Exploratory analysis of job offers

Aitor Pérez Pérez

Strategy

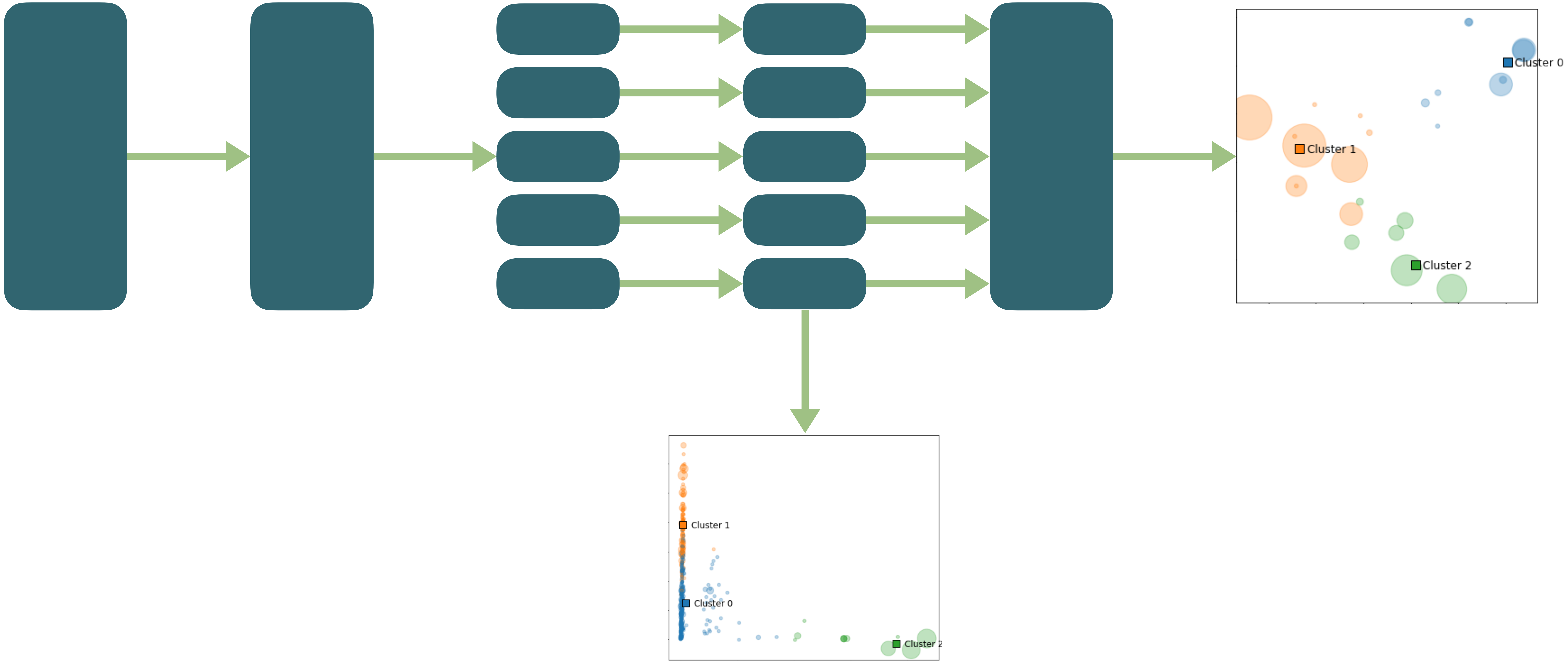
Data exploration

Vectorization

Clustering

Dimensionality reduction and visualization

Final classification



1 - Data exploration

Preprocessing:

Read data

Row counts

Missing values

Duplicates

id	Company	Title	Location	Responsibilities	Minimum Qualifications	Preferred Qualificatio
1	Google	Data Analyst, Product a...	New York, NY,...	Collect and analyze d...	Bachelor’s degree in Busi...	Experience partnering...
2	Google	Associate Account Strat...	Dublin, Ireland	Communicate with cust...	Bachelor's degree or equi...	Experience in sales, ...
3	Google	Solutions Architect, He...	New York, NY,...	Help compile customer...	BA/BS degree in Computer ...	Master's degree in Co...
4	Google	Associate Account Strat...	Dublin, Ireland	Implement creative wa...	Bachelor's degree or equi...	Experience in leading...
5	Google	Solution Architect, Goo...	Amsterdam, Ne...	Produce required desi...	BA/BS degree in a technic...	Experience with or de...

