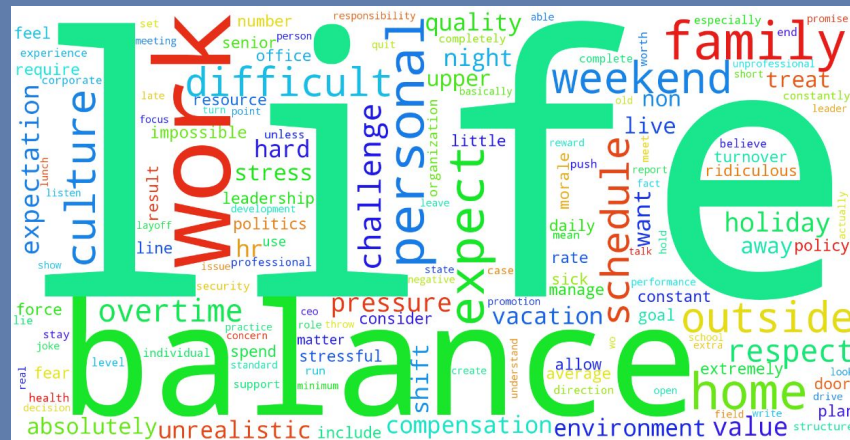
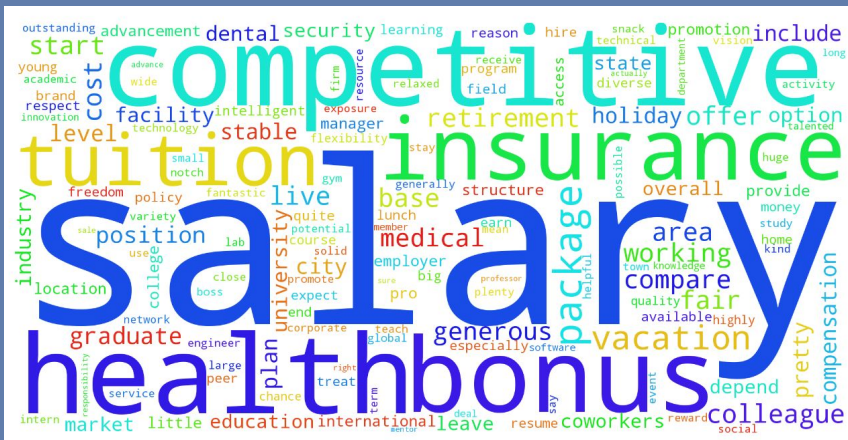


Quantifying Employee Satisfaction

Data Science Immersive Capstone Project



Author: Dave Hogue

Motivation



- Forbes “Best Places To Work” articles
- Personal experience as an employee
- Application of Natural Language Processing

Finding Source Data

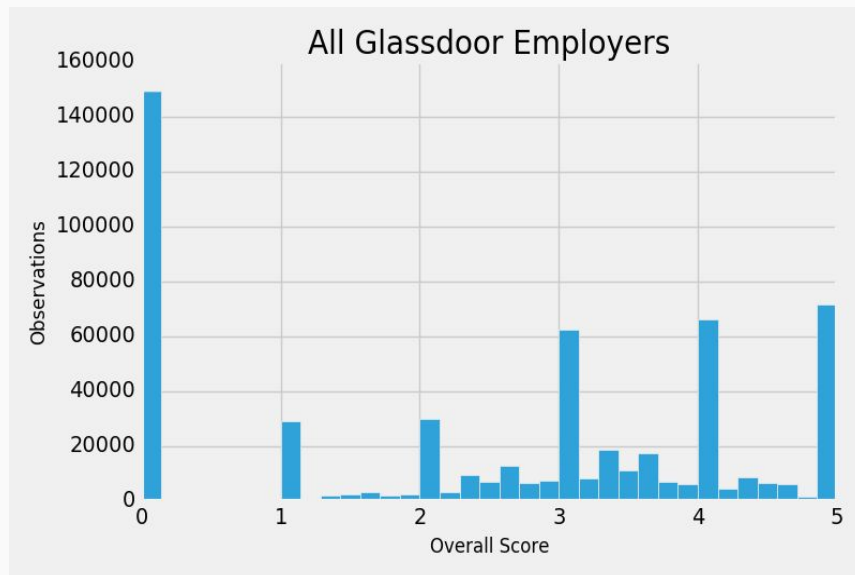


Employer Selection



Bird's Eye View:

