TIFFANY FRENCH

# THINKFUL FINAL CAPSTONE

# PROJECT GOALS & SCOPE

- To analyze job postings for potentially biased language, which may be a cause of very gender-skewed jobs.
- Scrape job postings, analyze with supervised and unsupervised NLP techniques.
- This could be the basis for a “Turn-it-in” Style tool that could take text input, and provide analysis and suggestions for neutralizing the language.

# Gender Gaps and AI Skills

WORLD ECONOMIC FORUM

## Skills where women outnumber men

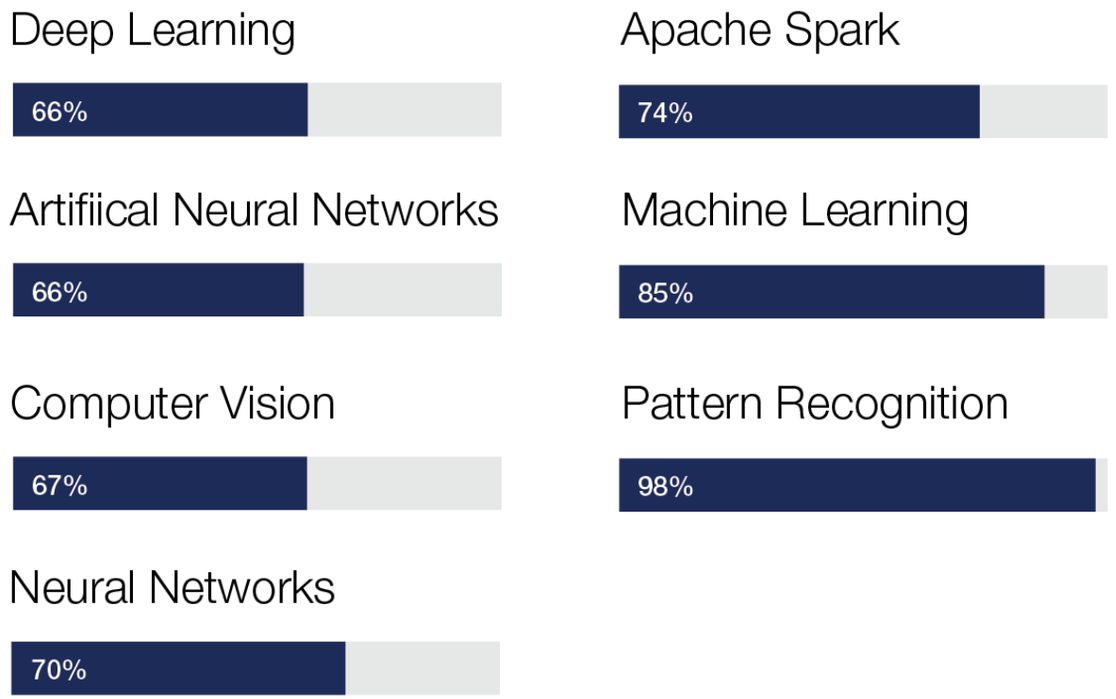
Text analytics

Text Mining

Speech Recognition

National Language Processing

## Skills where men outnumber women




Source: LinkedIn data featured in the
Global Gender Gap Report 2018, World Economic Forum

# Why Women Don't Apply for Jobs Unless They're 100% Qualified

by Tara Sophia Mohr

AUGUST 25, 2014

SAVE SHARE 212 COMMENT TEXT SIZE PRINT $8.95 BUY COPIES

You've probably heard the following statistic: Men apply for a job when they meet only 60% of the qualifications, but women apply only if they meet 100% of them.

The finding comes from a Hewlett Packard internal report, and has been quoted in *Lean In*, *The Confidence Code* and dozens of articles. It's usually invoked as evidence that women need more confidence. As one Forbes article put it, "Men are confident about their ability at 60%, but women don't feel confident until they've checked off each item on the list." The advice: women need to have more faith in themselves.

I was skeptical, because the times *I* had decided not to apply for a job because I didn't meet all the qualifications, faith in myself wasn't exactly the issue. I suspected I wasn't

SCIENCE

# The More Gender Equality, the Fewer Women in STEM

A new study explores a strange paradox: In countries that empower women, hey are less likely to choose math and science professions.

LGA KHAZAN FEB 18, 2018

rtist Tjaasa Gusfors poses next to her Jesus Christ ice statue near Stockholm in 2011. (TT NEWS AGENCY / REUTERS)

Male-oriented titles can inadvertently prevent women from clicking on your job in a list of search res
Avoid including words in your titles like "hacker," "rockstar," "superhero," "guru," and "ninja," and use ne
descriptive titles like "engineer," "project manager," or "developer."

## 2.) Check pronouns.

When describing the tasks of the ideal candidate, use "S/he" or "you." Example: "As Product Manager
XYZ, you will be responsible for setting the product vision and strategy."

## 3.) Avoid (or balance) your use of gender-charged words.

Analysis from language tool Textio found that the gender language bias in your job posting predicts th
gender of the person you're going to hire. Use a tool like Textio tool or the free Gender Decoder to ide
problem spots in your word choices. Examples: "Analyze" and "determine" are typically associated wit
traits, while "collaborate" and "support" are considered female. Avoid aggressive language like "crush