2 - Vectorization

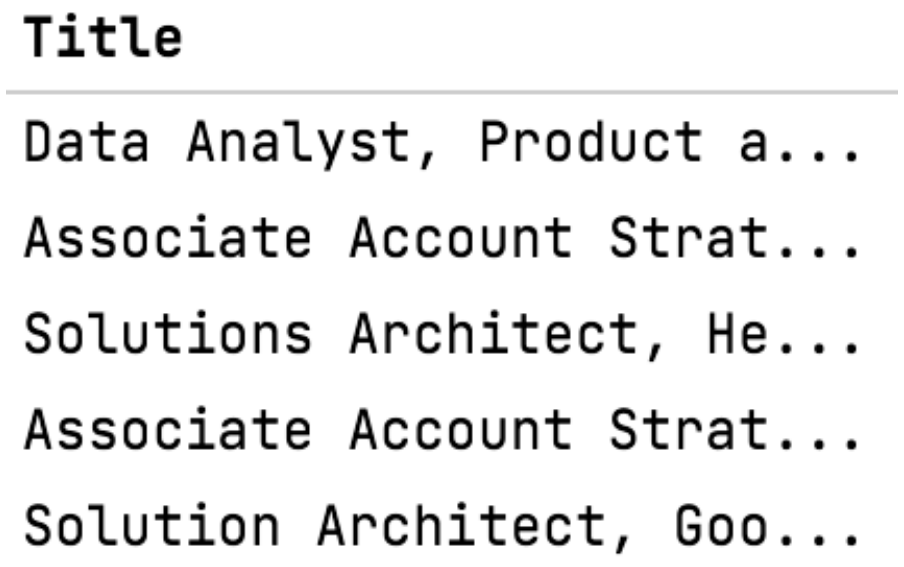
Text fields: Title, Location, Responsibilities, Min. Qualifications, Pref. Qualifications

Strategy: Vectorize fields **independently**

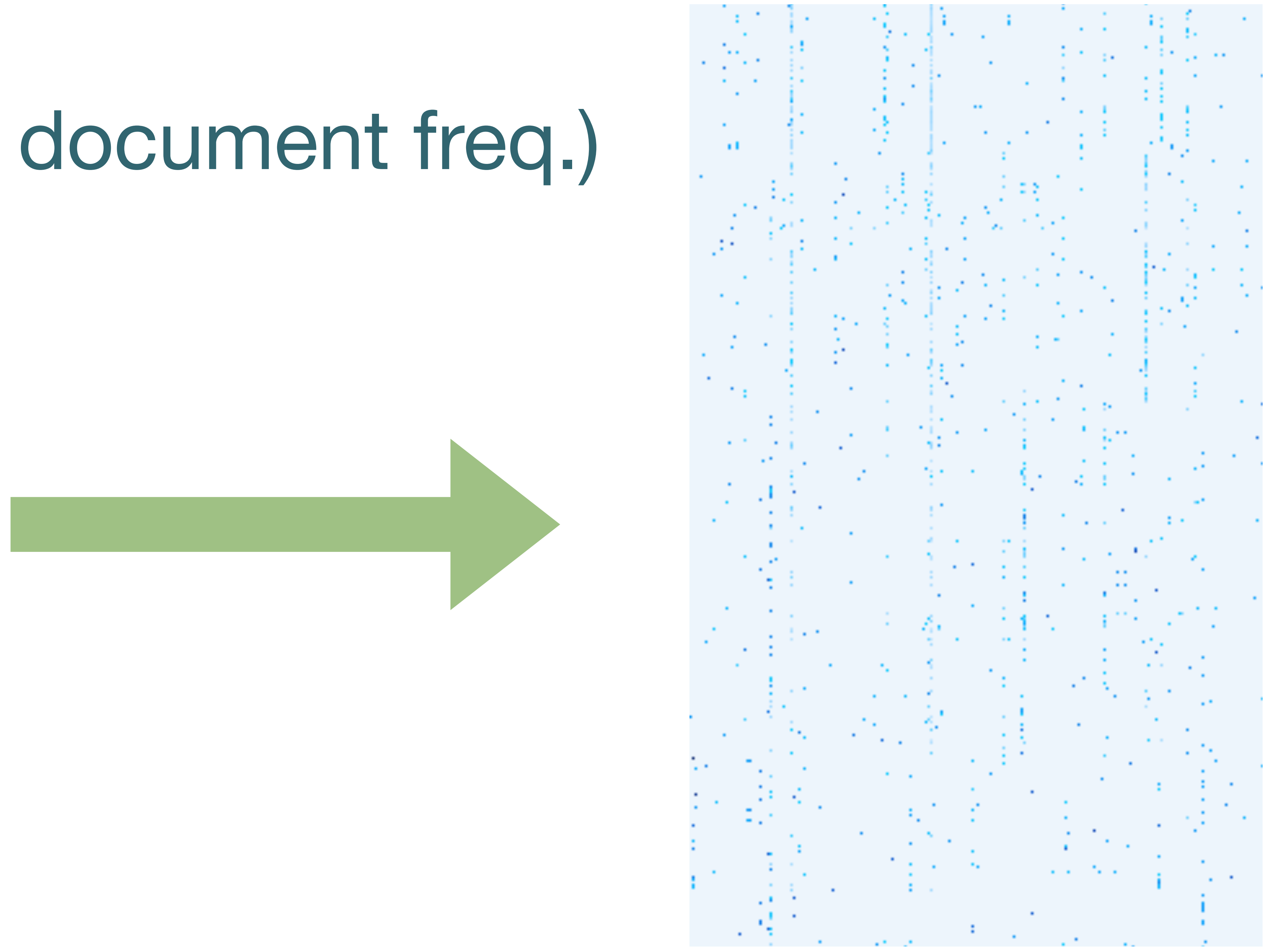
Vectorizer: **TfidfVectorizer** (term freq. + inverse document freq.)