- Over 560,000 total employers
- Most have 0 or 1 review

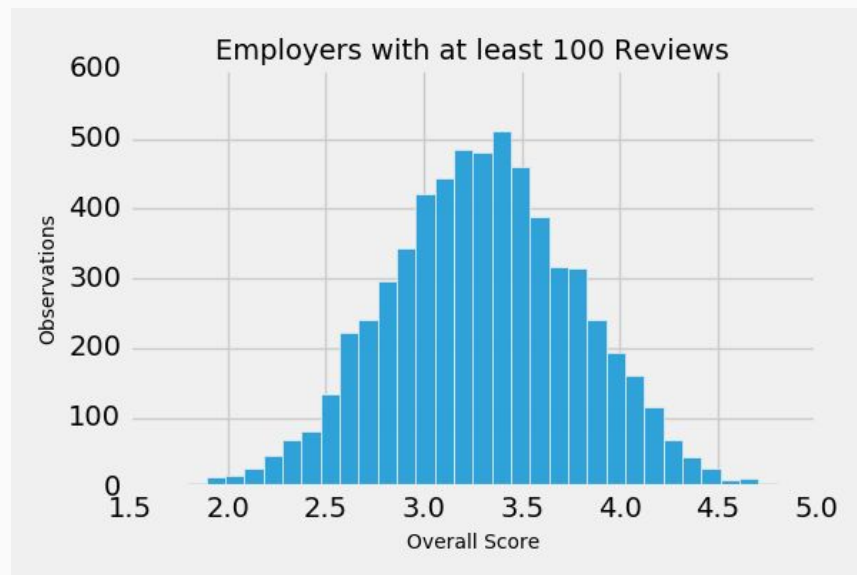


Employer Selection



Controlled for Volatility:

- 6,205 Employers
- Gaussian Distribution

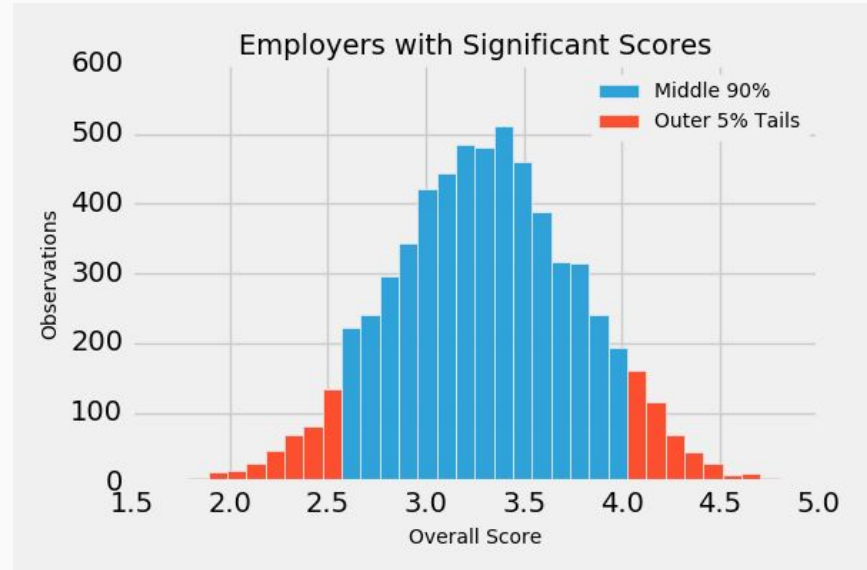


Employer Selection



Selected for Analysis:

- Above 95th percentile
- Below 5th percentile
- 847 total employers



Data Collection



- Time consuming, tedious
- 25,000+ pages of reviews
- Manual separation of workload
- Multithreading for performance gains



Pluralsight Reviews | Glassdoor

Pluralsight

102 Reviews

4.2

Recommend to a friend: 94%

Approve of CEO: 90%

Reviews by Job Title

102 Employee Reviews

Sort: Popular Rating Date

Reviews by Job Title

- Sales (8)
- Software Craftman (4)
- Account Manager (3)
- Sales Repres (1)

Work in HR or Recruiting?