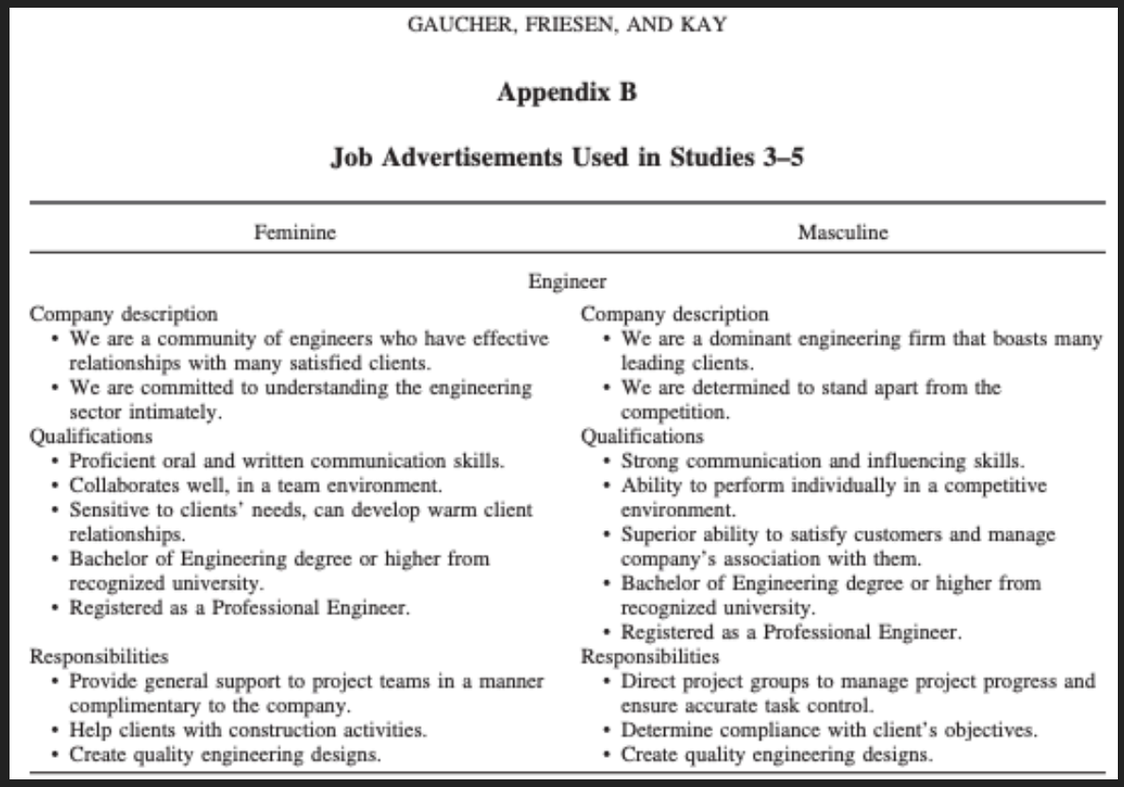
## 4.) Avoid superlatives.

Excessive use of superlatives such as "expert," "superior," "world class" can turn off female candidates
are more collaborative than competitive in nature. Research also shows that women are less likely tha
men to brag about their accomplishments. In addition, superlatives related to a candidate's backgrou
limit the pool of female applicants because there may be very few females currently in leading positio
"world class" firms.

## 5.) Limit the number of requirements.

# EXAMPLES OF GENDERED LANGUAGE

| **Masculine:** | **Feminine:** |
|---|---|
| Active | Communal |
| Domina* | Connect* |
| Decisive | Cooperative |
| Analy* | Interdepend* |
| Objective | Support* |
| Self-reliant | Together* |



GAUCHER, FRIESEN, AND KAY

**Appendix B**

**Job Advertisements Used in Studies 3–5**

| Feminine | Masculine |
|---|---|
| Engineer | |
| Company description<br>• We are a community of engineers who have effective relationships with many satisfied clients.<br>• We are committed to understanding the engineering sector intimately. | Company description<br>• We are a dominant engineering firm that boasts many leading clients.<br>• We are determined to stand apart from the competition. |
| Qualifications<br>• Proficient oral and written communication skills.<br>• Collaborates well, in a team environment.<br>• Sensitive to clients' needs, can develop warm client relationships.<br>• Bachelor of Engineering degree or higher from recognized university.<br>• Registered as a Professional Engineer. | Qualifications<br>• Strong communication and influencing skills.<br>• Ability to perform individually in a competitive environment.<br>• Superior ability to satisfy customers and manage company's association with them.<br>• Bachelor of Engineering degree or higher from recognized university.<br>• Registered as a Professional Engineer. |
| Responsibilities<br>• Provide general support to project teams in a manner complimentary to the company.<br>• Help clients with construction activities.<br>• Create quality engineering designs. | Responsibilities<br>• Direct project groups to manage project progress and ensure accurate task control.<br>• Determine compliance with client's objectives.<br>• Create quality engineering designs. |

Gaucher, D., Friesen, J., & Kay, A. C. (2011, March 7). Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality. Journal of Personality and Social Psychology

# DATASET

- Text analysis of job postings from indeed.com to assess for possible gender-biased language
- The job types are:
  - **Female**: Text Analytics, Text Mining, Speech Recognition, NLP
  - **Male**: Machine Learning, Apache Spark, Pattern Recognition, Neural Networks
- Techniques used:
  - Beautiful Soup
    - I scraped over 7,800 job postings from indeed.com with an iterative scraper that worked through hundreds of pages of job postings.
  - Due to duplicates (I.e. an NLP/Machine Learning posting) the dataset was reduced to 4,300.
  - Additionally, I removed one of the job types (computer vision) to reduce the possibility of class imbalance.  Female fields represented 34% of the dataset. The dataset ultimately consisted of 3700 postings.

# NOTEBOOKS AND CODE

```
starts = list(range(700, 1000, 10))
requests = 0
start = time.time()

baseurl = 'https://www.indeed.com/'

nlp_jobs = []
for start in starts:
    my_urls = ('https://www.indeed.com/jobs?q=%22machine+learning%22&start=' + str(start),)
    my_url = my_urls[0]
    for my_url in my_urls:
        uClient = urlopen(my_url)
        html_input = uClient.read()
        uClient.close()
        soup = BeautifulSoup(html_input, "html.parser")
        cards = soup.findAll('div', {'class':'jobsearch-SerpJobCard'})
        it = iter(cards)
        next(it) # ads
        next(it) # ads
        #next(it)
        for curr in it:
            try:
                link = curr.find('h2').find('a', href=True)['href']
            except:
                pass
            with urlopen(baseurl + link) as uClient:
                list_url = uClient.read()
            listing = BeautifulSoup(list_url, 'html.parser')
            title = listing.find('h3',
                                 {'class': 'icl-u-xs-mb--xs icl-u-xs-mt--none jobsearch-JobInfoHeader-t
itle'})
            if not title:
                    print('missing content @ ' + baseurl + link)
            body = listing.find('div',
                                {'class': 'jobsearch-JobComponent-description icl-u-xs-mt--md'}
                                )
            if not body:
                print('missing content @ ' + baseurl + link)
            requests += 1
            sleep(randint(5,7))
            end = time.time()
            #print("Done in", end, "seconds")
            print('Request: {}; Frequency: {} requests/s'.format(requests, requests/end))
            clear_output(wait = True)
            with db_session:
                Job(title=str(title),
                job_description=str(body),
                job_class='Machine Learning')
```

```
GET CONNECTION FROM THE LOCAL POOL
BEGIN IMMEDIATE TRANSACTION
INSERT INTO "Job" ("title", "job_description", "job_class") VALUES (?, ?, ?)
```

