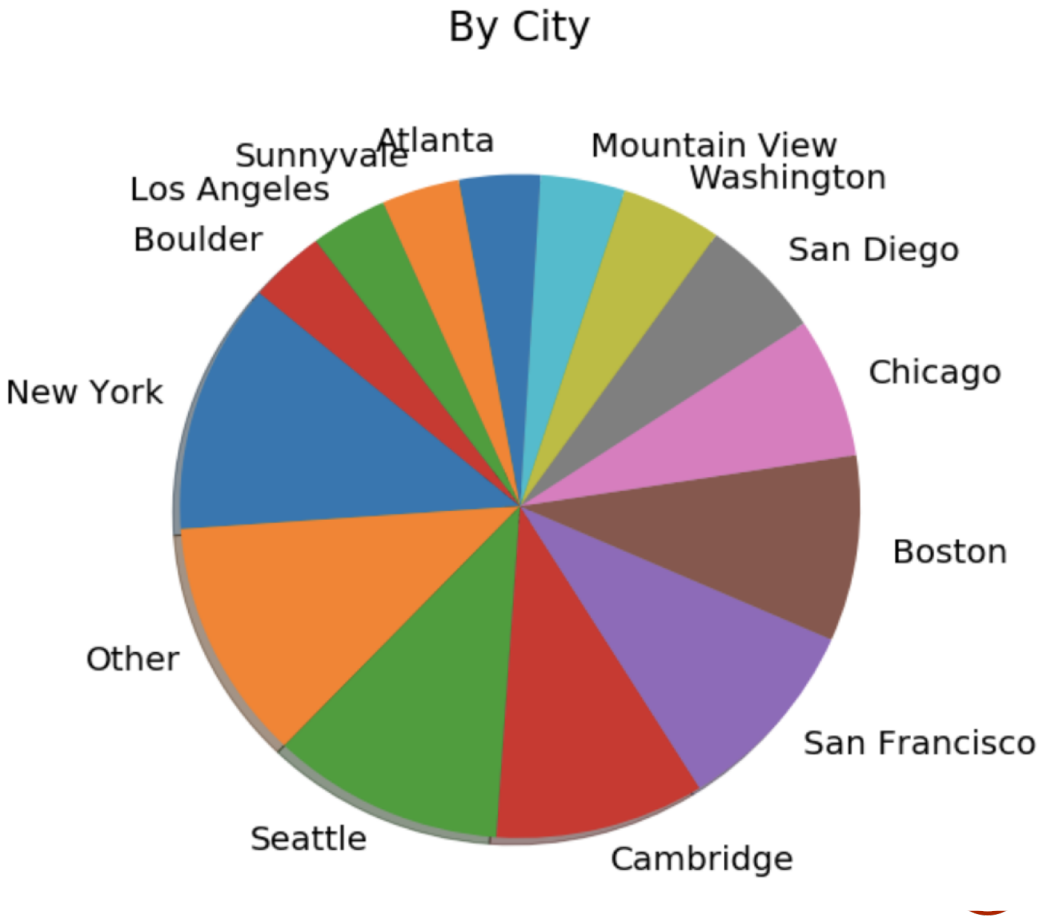
HOW ABOUT OFFICE PERKS?

