Kate Crawford's GA Capstone

PORTFOLIO EVALUATION ENGINE

Project Goal:

Measure up portfolio projects to available jobs in the market.

Data Collection

Data Collection: Attempt to Scrape Data

Major win for our members in court ruling against personal data scraping

Sarah Wight November 4, 2022

Share Share

Today, lead attorney Sarah Wight shared an update on LinkedIn about the hiQ court ruling that LinkedIn may enforce its User Agreement against data scraping and fake accounts. This is a significant step in helping us keep our platform and members safe, so we're sharing it with you in full text here.

> *Today in the hiQ legal proceeding, the Court announced a significant win for LinkedIn and our members against personal data scraping, among other platform abuses. The Court ruled that LinkedIn's User Agreement unambiguously prohibits scraping and the unauthorized use of scraped data as well as fake accounts, affirming LinkedIn's legal positions against hiQ for the past six years. The Court also found that hiQ knew for years that its actions violated our User Agreement, and that LinkedIn is entitled to move forward with its claim that hiQ violated the Computer Fraud and Abuse Act.*
>
> *The Court's ruling helps us better protect everyone in our professional community from unauthorized use of profile data, and it establishes important precedent to stop this kind of abuse in the future. We will continue to fight on behalf of our members to stop illegal scraping. From taking legal action against unauthorized scraping to making significant investments in technical defenses, we are committed to keeping the control of data where it belongs - with our members.*

Data Wrangling

Data Wrangling: Multiple Sources

47915 rows × 11 columns

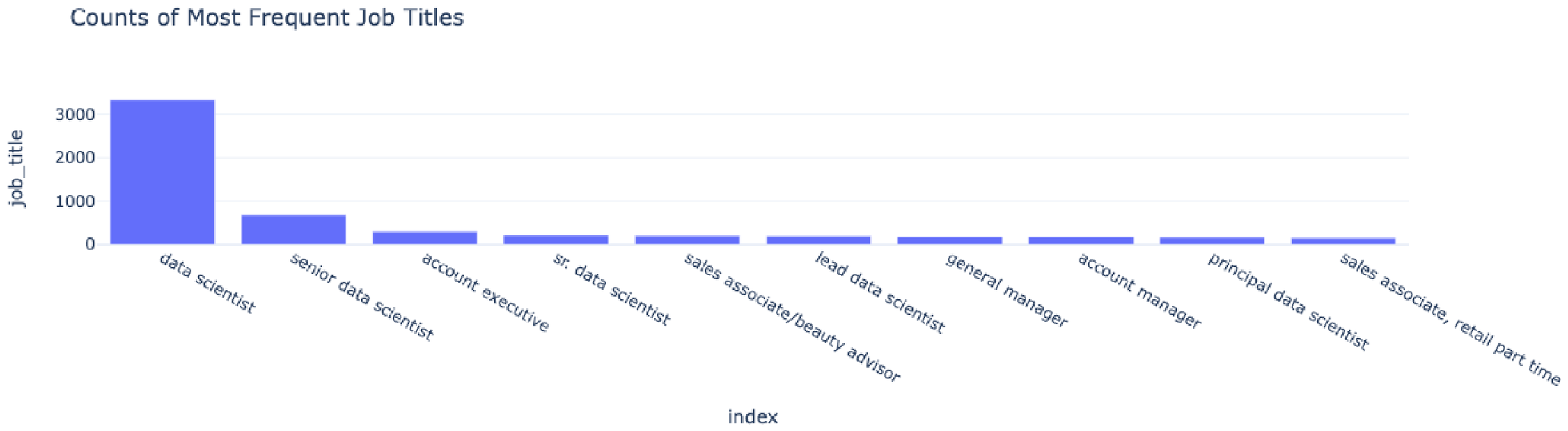
Indeed Job Posting Dataset from PromptCloud and DataStock
Number of Features: 28848

Job Posts & Online Courses Study found on Mendeley
Number of Features: 9067

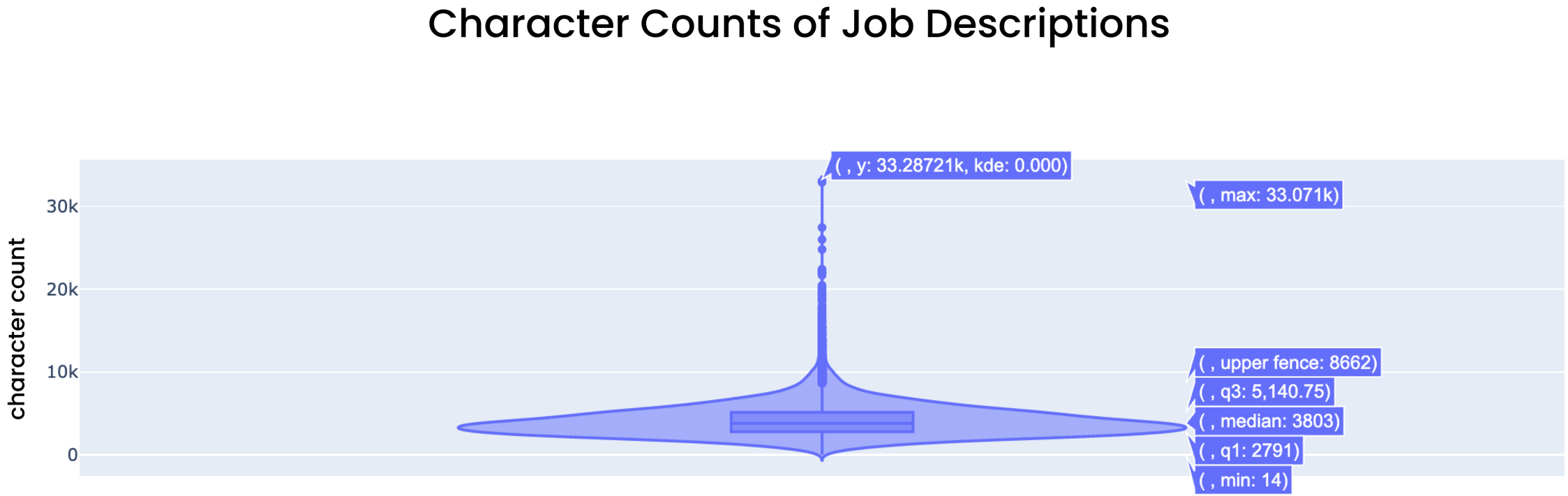
Data Scientist Job Postings found on Kaggle from JobsPikr.
Number of Features: 28848