Tokens: Words

Stopwords: English



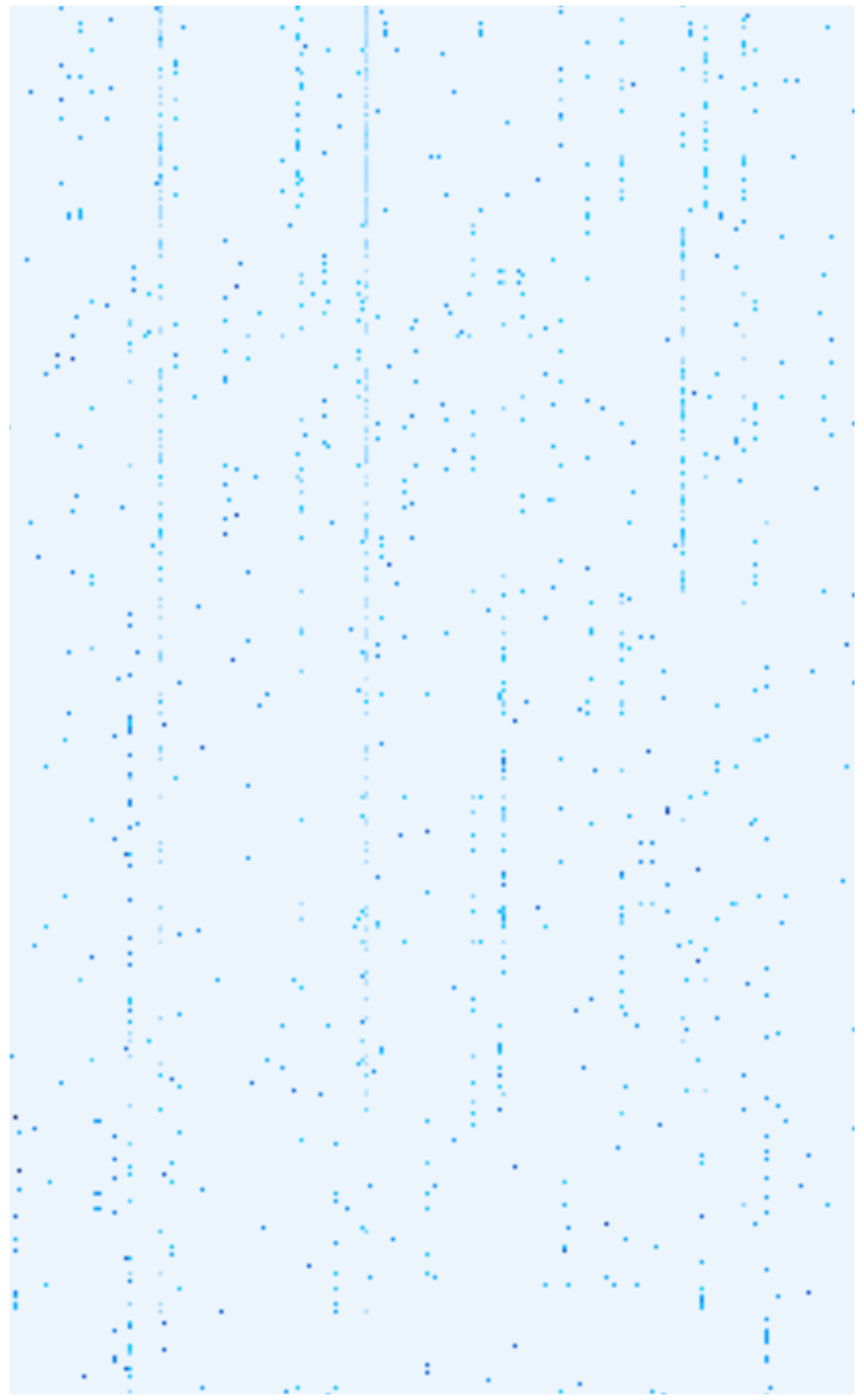
Title
Data Analyst, Product a...
Associate Account Strat...
Solutions Architect, He...
Associate Account Strat...
Solution Architect, Goo...

