Industry Clusters
Final Project – CS 505: Computational Tools for Data Science

Shreya Ramesh & Jia Yao

December 14, 2016

Introduction

Within general industries, there are further details and specializations, which are crucial to candidates' job searches. For example, within the all-encompassing title of "Software Development/IT," there are job specific details such as programming languages, database skills, etc. which will highly influence a candidate's employability and eligibility as well as the position's attractiveness to a candidate.

Using Latent Semantic Analysis, which is the application of Principal Component Analysis to term-document matrices, this project will identify clusters of jobs within each industry, in the ten cities with the most job listings for that industry, based on job descriptions available on indeed.com. The results of this project will likely have applications in several areas including candidates' decision-making to pursue certain specializations or certifications, and identifying trends within industries to further understand the job market.

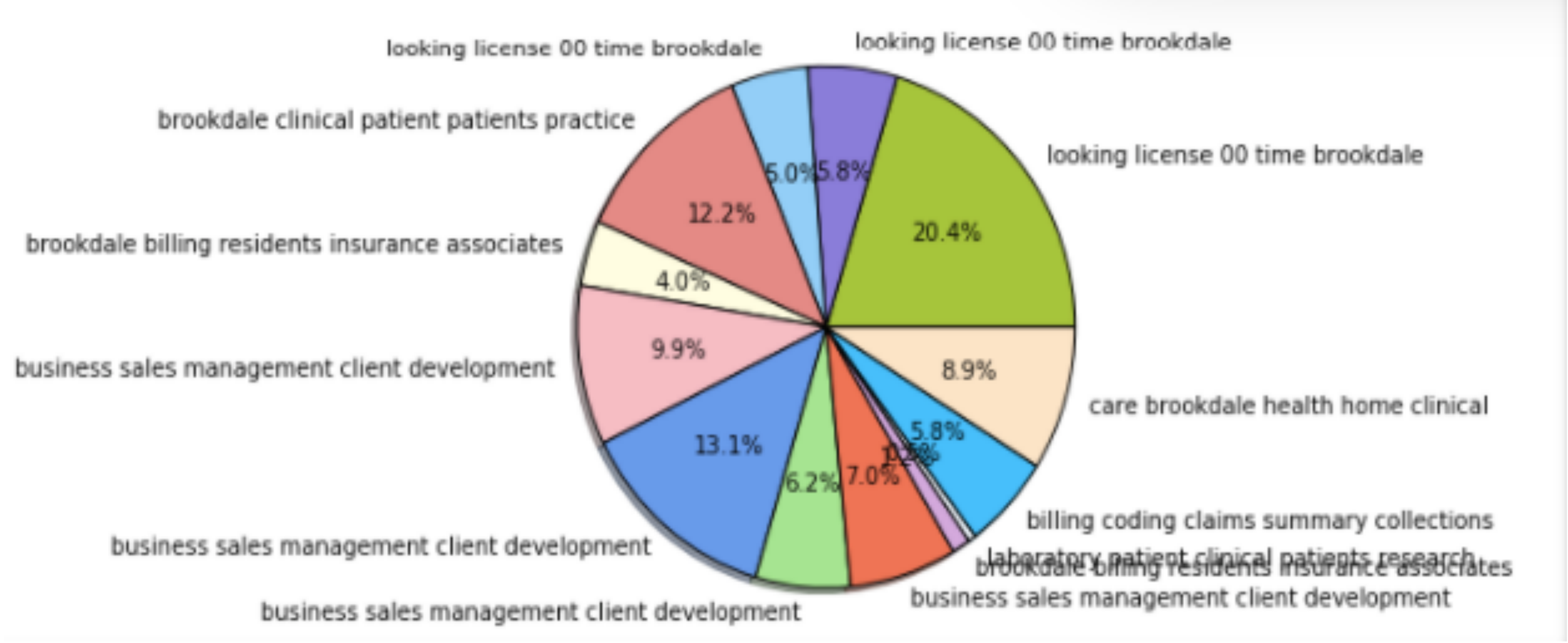
Datasets

The dataset is generated as a result of information from Glassdoor's API, and more detailed content from indeed.com. In order to determine which cities' data to scrape, the Glassdoor API provides the ability to pull the cities with the most jobs per industry. For example, the top ten cities with job openings in the field of Software Development are New York, San Francisco, Washington DC, Chicago, Atlanta, Austin, San Jose, San Diego, Boston, and Santa Clara.