# 8 DIFFERENT JOB TITLES

# BEAUTIFUL SOUP SCRAPER

21 lines (14 sloc) | 281 Bytes

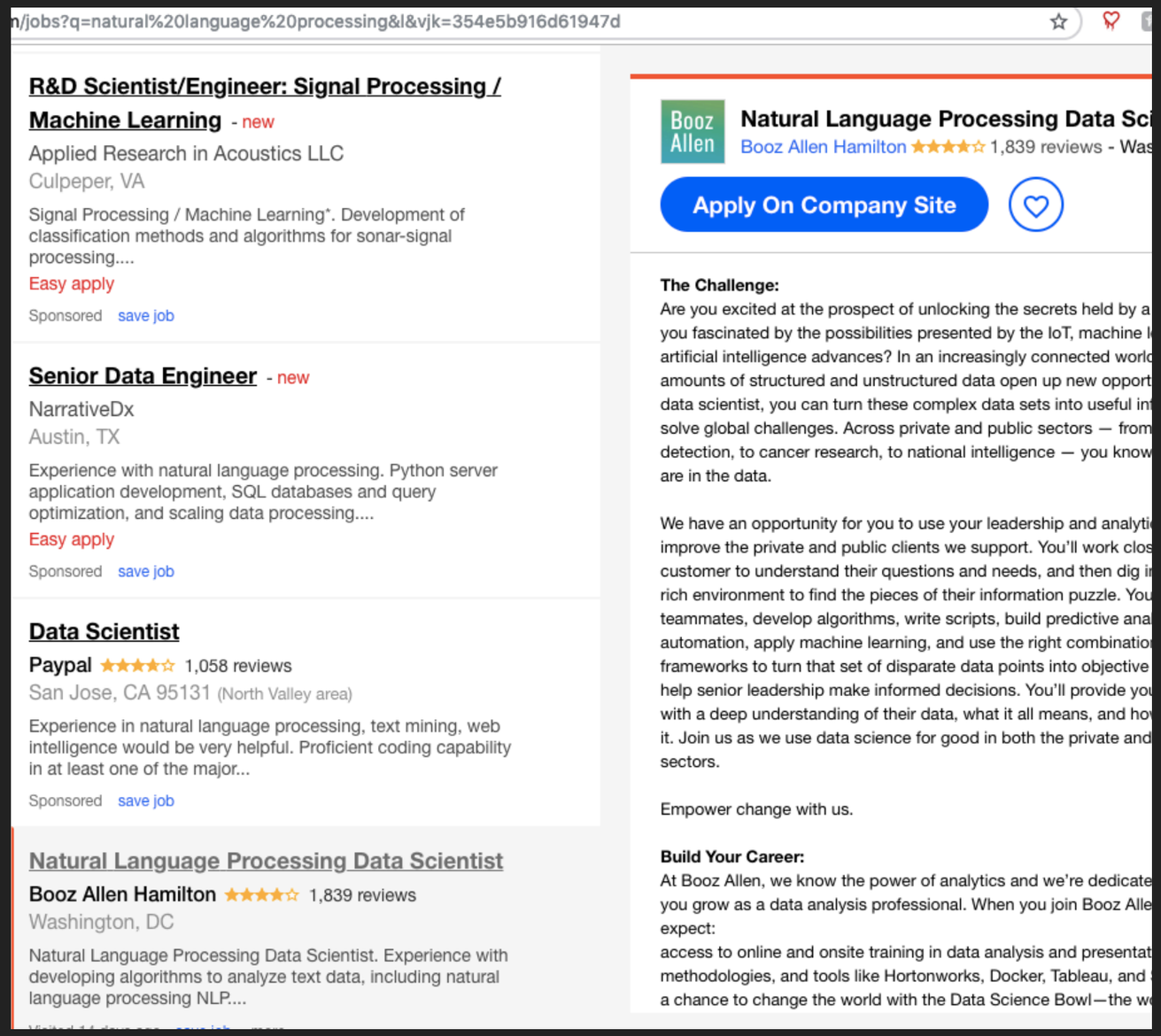
```
1
2  # coding: utf-8
3
4  # In[1]:
5
6
7  from pony_orm_model import *
8  import csv
9
10 @db_session
11 def add_job(title, job_description):
12     d = Job()
13     d.title = title
14     commit()
15     d.job_description = job_description
16     commit()
17     d.job_class = job_class
18     commit()
19
20 populate_database()
```