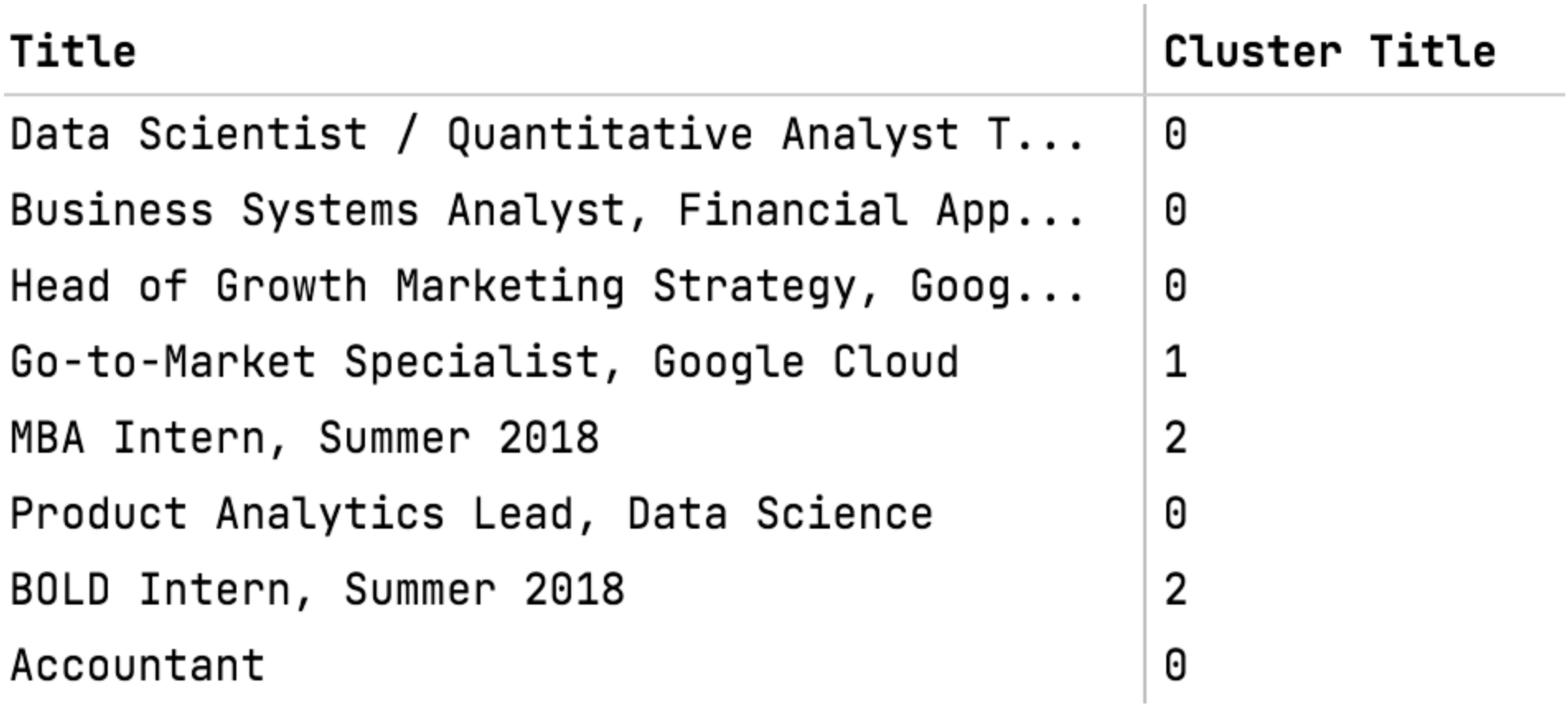


3 - Clustering

Algorithm: **K-means**

Number of clusters: Prescribed for each text field



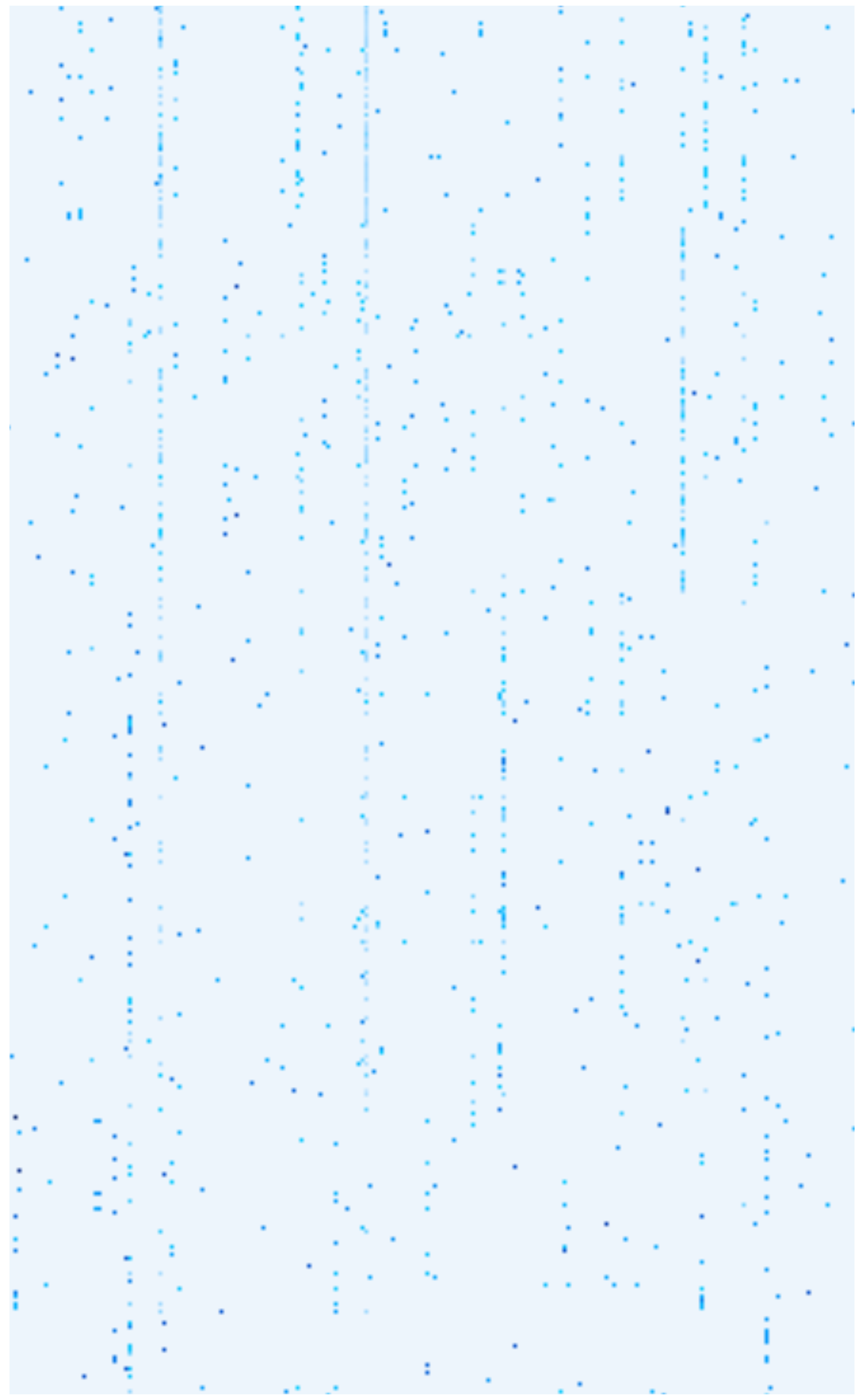


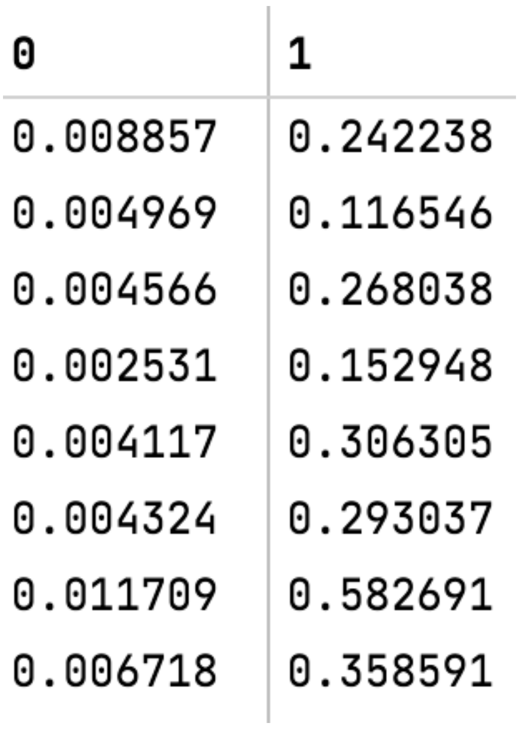
Title	Cluster Title
Data Scientist / Quantitative Analyst T...	0
Business Systems Analyst, Financial App...	0
Head of Growth Marketing Strategy, Goog...	0
Go-to-Market Specialist, Google Cloud	1
MBA Intern, Summer 2018	2
Product Analytics Lead, Data Science	0
BOLD Intern, Summer 2018	2
Accountant	0