Shopify Reviews | Glassdoor

Shopify

104 Reviews

4.5

Recommend to a friend: 94%

Approve of CEO: 90%

Reviews by Job Title

104 Employee Reviews

Sort: Popular Rating Date

Reviews by Job Title

- Sales (8)
- Software Craftman (4)
- Account Manager (3)
- Sales Repres (1)

Work in HR or Recruiting?

DayNine Consulting Reviews | Glassdoor

DayNine Consulting

100 Reviews

4.4

Recommend to a friend: 94%

Approve of CEO: 90%

Reviews by Job Title

100 Employee Reviews

Sort: Popular Rating Date

Reviews by Job Title

- Consultant (11)
- Integration Consultant (8)
- Senior Consultant (5)
- Principal Consultant (5)
- Project Manager (4)

Company Updates

Lettuce Entertain You Reviews | Glassdoor

Lettuce Entertain You

103 Reviews

4.0

Recommend to a friend: 94%

Approve of CEO: 90%

Reviews by Job Title

103 Employee Reviews

Sort: Popular Rating Date

Reviews by Job Title

- Server (25)
- Host (8)
- Manager (5)
- Hostess (2)

Work in HR or Recruiting?

University of Melbourne Reviews | Glassdoor

University of Melbourne

104 Reviews

4.0

Recommend to a friend: 94%

Approve of CEO: 90%

Reviews by Job Title

104 Employee Reviews

Sort: Popular Rating Date

Reviews by Job Title

- Research Assistant (7)
- Tutor (5)
- Research Fellow (4)
- Casual Tutor (1)

Work in HR or Recruiting?

SUNY Albany Reviews | Glassdoor

SUNY Albany

99 Reviews

4.2

Recommend to a friend: 94%

Approve of CEO: 90%

Reviews by Job Title

99 Employee Reviews

Sort: Popular Rating Date

Reviews by Job Title

- Graduate Research Assistant (8)
- Research Assistant (4)
- Lecturer (3)
- Adjunct Inst (1)

Work in HR or Recruiting?



Text Preprocessing Considerations

- **Stop words** - you, your, me, I, the, a, an, etc.
- **Part of speech tagging** - using context to filter text
- **Lemmatization** - am, are, is, was → be



TF-IDF Matrix

Term Frequency - Inverse Document Frequency

- Each row represents one document, each column represents one term
- TF score - count of instances in one document
- IDF score - penalize for terms in many documents