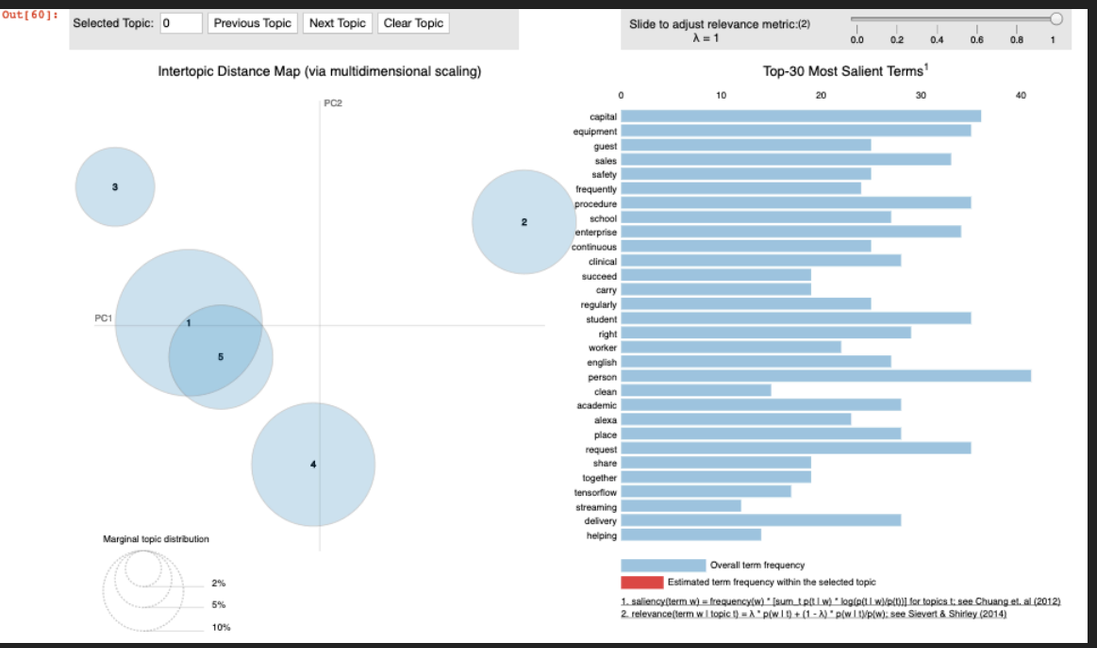
# DATABASE MANAGEMENT

# ORM AND SQLITE STORAGE

# SCRAPING AND STORAGE

- I scraped hundreds of sites like you see here
- Time intensive.
- I used an Object Relational Manager to feed the scraped data into a SQLite database, which was more reliable for larger datasets than a JSON file.



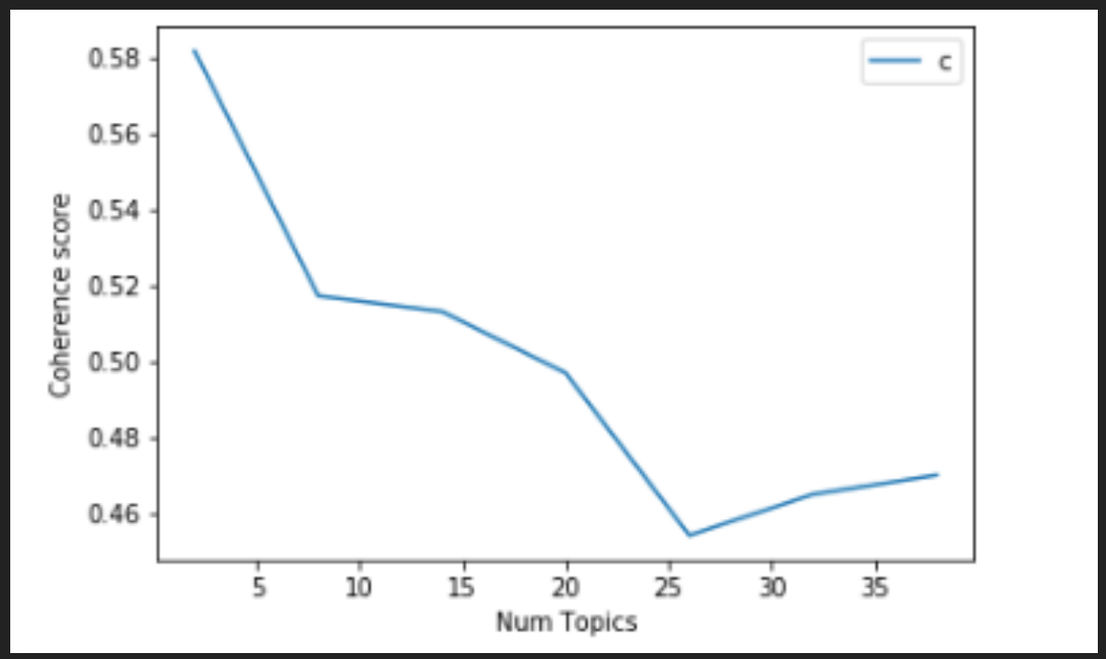



# GENSIM AND PYLDAVIZ

# UNSUPERVISED APPROACH

# EVALUATING COHERENCE

- After evaluating the coherence of the LDA, it would be unwise to go above about 10 topics since there is a plateau and drop-off at that point.



# UNSUPERVISED APPROACH–TECHNIQUES USED

- Gensim
  - Open source Python (& Cython) product for unsupervised topic modeling and NLP.
- Latent Dirichlet Allocation
  - Topic Modeling that projects the data into a space, and produces the most salient terms.
- PyLDAviz
  - Visualization of the above.