Exploratory Data Analysis

EDA: Job Titles

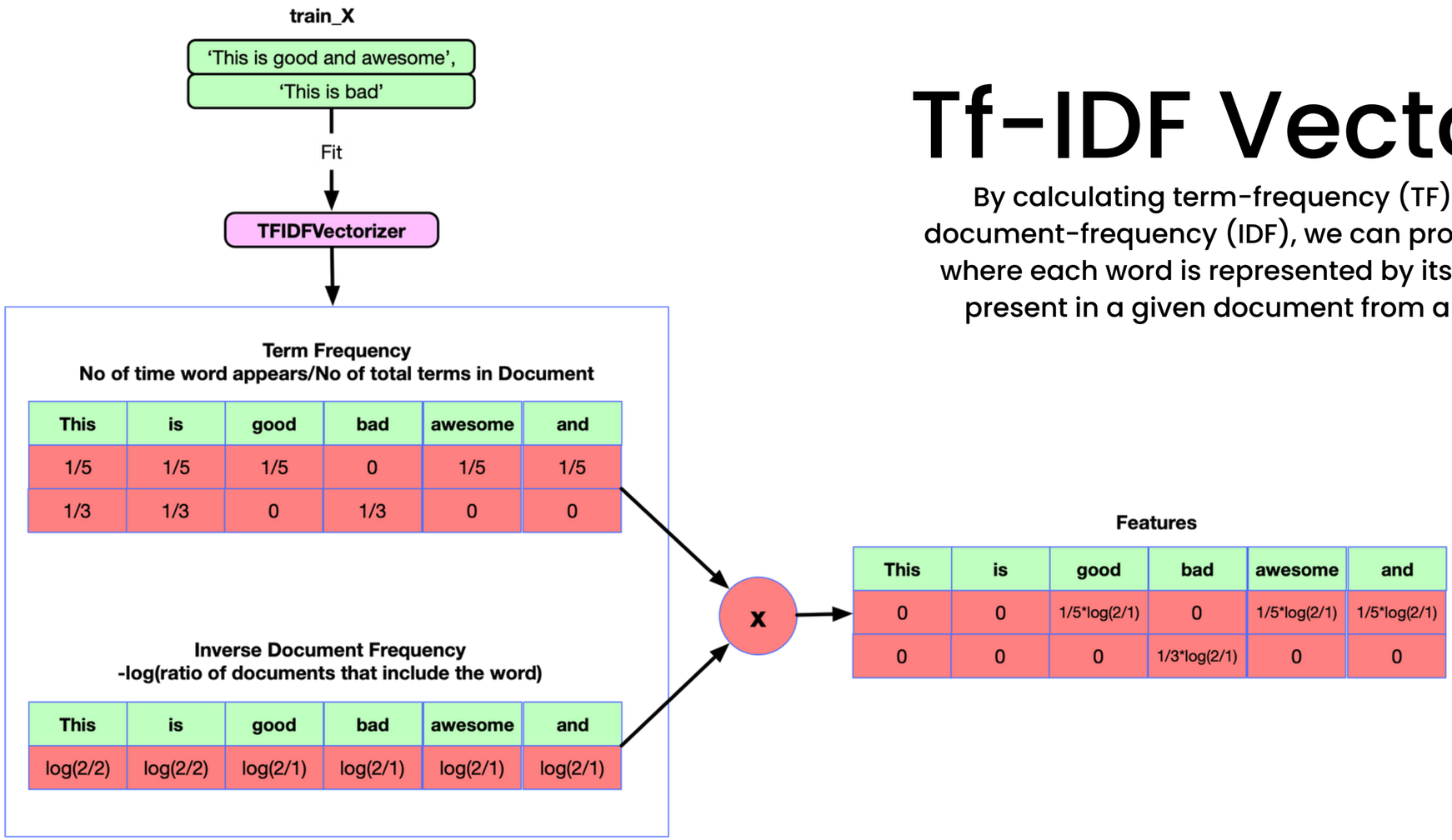


EDA: Job Posting Length



Natural Language Processing

NLP: Vectorizing



Source

Tf-IDF Vectorizing

By calculating term-frequency (TF) times the inverse document-frequency (IDF), we can produce a sparse matrix where each word is represented by its probability of being present in a given document from a set of documents.

NLP: Ngrams

Top Bigrams

data scientist	1880.396324
customer service	1601.332116
year experience	1520.429904
machine learning	1430.089486
communication skill	1214.031269