$$w_{x,y} = \text{tf}_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

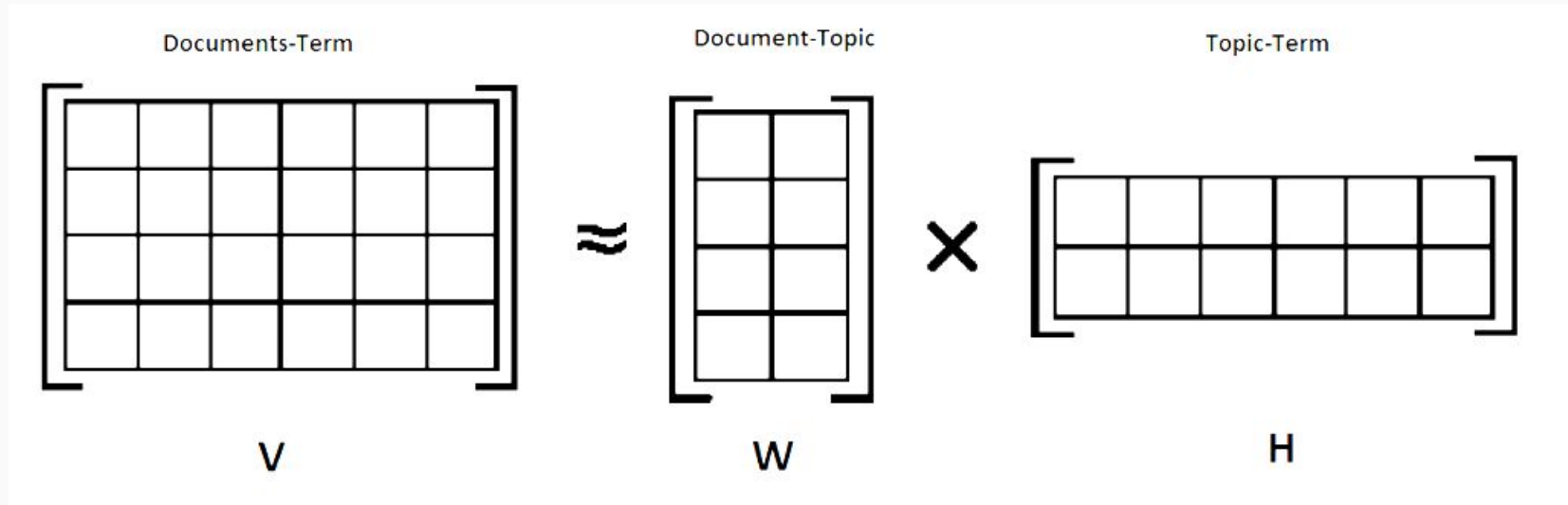


NMF Clustering

Non-Negative Matrix Factorization

- Matrix (W) maps documents to latent topics
 - Each row represents a document
 - Each column represents a topic
- Matrix (H) maps latent topics to terms
 - Each row represents a topic
 - Each column represents a term