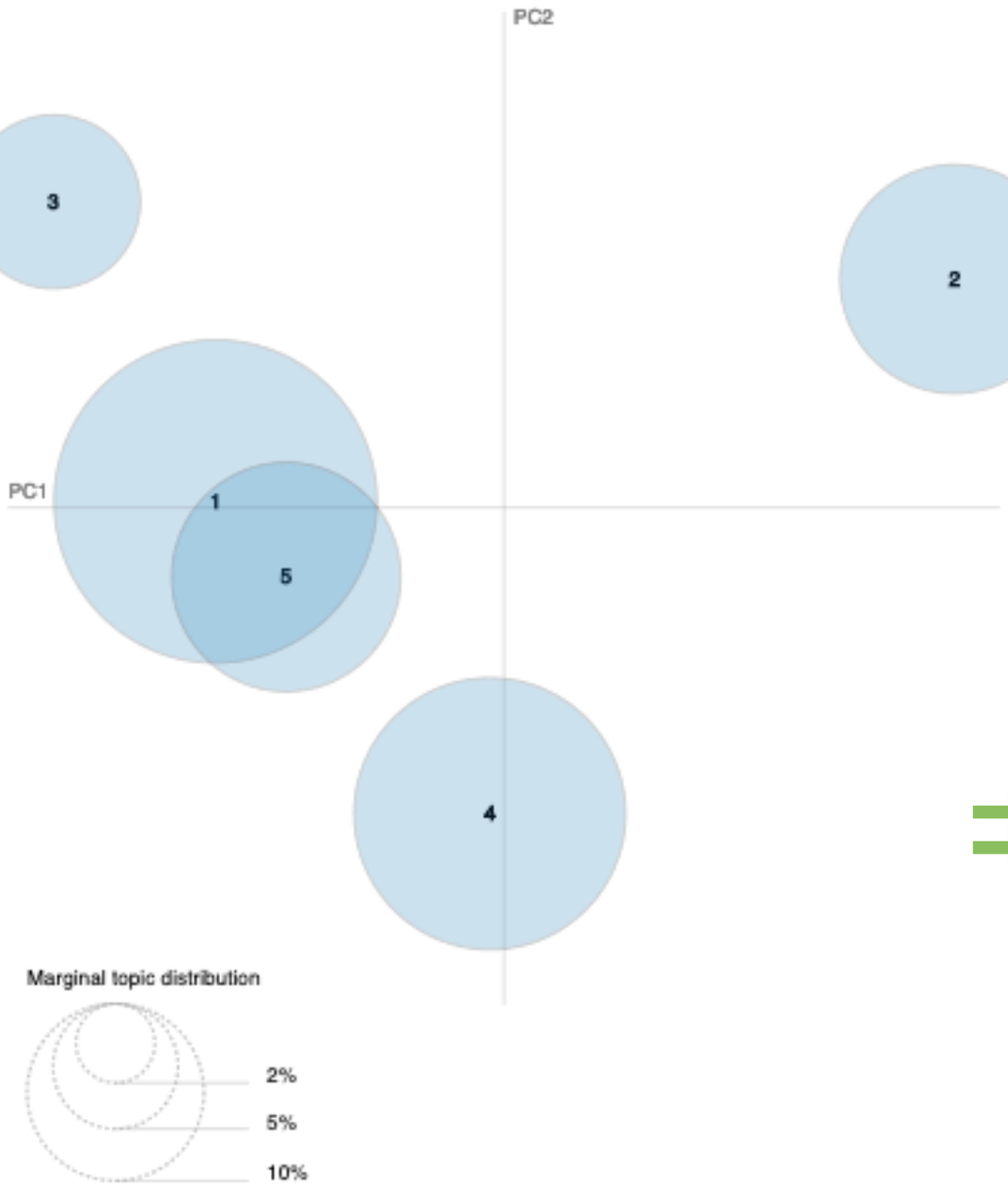
Out[60]:

Selected Topic: 0 | Previous Topic | Next Topic | Clear Topic

Slide to adjust relevance metric:(2)
λ = 1

0.0 0.2 0.4 0.6 0.8 1

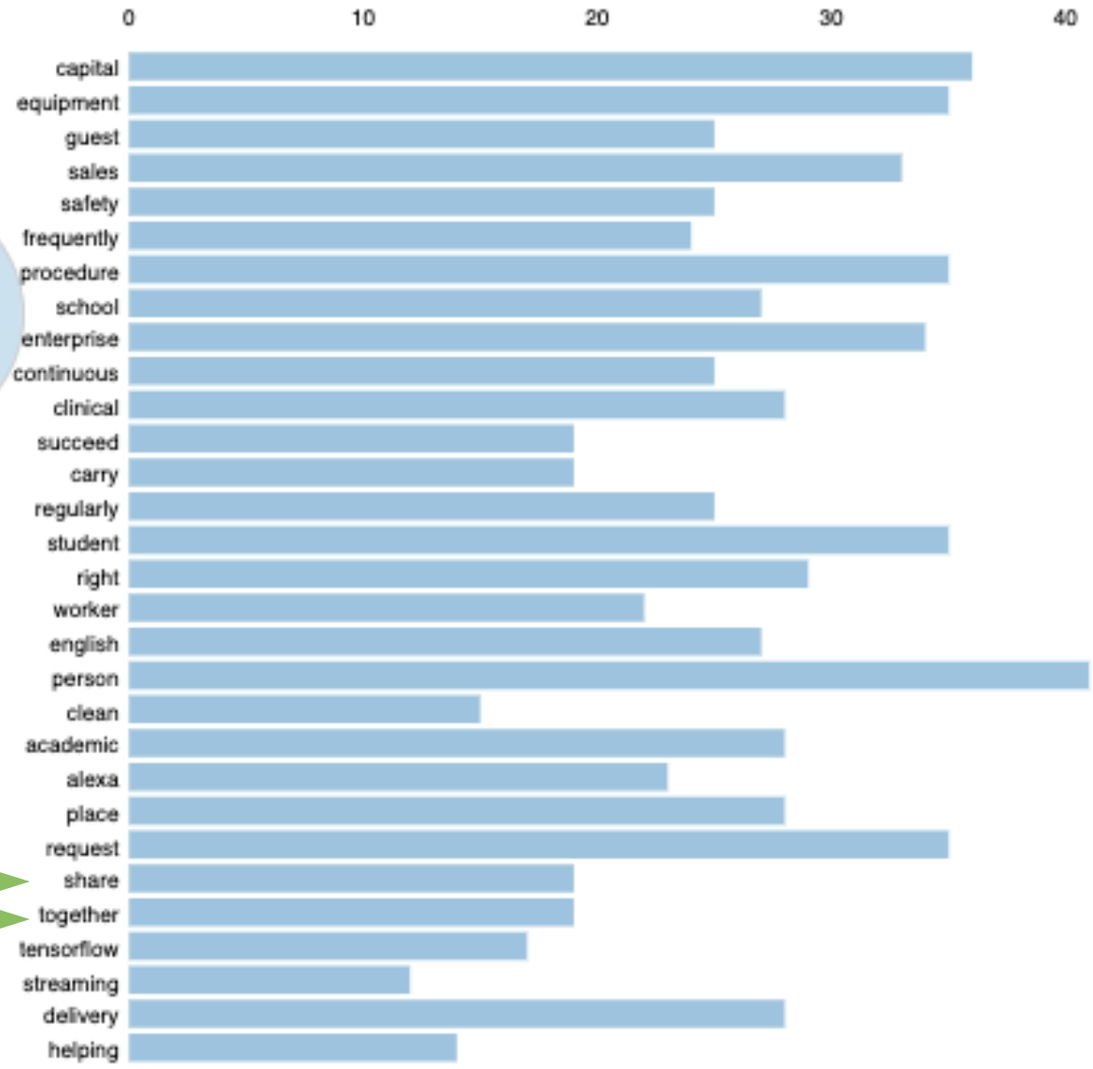
## Intertopic Distance Map (via multidimensional scaling)