Bigrams led by Verbs

ability +
apply
build
communicate
develop

grow +
business
company
team
opportunity

perform +
duty
essential
job
management

Bigrams led by Nouns

data +

engineer
management
mining
science
scientist

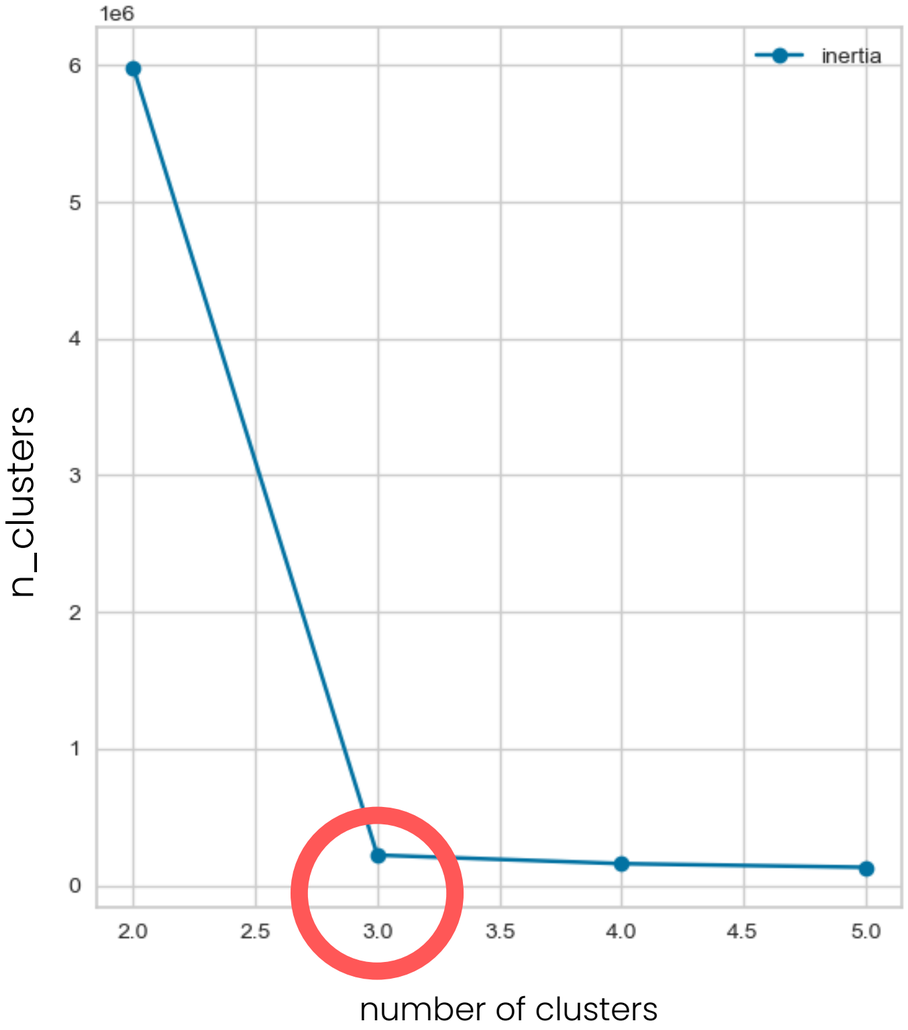
business +