4 - Dimensionality reduction and visualization

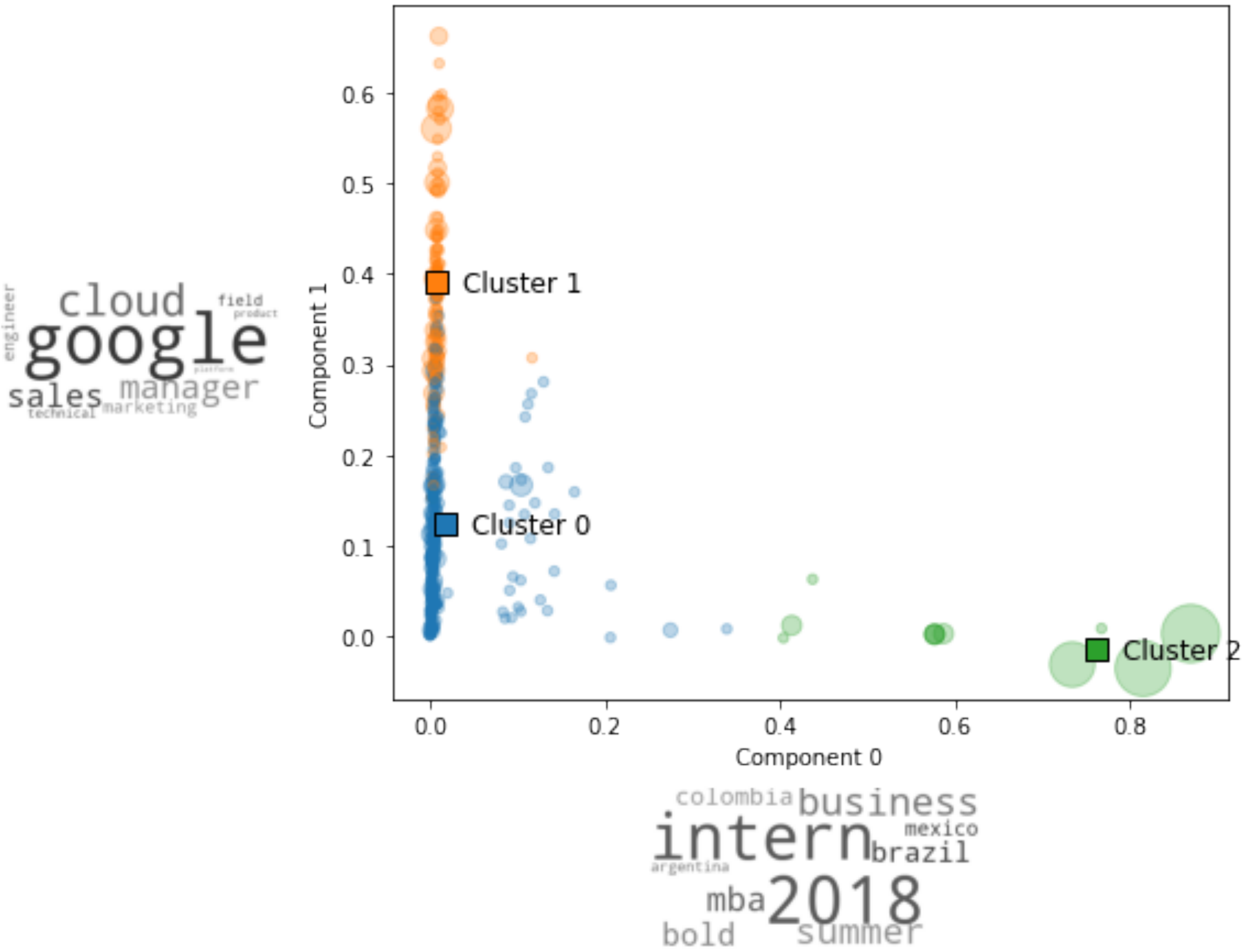
Algorithm: **Truncated Singular Value Decomposition (SVD)**

Number of components: 2





0	1
0.008857	0.242238
0.004969	0.116546
0.004566	0.268038
0.002531	0.152948
0.004117	0.306305
0.004324	0.293037
0.011709	0.582691
0.006718	0.358591