Marginal topic distribution




## Top-30 Most Salient Terms[1]




Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Cluster: 1

| job_class | job_description | MiniBatchLabels |
|---|---|---|
| Apache Spark | 19 | 19 |
| Machine Learning | 183 | 183 |
| Natural Language Processing | 86 | 86 |
| Neural Networks | 166 | 166 |
| Pattern Recognition | 45 | 45 |
| Speech Recognition | 45 | 45 |
| Text Analytics | 8 | 8 |
| Text Mining | 4 | 4 |

Cluster: 2

| job_class | job_description | MiniBatchLabels |
|---|---|---|
| Apache Spark | 157 | 157 |
| Machine Learning | 82 | 82 |
| Natural Language Processing | 103 | 103 |
| Neural Networks | 188 | 188 |
| Pattern Recognition | 73 | 73 |
| Speech Recognition | 8 | 8 |
| Text Analytics | 75 | 75 |
| Text Mining | 118 | 118 |

Cluster: 3

| job_class | job_description | MiniBatchLabels |
|---|---|---|
| Apache Spark | 3 | 3 |
| Machine Learning | 48 | 48 |
| Natural Language Processing | 57 | 57 |
| Neural Networks | 56 | 56 |
| Pattern Recognition | 24 | 24 |
| Speech Recognition | 27 | 27 |
| Text Analytics | 2 | 2 |
| Text Mining | 7 | 7 |

Cluster: 4

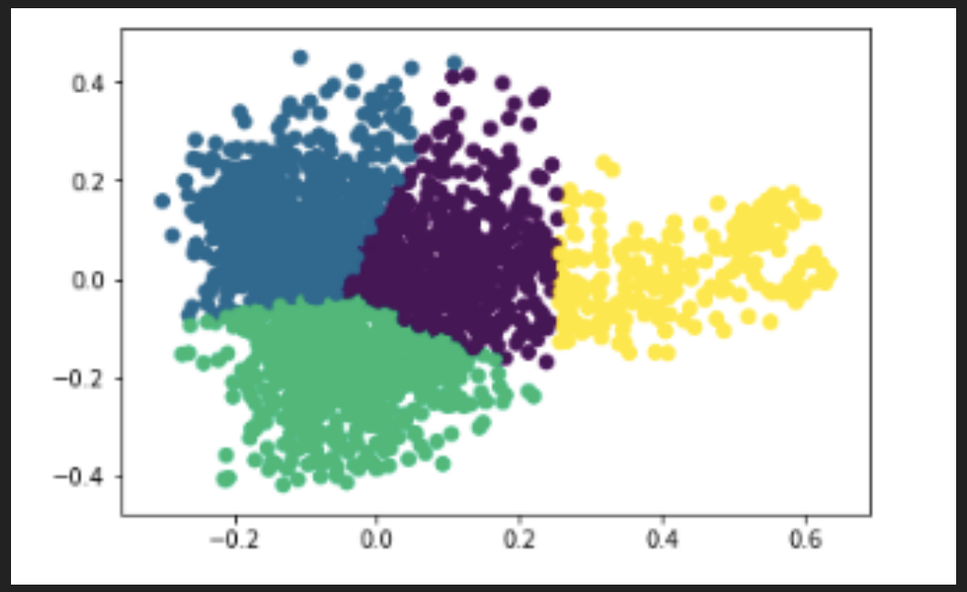
| job_class | job_description | MiniBatchLabels |
|---|---|---|
| Apache Spark | 358 | 358 |
| Machine Learning | 14 | 14 |

# TF-IDF AND BOW

# SUPERVISED APPROACH

# BAG OF WORDS

- Models used:
  - K-Means: Unfortunately, did not perform well.
  - LSA with Bow: Much more helpful.
  - LSA with Bigrams: Even better
    - Best model with 83% Cross-Val score.



# TF-IDF

- Models used:
  - K-Means Mini-batch: Helpful in generating understanding “behind the scenes” batches 1 & 2 appear very balanced.
- Gradient Boosting Classifier performed best here, and was able to match the job to its type (1 of the 8) with a score of 87.

# K-MEANS MINI-BATCH

- Speech Recognition was heavily clustered in two of the batches.
- Cluster 1 has many male-dominated fields clustered within it. The female fields are not highly represented.
- Cluster two is relatively balanced.
- Cluster 5 is an example of one of the more Speech Recognition skewed clusters.

Cluster: 1

| job_class | job_description | MiniBatchLabels |
|---|---|---|
| Apache Spark | 19 | 19 |
| Machine Learning | 183 | 183 |
| Natural Language Processing | 86 | 86 |
| Neural Networks | 166 | 166 |
| Pattern Recognition | 45 | 45 |
| Speech Recognition | 45 | 45 |
| Text Analytics | 8 | 8 |
| Text Mining | 4 | 4 |

Cluster: 2

| job_class | job_description | MiniBatchLabels |
|---|---|---|
| Apache Spark | 157 | 157 |
| Machine Learning | 82 | 82 |
| Natural Language Processing | 103 | 103 |
| Neural Networks | 188 | 188 |
| Pattern Recognition | 73 | 73 |
| Speech Recognition | 8 | 8 |
| Text Analytics | 75 | 75 |
| Text Mining | 118 | 118 |

Cluster: 3

Cluster: 5

| job_class | job_description | MiniBatchLabels |
|---|---|---|
| Apache Spark | 19 | 19 |
| Machine Learning | 65 | 65 |
| Natural Language Processing | 68 | 68 |
| Neural Networks | 75 | 75 |
| Pattern Recognition | 96 | 96 |
| Speech Recognition | 126 | 126 |
| Text Analytics | 41 | 41 |
| Text Mining | 31 | 31 |