analytic
case
decision
development
goal

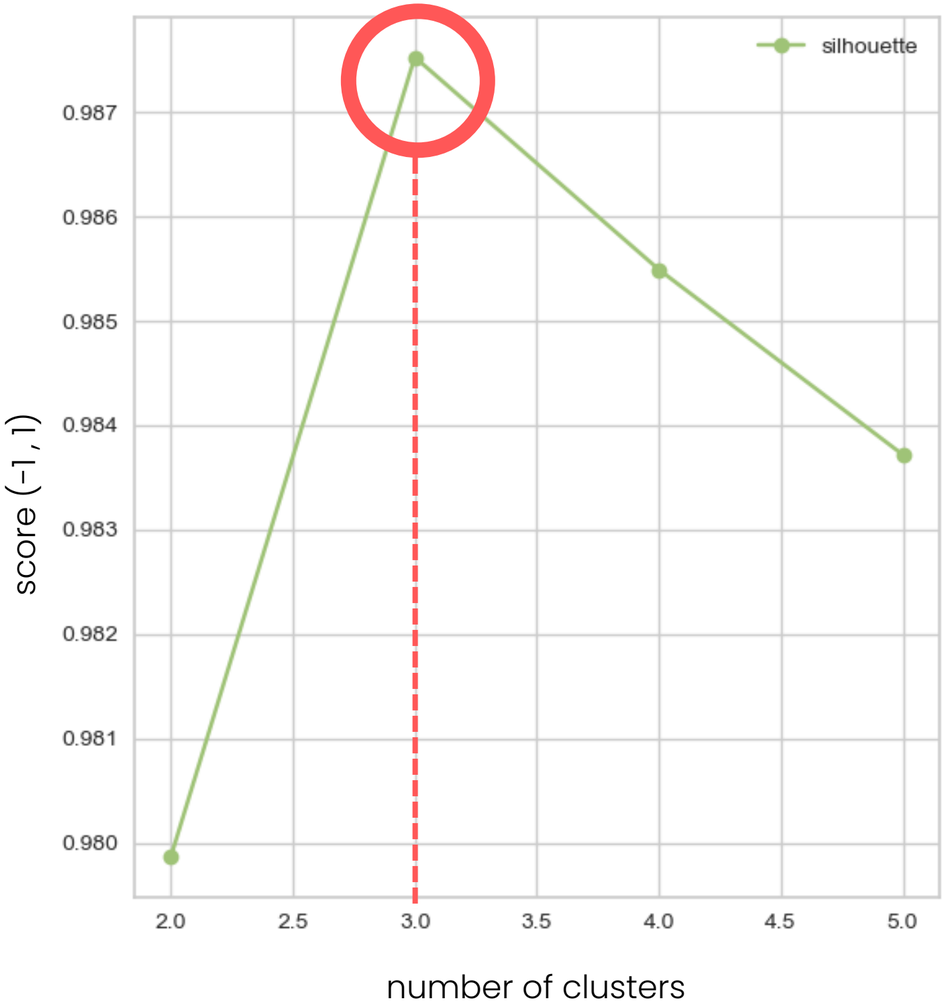
Data Mining

Cluster Model: KMeans

Elbow Method

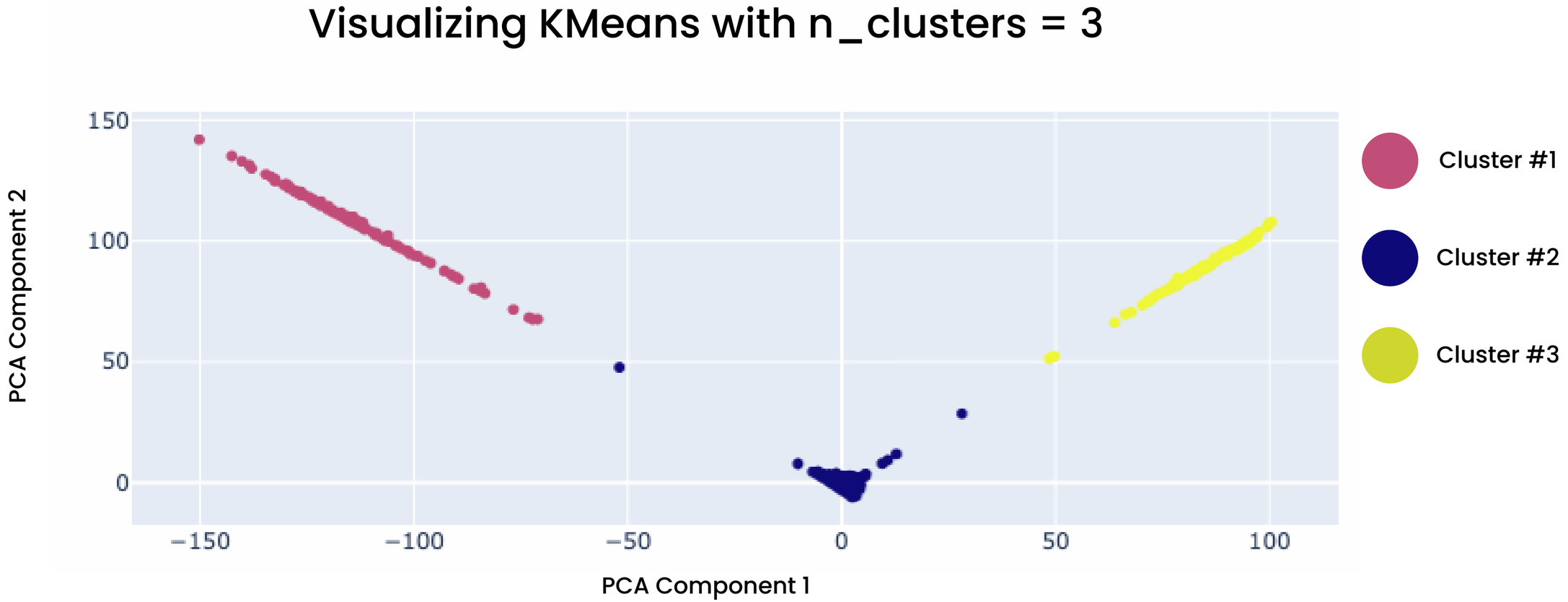


Silhouette Scoring

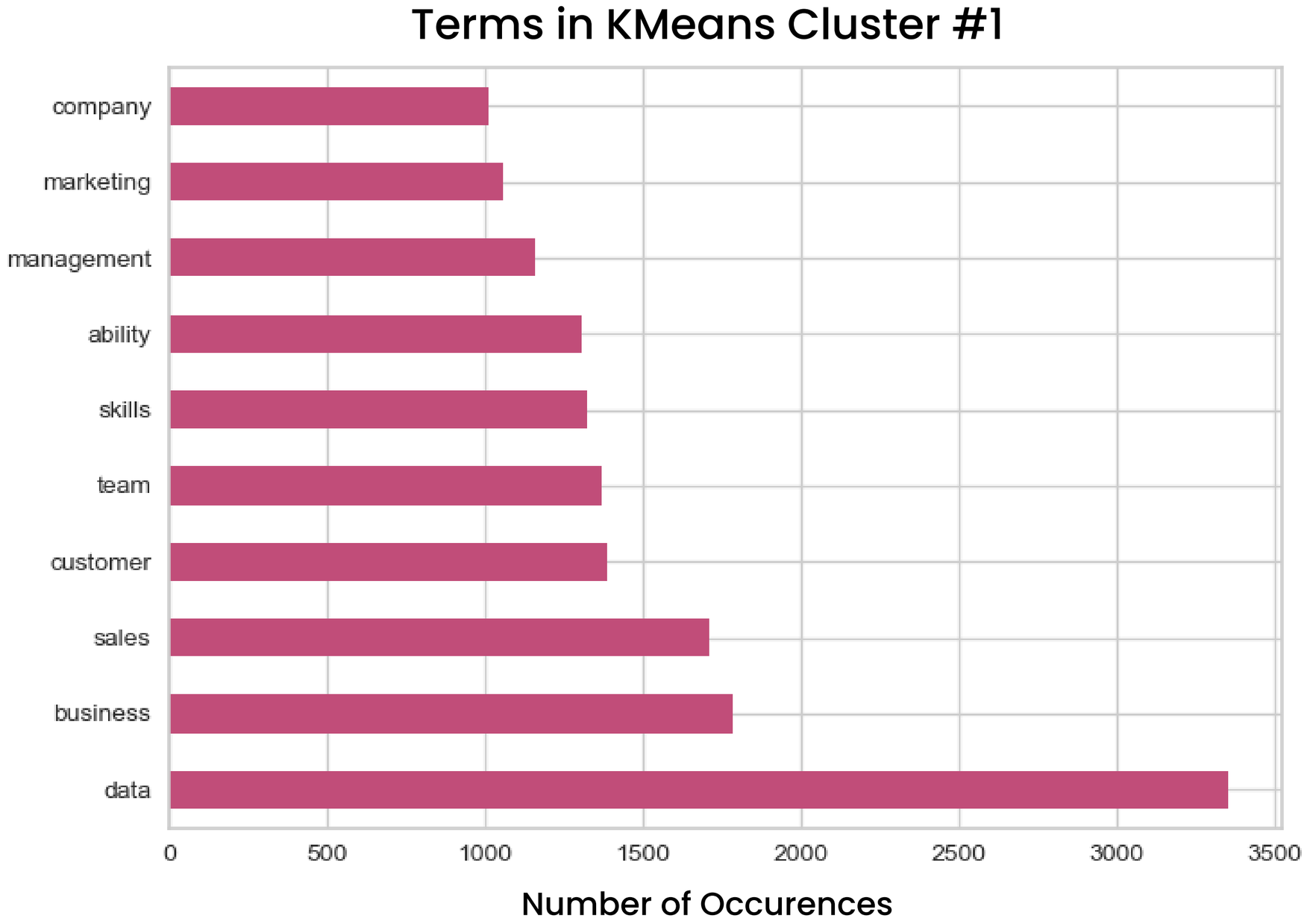


Choosing the Optimal N_Clusters