NMF Clustering

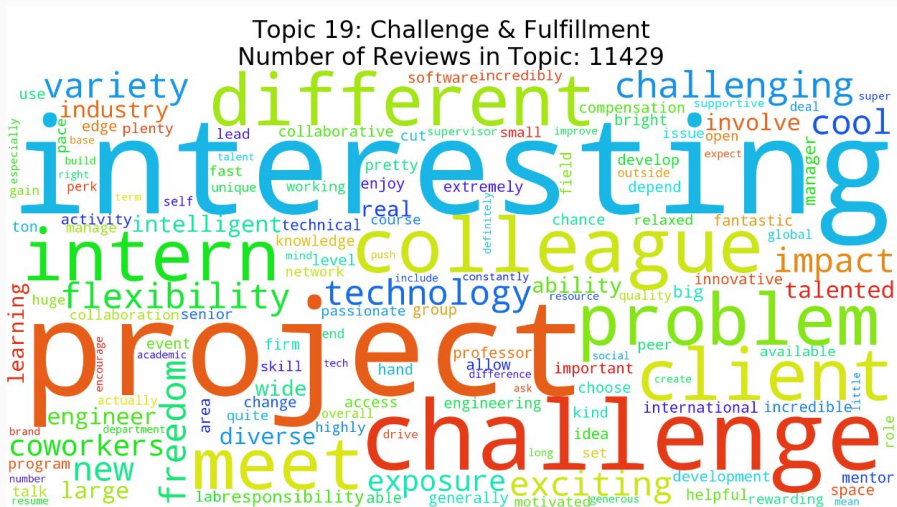


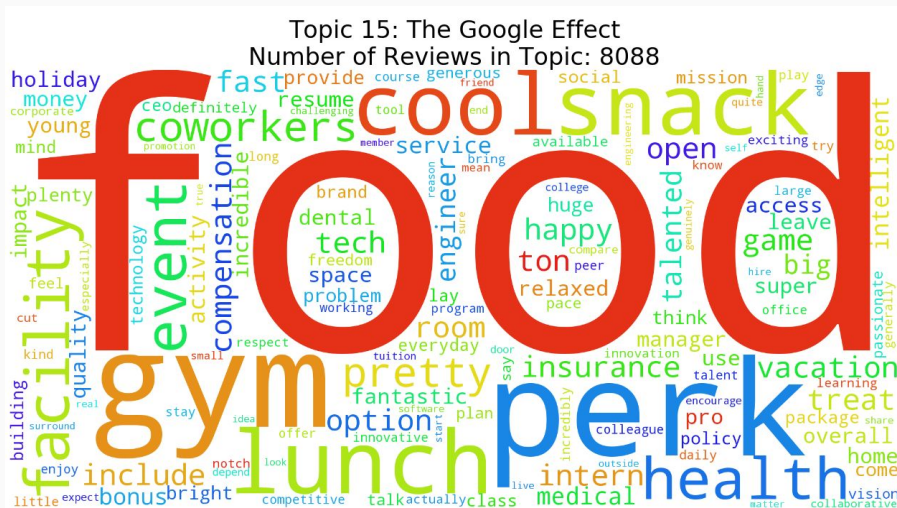


NMF Clustering

Blurry Clustering

- Allows each document to exist in multiple clusters
- Gives a score to each document for topic attribution
- Allows for discretion on a threshold for topic attribution









Challenges



Bot Detection

- Request Headers & Selenium
- Rate Limits
- CAPTCHA Challenge Images

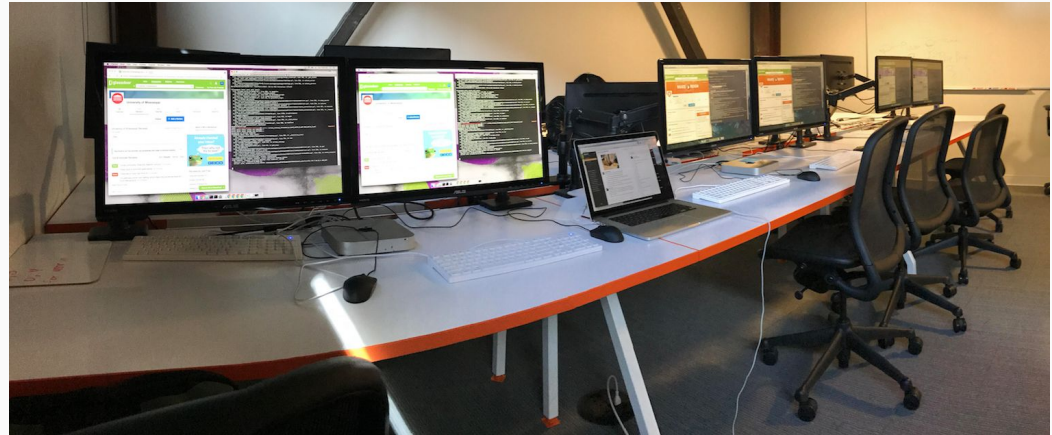


Challenges

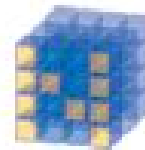
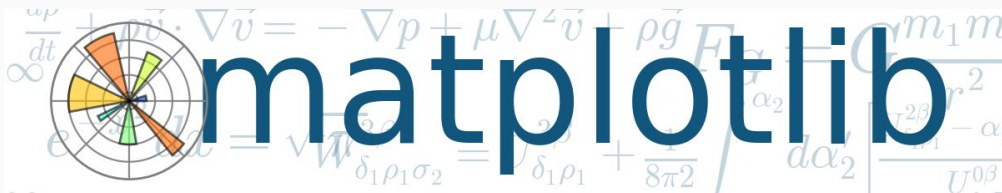


Scraping & Volume of Data

- HTTP Errors
- Multithreading in Selenium
- Multiprocessing







Technologies Used



NumPy

Contact Information



-  thedavehogue@gmail.com
-  thedavehogue
-  in/thedavehogue
-  @thedavehogue



"This is the Dave Hogue you are looking for . . ."