**The distance between the Deep Learning Engineer (66% male) and the NLP Position (mostly women).**

In [82]:
```
scipy.spatial.distance.pdist([one_topics, two_topics])
```

Out[82]:
```
array([0.25583754])
```

**The distance between the Deep Learning and Pattern Recogition (98% male).**

In [83]:
```
scipy.spatial.distance.pdist([one_topics, three_topics])
```

Out[83]:
```
array([0.41192855])
```

**The distance between NLP and Pattern Recognition.**

In [84]:
```
scipy.spatial.distance.pdist([two_topics, three_topics])
```

Out[84]:
```
array([0.3271281])
```

**This is very interesting. What we can see here is that a job description from a relatively balanced field has the greatest distance from the very-male dominated field. While the NLP and Pattern Recognition aren't as far apart as the Deep Learning and Pattern Recognition, I think it is still of note. What we see here is that a job description from a balanced field is different from one that is not balanced at all. This makes a great case that even job descriptions that might be skewed toward one gender or another are more similar than we realize as well.**

**In the end, I think this makes a great case for gender-balanced job descriptions that can attract the best and brightest from any gender (I'd, of course, like to broaden this to people who identify as gender non-binary).**

**I think something like this projection could help create a program that analyzes a job description, and gives feedback to the client about how balanced it might be. It would take web development, and extensive model training, but I think something like that could be valuable.**

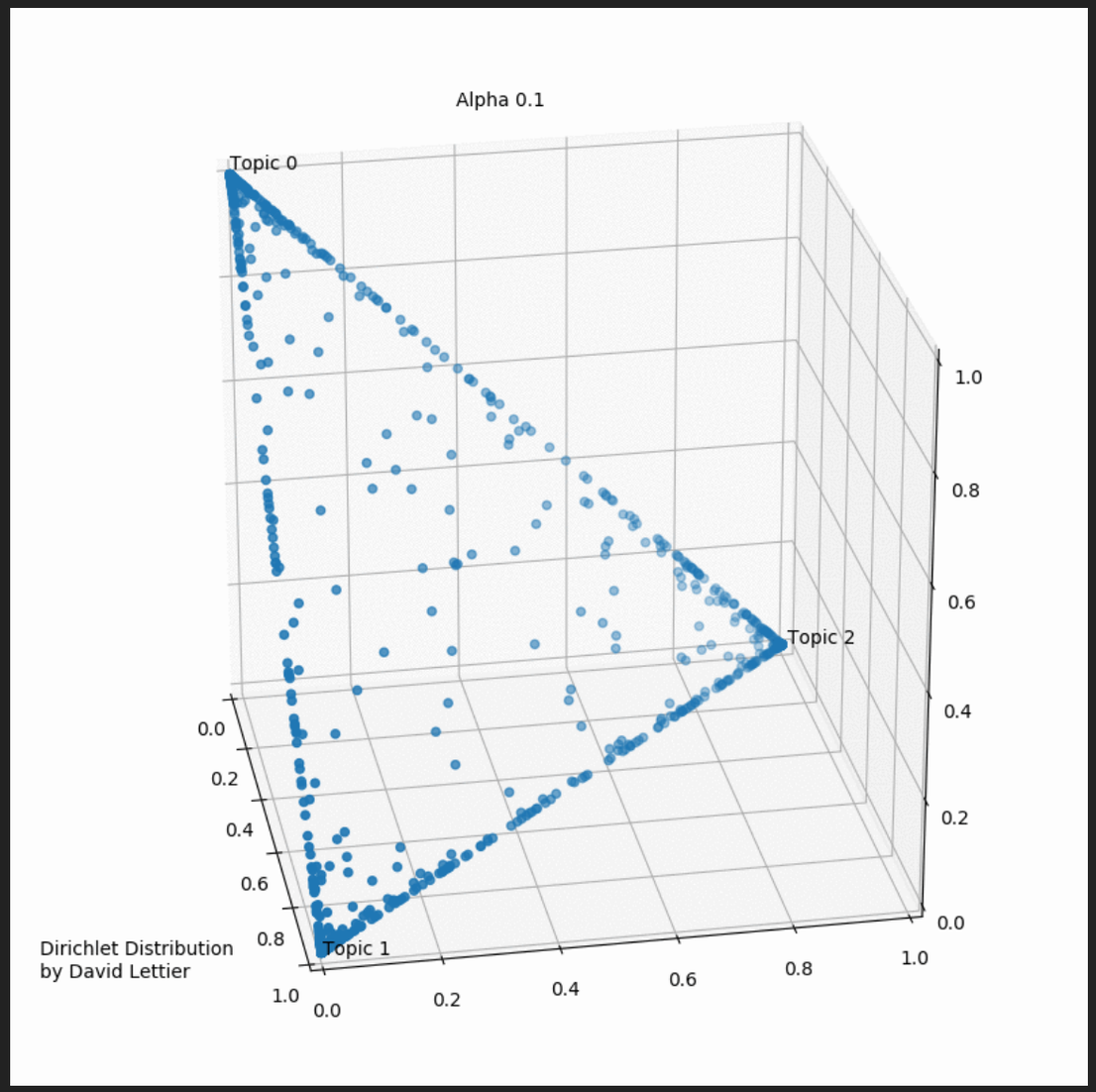
**This projection model was created with extensive help from my mentor, Philip Robinson. So I'd like to give him credit here where it's due!**

# LDA PROJECTION

# FURTHER UNSUPERVISED APPROACH

# LDA PROJECTION

- With this, we project three new job descriptions into the LDA space, and measure the distance between the new postings, based on the understanding of the Latent Dirichlet Allocation.
- Each topic (word in our case) has an associated alpha which is plotted in a shape like this.




https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d

**The distance between the Deep Learning Engineer (66% male) and the NLP Position (mostly women).**

```
scipy.spatial.distance.pdist([one_topics, two_topics])
```

```
array([0.25583754])
```

**The distance between the Deep Learning and Pattern Recogition (98% male).**

```
scipy.spatial.distance.pdist([one_topics, three_topics])
```

```
array([0.41192855])
```

**The distance between NLP and Pattern Recognition.**

```
scipy.spatial.distance.pdist([two_topics, three_topics])
```

```
array([0.3271281])
```

This shows that the difference between the balanced and female descriptions and the difference between the Deep Learning is much larger than the difference between the NLP and Pattern Recognition.