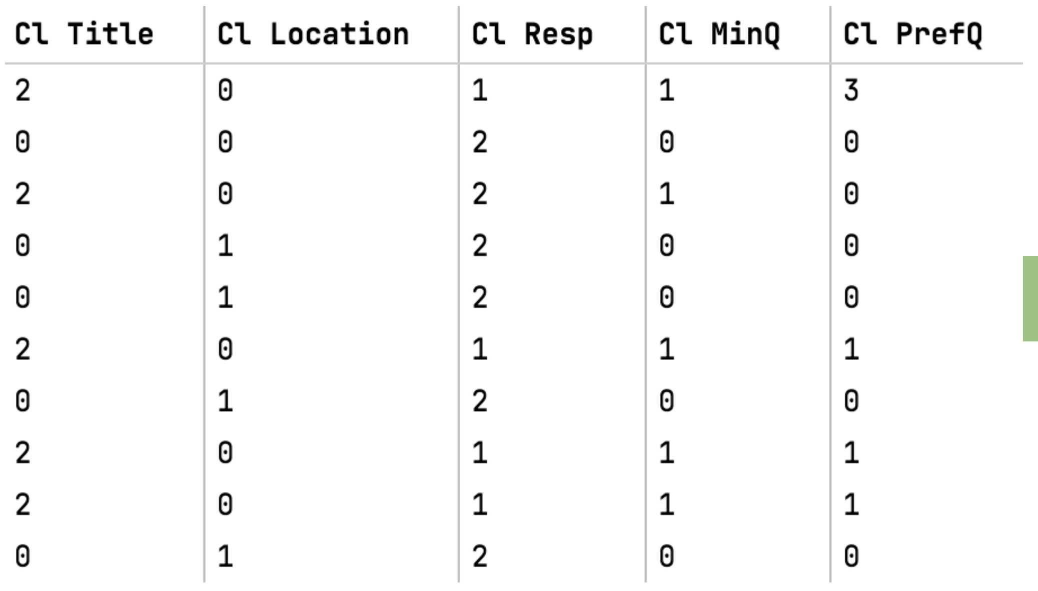


5 - Final classification

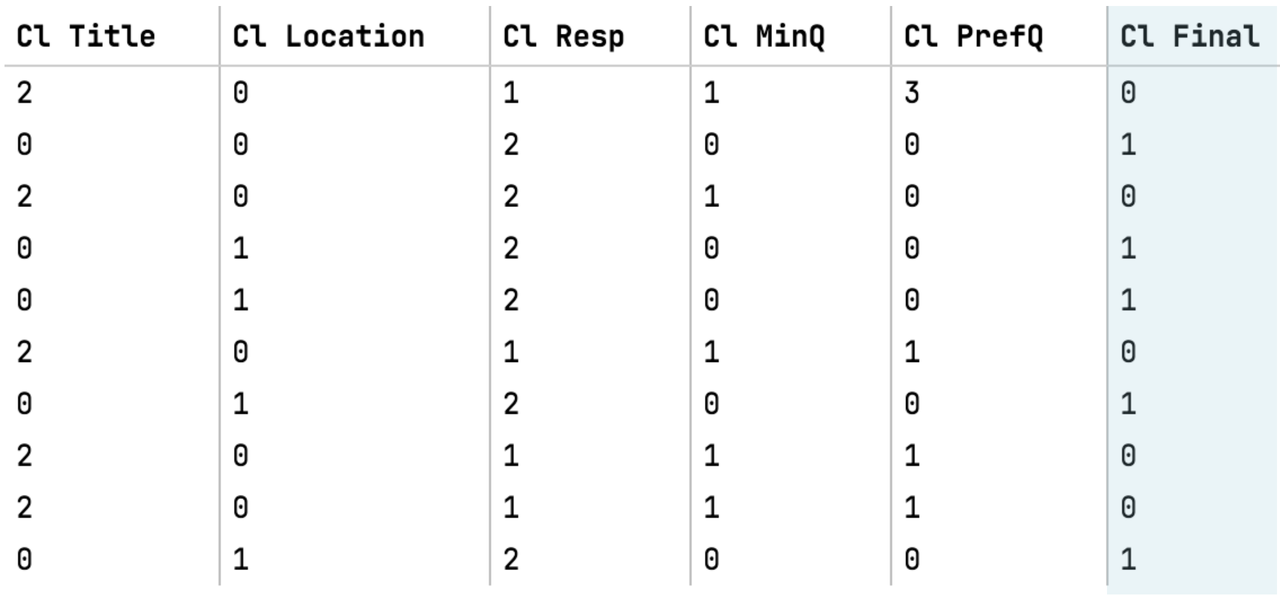
Goal: **Combine cluster labels** for all text fields

Step 1: Clustering cluster labels (**K-means**)