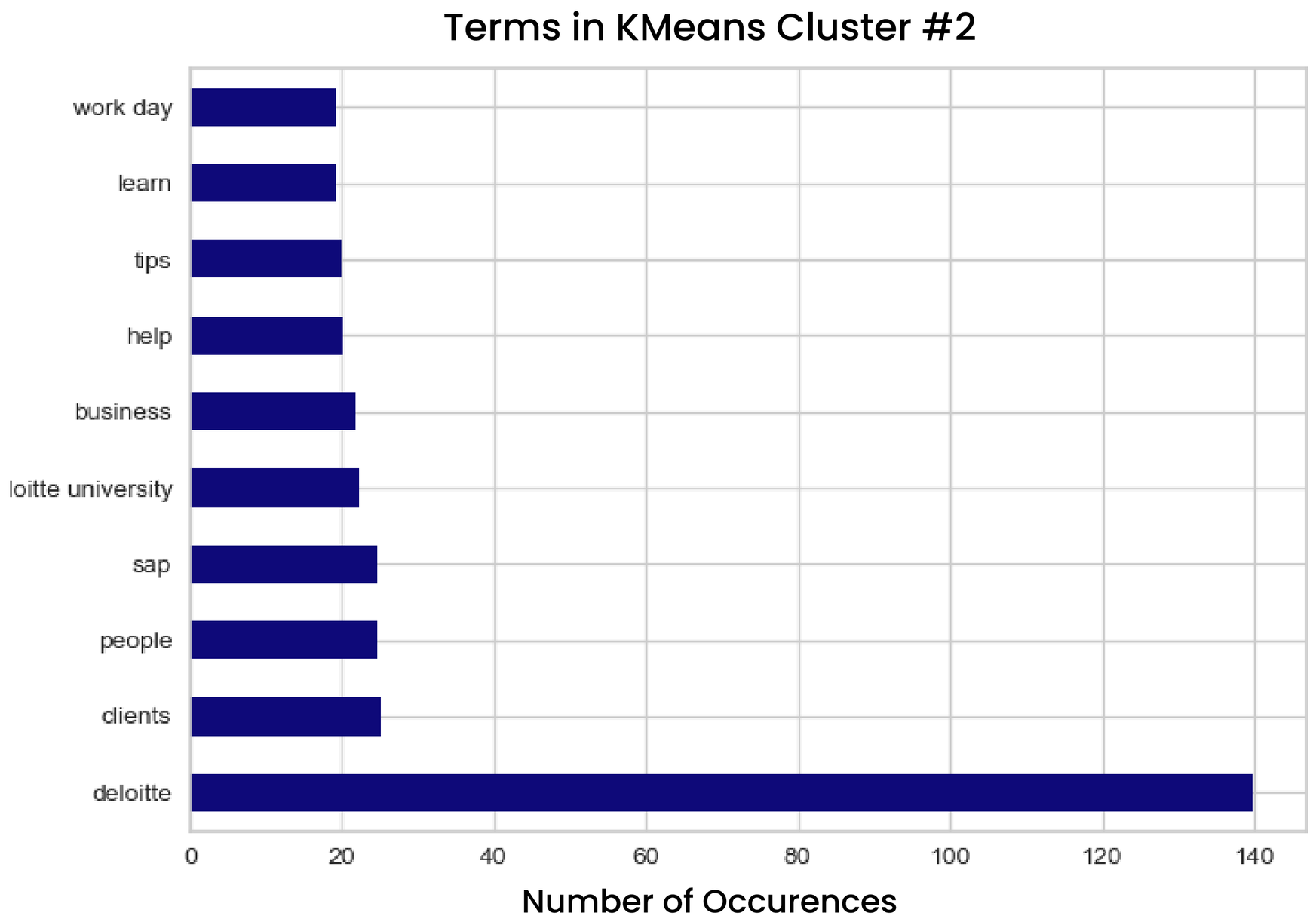
Cluster Model: KMeans



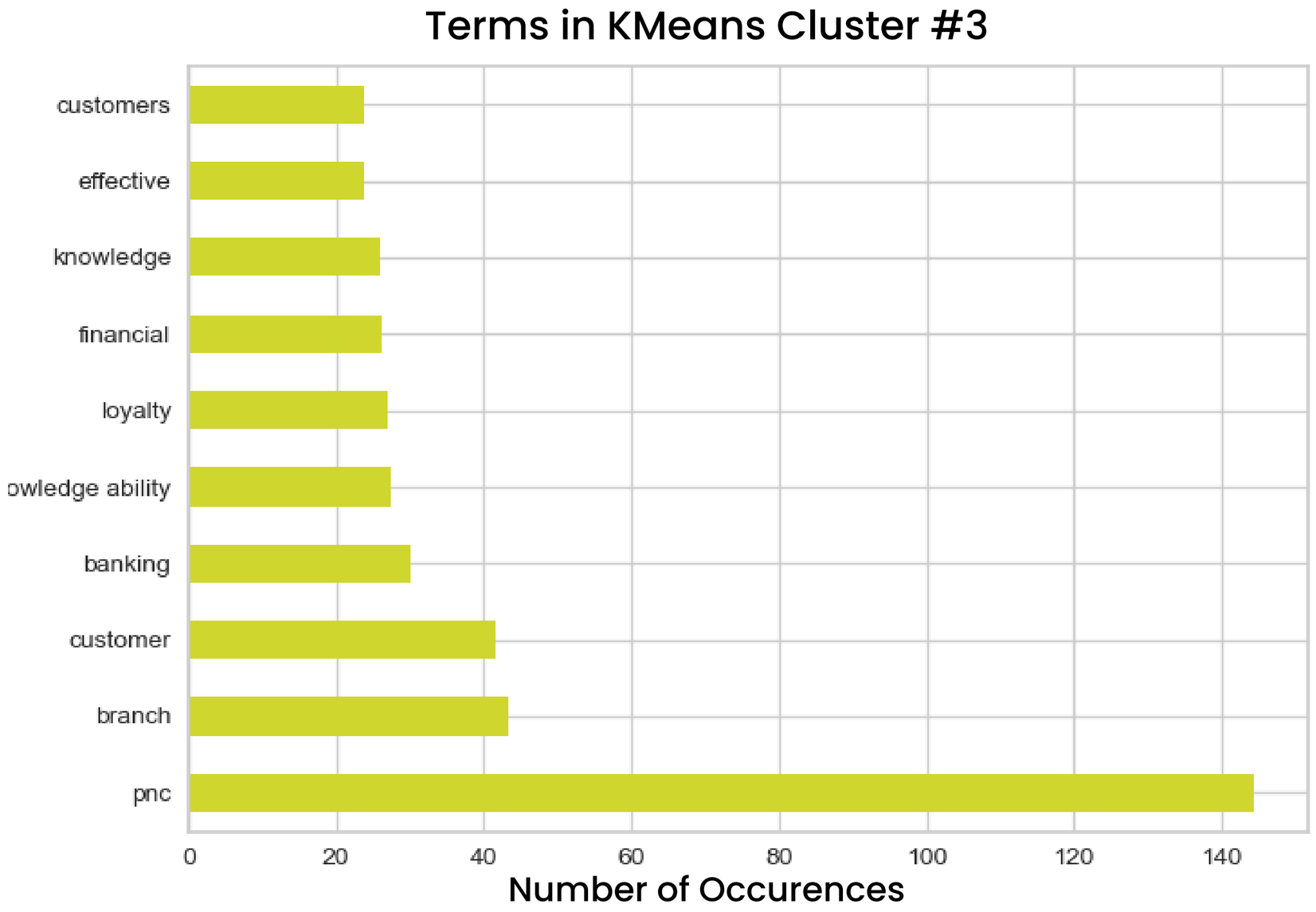
Cluster Model: KMeans



Cluster Model: KMeans



Cluster Model: KMeans



Cluster Model: KMEANS

Cluster #1:
A mix of sales and technical positions

1. Customer Service Representative