**This is very interesting. What we can see here is that a job description from a relatively balanced field has the greatest distance from the very-male dominated field. While the NLP and Pattern Recognition aren't as far ap the Deep Learning and Pattern Recognition, I think it is still of note. What we see here is that a job descriptio a balanced field is different from one that is not balanced at all. This makes a great case that even job descrip that might be skewed toward one gender or another are more similar than we realize as well.**

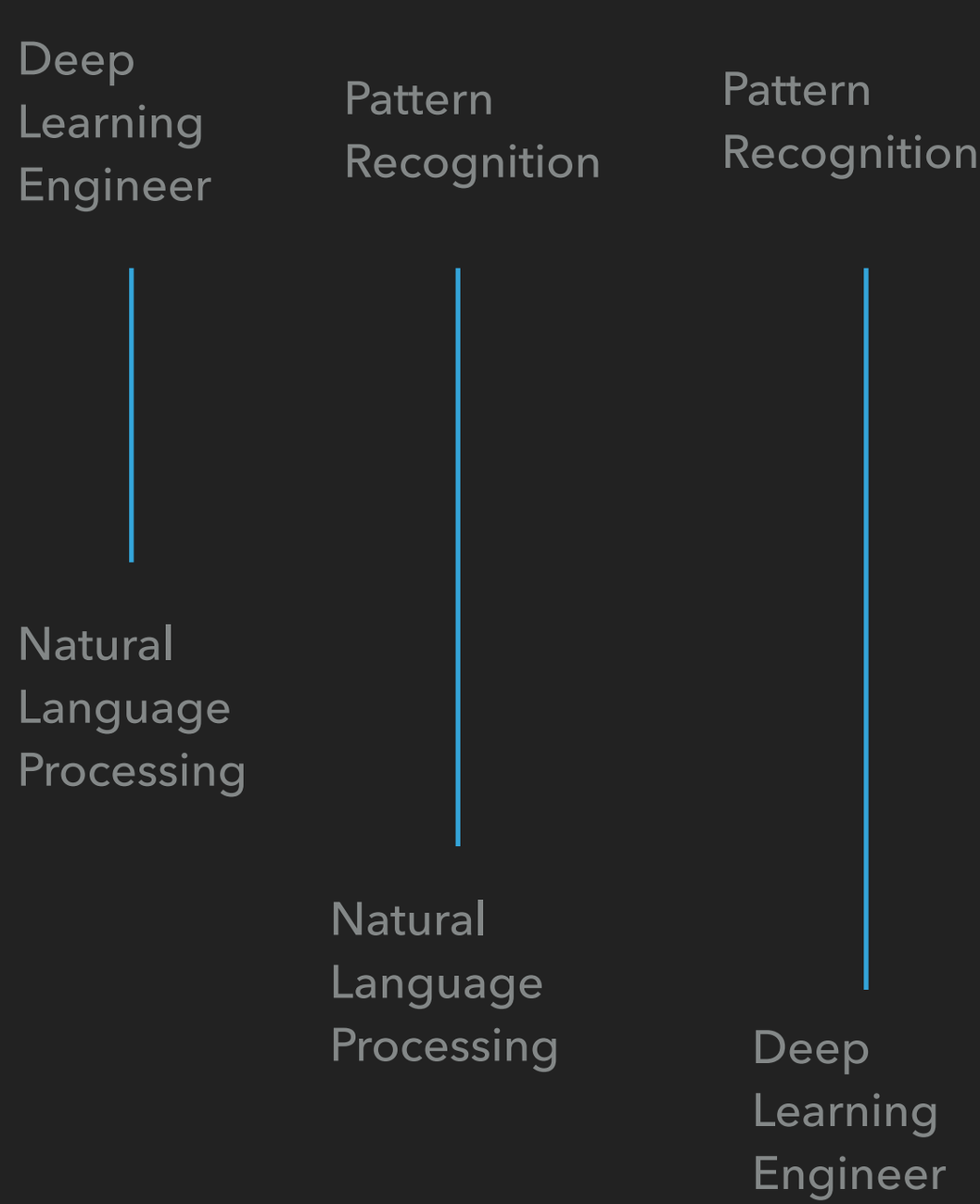
**In the end, I think this makes a great case for gender-balanced job descriptions that can attract the best and brightest from any gender (I'd, of course, like to broaden this to people who identify as gender non-binary).**

**I think something like this projection could help create a program that analyzes a job description, and gives feedback to the client about how balanced it might be. It would take web development, and extensive model training, but I think something like that could be valuable.**

**This projection model was created with extensive help from my mentor, Philip Robinson. So I'd like to give hi credit here where it's due!**

# ANALYSIS OF LDA PROJECTION

- Three new job descriptions were projected into the LDA Space:
  - NLP (Female-dominated)
  - Deep Learning Engineer (Balanced)
  - Pattern Recognition (Male-Dominated)
- As it turns out, the least similar descriptions are the balanced and male-dominated description, whereas the most similar descriptions are the balanced, and female-dominated fields.
- This suggests that the female, and balanced fields might have the most similar postings, and the male-dominated fields would be the most different of the group.



# OUTCOMES AND FURTHER RESEARCH

- In short, the project did not produce some of the definitive results I was looking for.  However, I still think it had some valuable outcomes
  - LDA Projection
  - Modeling and classification
  - PyLDAviz
- A larger corpus could help promote understanding, so to improve the project, I would increase the corpus size and try some of the same approaches.

# OUTCOMES AND FURTHER RESEARCH

- Something like this would be the ideal outcome from this project. However, I think just creating awareness with the project helps us not to skew a posting either way, but perhaps promote an equitable work environment that brings together all of the best talent available.

@tshaefrench

@datatf

# THANKS FOR ATTENDING!

# QUESTIONS?