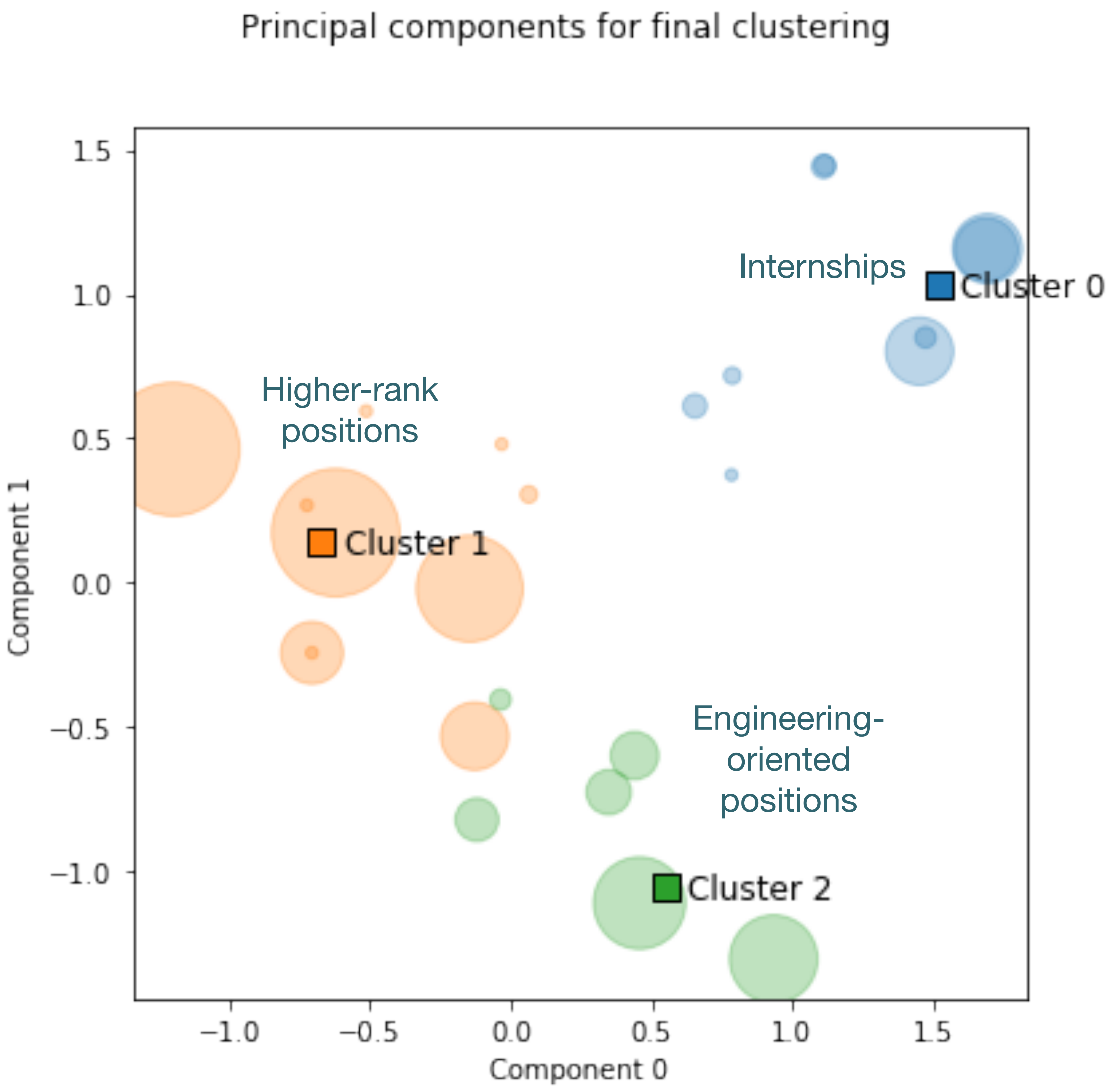
Step 2: Dimensionality reduction (**PCA**)



Cl Title	Cl Location	Cl Resp	Cl MinQ	Cl PrefQ
2	0	1	1	3
0	0	2	0	0
2	0	2	1	0
0	1	2	0	0
0	1	2	0	0
2	0	1	1	1
0	1	2	0	0
2	0	1	1	1
2	0	1	1	1
0	1	2	0	0



Cl Title	Cl Location	Cl Resp	Cl MinQ	Cl PrefQ	Cl Final
2	0	1	1	3	0
0	0	2	0	0	1
2	0	2	1	0	0
0	1	2	0	0	1
0	1	2	0	0	1
2	0	1	1	1	0
0	1	2	0	0	1
2	0	1	1	1	0
2	0	1	1	1	0
0	1	2	0	0	1



Principal components for final clustering

Improvement points

Technical:

Use **2-grams** instead of words as tokens

Change clustering (e.g. choice of K, hierarchical clustering)

Account for **negative correlation** in principal components

Replace one-hot encoding with **distance between centroids**

General:

Field-specific treatment (e.g. use **geospatial data** for Location)

Further cleanup of data (e.g. remove common expressions)

Use **3 components and 3D plots**

Talk to an expert and incorporate feedback!