2. Senior Data Scientist
3. Software Engineer (Data Engineer)

Cluster #2:
Specializations and Managerial Roles

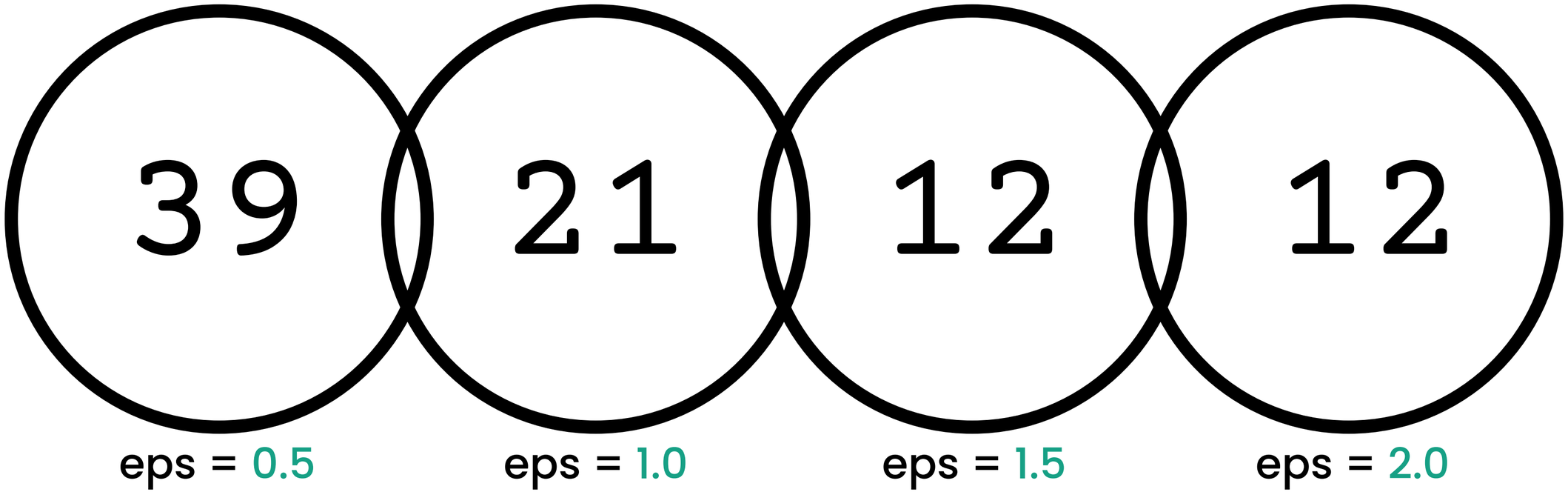
1. SAP Data Architect - Manager
2. Technical Business Analyst - Manager
3. Senior Manager - Oracle Cloud