- 250 out the 7000 jobs offered daily catered lunches

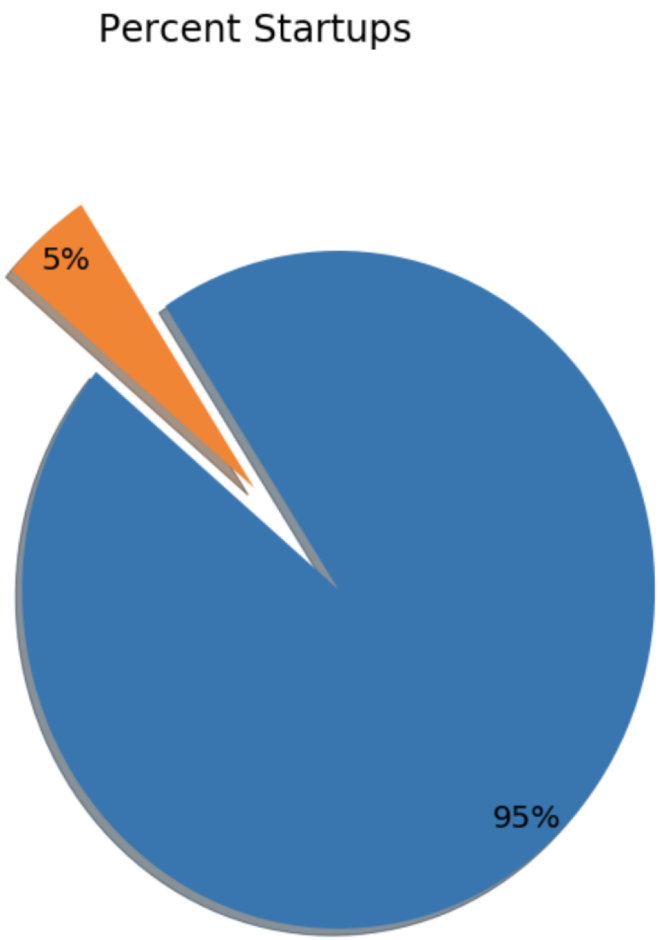


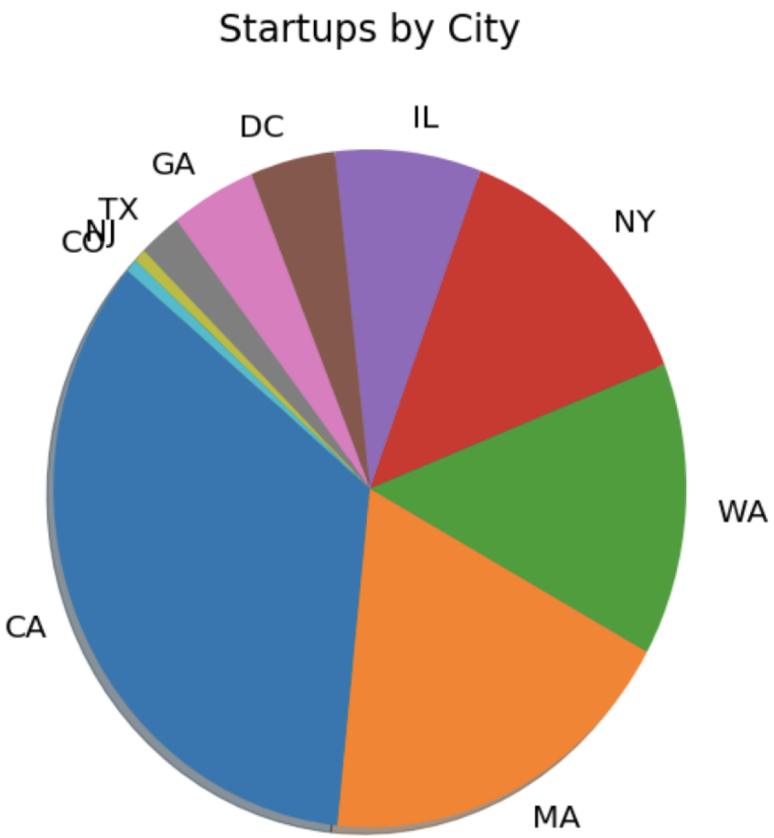
WHERE ARE THE DATA SCIENCE JOBS?





DO START UPS SEEK DATA SCIENTISTS?





CONCLUSION

- Gaining insights from unstructured natural language is a task for machine learning

- We were able to gather insights using tools such as:
  - Bag of Words combined with Dimensionality Reduction
  - Word2vec Similarity
  - Nltk ngrams
  - Regular expressions

at yapstone, we approach payments with the same startup mentality that we had when we launched our first payment solution in 1999. we are now focused on combining our entrepreneurial spirit with our immense payment expertise to take our company and our partners to the forefront of innovation. as one of the leading payment companies, yapstone is continually searching for passionate thinkers to join us in changing how the world pays.\nwe are looking for a data scientist who will support our risk and technology teams with insights gained from analyzing company data. the ideal candidate is adept at using large data sets to find opportunities for product and process optimization and using models to test the effectiveness of different courses of action. they must have strong experience using a variety of data mining data analysis methods, using a variety of data tools, building and implementing models, using creating algorithms and creating running simulations. they must have a proven ability to drive business results with their data based insights. they must be comfortable working with a wide range of stakeholders and functional teams. the right candidate will have a passion for discovering solutions hidden in large data sets and working with stakeholders to improve business outcomes.\nprimary responsibilities:\ndesigns, develops and programs methods, processes, and systems to consolidate and analyze unstructured, diverse "big data" sources to generate actionable insights and solutions for client services and product enhancement.\nanalyzes complex business problems and issues using data from internal and external sources to provide insight to decision makers.\nidentifies and interprets trends and patterns in datasets to locate influences.\nconstructs forecasts, recommendations and strategic tactical plans based on business data and market knowledge.\ncreates specifications for reports and analysis based on business needs and required or available data elements.\nmay provide consultation to users and lead cross functional teams to address business issues.\nmay directly produce datasets and reports for analysis using system reporting tools.\nuses advanced mathematical and statistical concepts and theories to analyze and collect data and construct solutions to business problems.\nperforms complex statistical analysis on experimental or business data to validate and quantify trends or patterns identified by business analysts.\nconstructs predictive models, algorithms and probability engines to support data analysis or product functions; verifies model and algorithm effectiveness based on real world results.\ndesigns experiments and methodologies to generate and collect data for business use.\nprojects may include a focus on "quantitative finance" or help identify new business opportunities.\nrequirements:\nwe're looking for someone with 5 7 years of experience manipulating data sets and building statistical models, and has a masters's or phd in statistics, mathematics, computer science or another quantitative field.\nprevious experience with either fraud or payments required.\nknowledge and experience in statistical and data mining techniques\nstrong problem solving skills with an emphasis on product development.\nexperience using statistical computer languages (r, python, slq, etc.) to manipulate data and draw insights from large data sets.\nexperience working with and creating data architectures.\nknowledge of a variety of machine learning techniques (clustering, decision tree learning, artificial neural networks, etc.) and their real world advantages drawbacks.\nknowledge of advanced statistical techniques and concepts (regression, properties of distributions, statistical tests and proper usage, etc.) and experience with applications.\nexcellent written and verbal communication skills for coordinating across teams.\na drive to learn and masters new technologies and techniques.\nexperience querying databases and using statistical computer languages: r, python, slq, etc.\nchanging how the world pays is a mission that inspires us daily. it gets us up in the morning and keeps us up at night. and just in case we need an extra dose of inspiration during the day, our team can take in the stellar view from our santa monica office or visit the yap café in our walnut creek office to choose from a notably large selection of snacks. from comprehensive health insurance to gym memberships, you'll find plenty of benefits and perks that reflect our appreciation for all of the thinking and doing that goes on at yapstone.\n\nyapstone is a global provider of full stack payment solutions for global marketplaces and large vertical markets. yapstone powers online and mobile payments for homeaway®, vrbo®, and thousands of apartment and vacation rental companies, homeowners' associations, self storage companies, hospitality establishments and non profits. yapstone processes over $14b in payment volume annually and has been recognize...