Cluster #3:
Sales and Finance Industries

1. Branch Sales & Service Associate I
2. Private Equity Associate
3. Relationship Manager

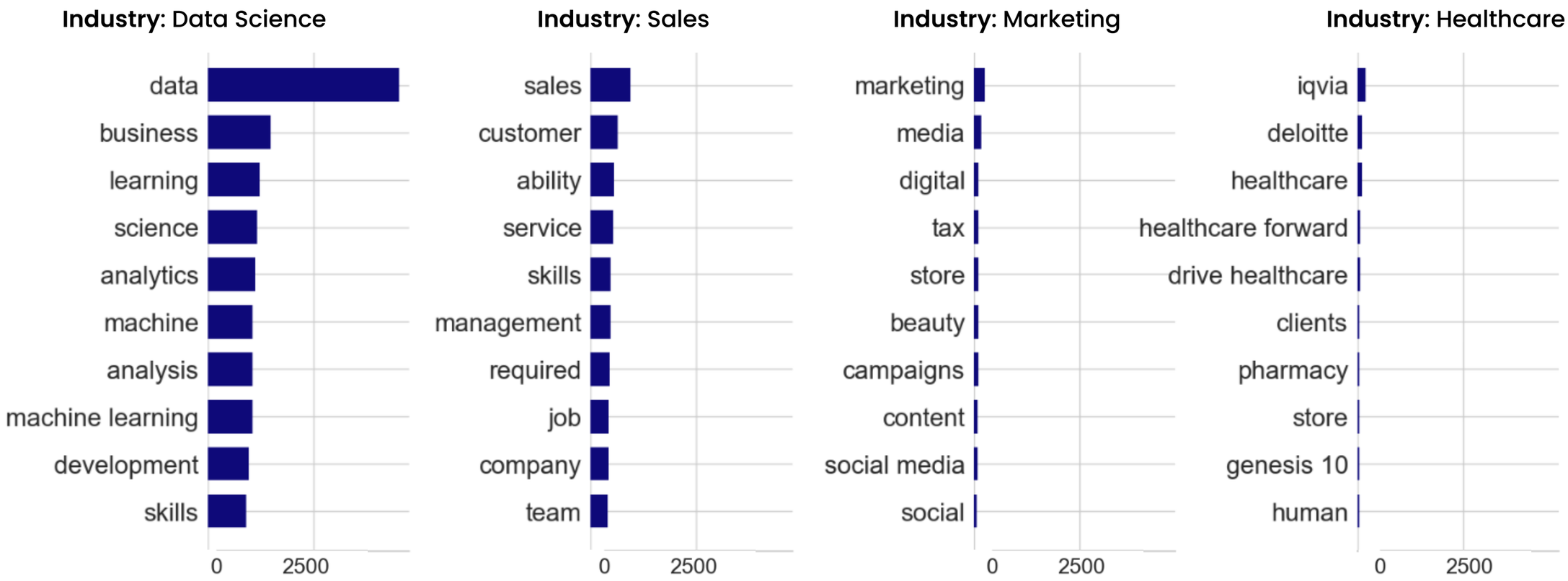
Cluster Model: DBSCAN

n_clusters per epsilon value



Topic Modeling: LDA

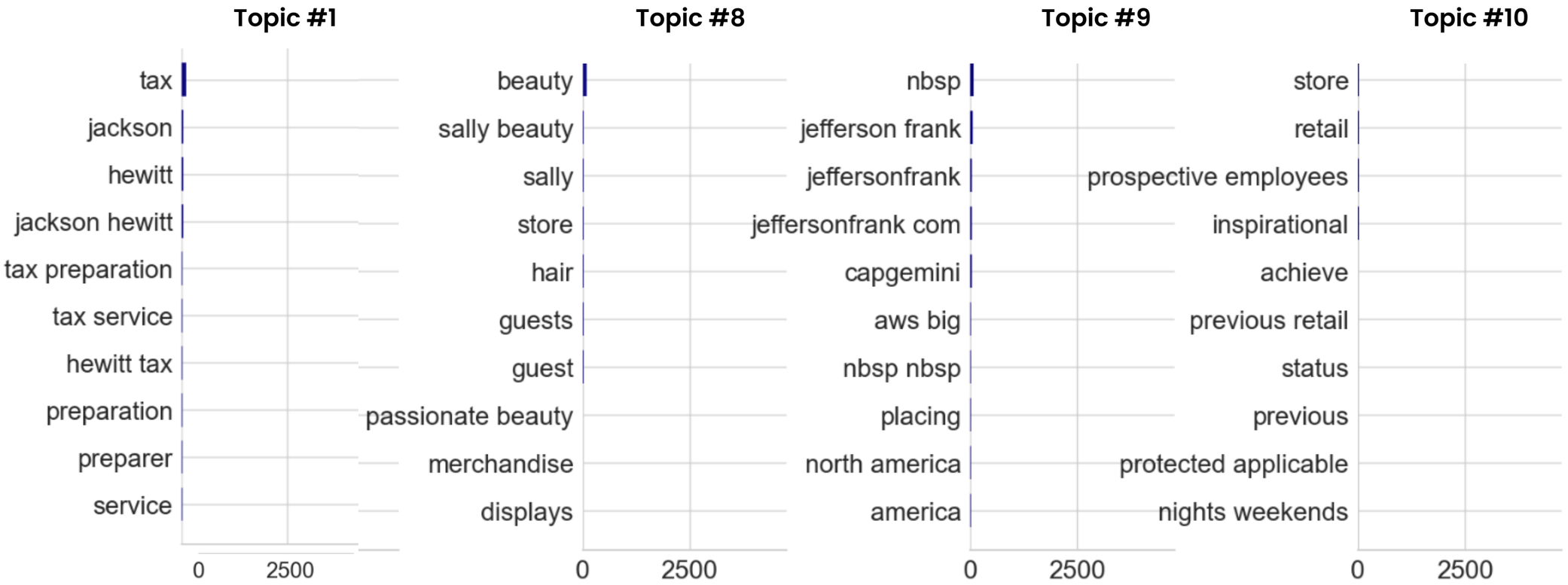
Grouping into 4 Topics



x-axis: Term Frequency

Topic Modeling: LDA

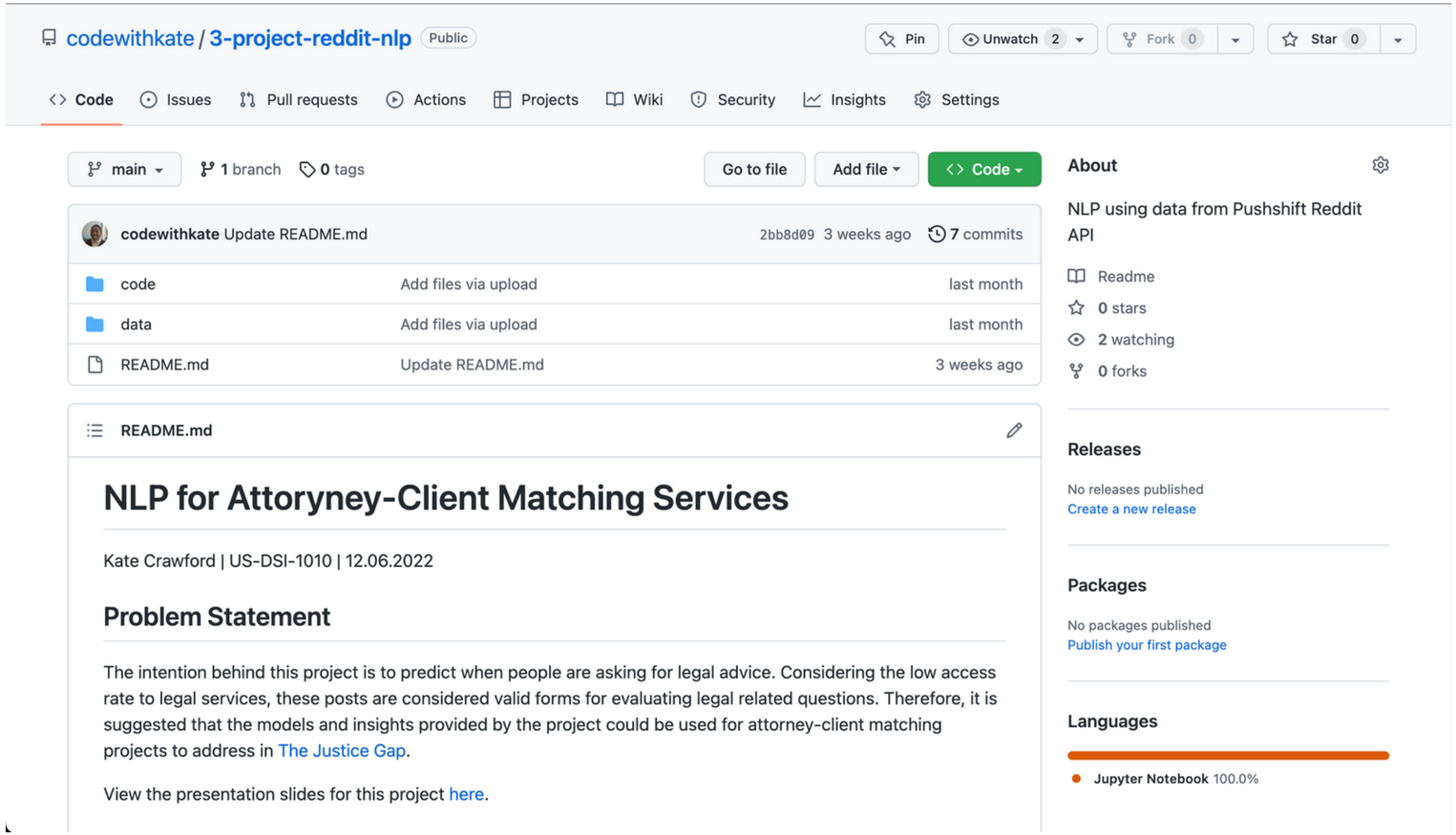
Sample from Grouping into 12 Topics



x-axis: Term Frequency

Recommender Engine

Recommender Engine Testing



Generated Recommendations

1. Legal Counsel
2. Counsel Sr.
3. Assistant General Counsel
4. Software Development Team Lead
5. Associate Corporate Counsel
6. Corporate Counsel
7. Associate Vice President, Corporate Counsel
8. Contract Technical Recruiter
9. Assistant General Counsel
10. Business Intelligence Analyst, Legal

NOTE: Observed outputs from only one README.md

app demo

Job Recommendation App

This Recommender Engine generates job recomendations and language to use if you apply to these positions. To begin, connect to GitHub:

Enter GitHub Username:

codewithkate

Get Projects

README.md files retrieved!

Your Best Matches

1 2 3 4 5 6 7 8 9 10

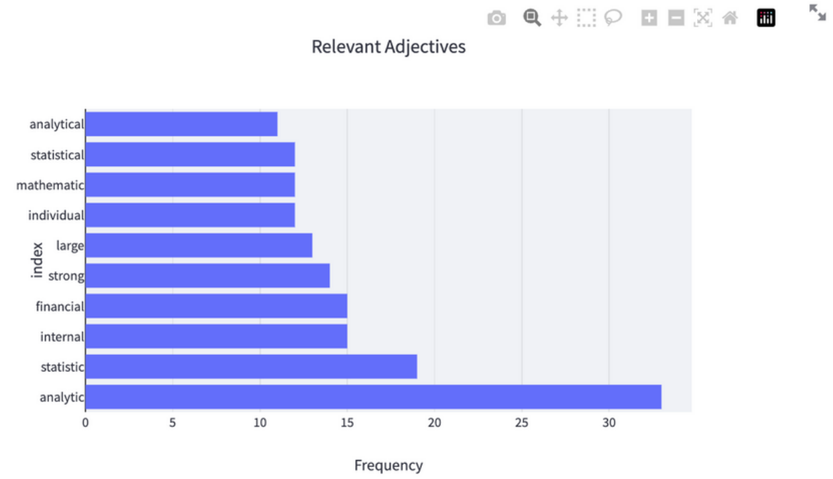
Data Scientist, Ar/vr Sales

Menlo Park

CA

Https://www.linkedin.com/jobs/search/?keywords=data%20scientist,%20ar/vr%20sales&location=Menlo%20Park

Language Recommendations for Applications



Conclusions
&
Recommendations

Conclusions & Recommendations

From the Engine:

- **Does not consider qualifications**, rather it outputs based on textual similarities.
- **Context-independent vectorization** minimizes the shape of data, yet strips meaning from text vectors.
- Recommended Jobs may provide insights into job locations that were not previously considered but are feasible in our evolving **remote-work environment**.
- **Project topics** may align with certain jobs based on titles and synonymous terms.
- Intentionally expand the **scope of industries** included in training data used for creating the vector space and subsequent recommendations.

Thank you!

ANY QUESTIONS?