For each city in each industry, the job descriptions available on indeed.com were pulled using web-scraping techniques. In each of the 29 industries in our dataset, there are between 2000 to 3000 job descriptions.

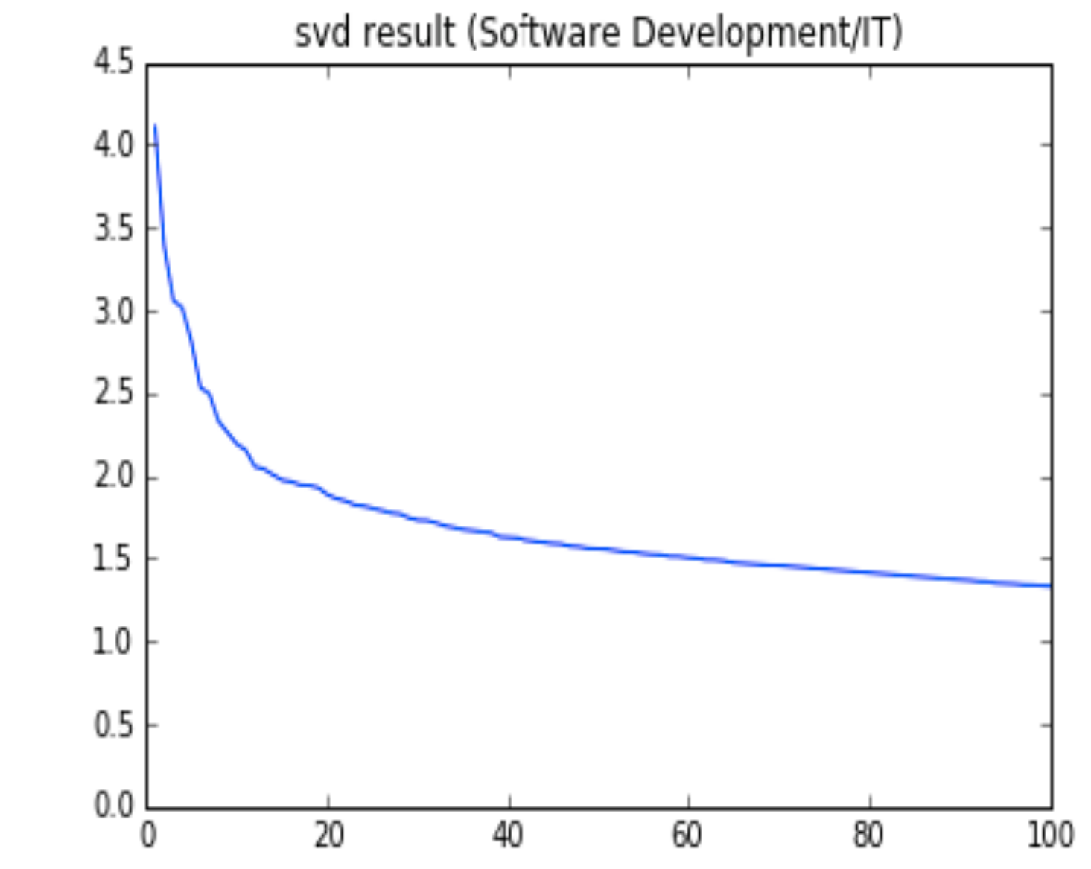
Limitations

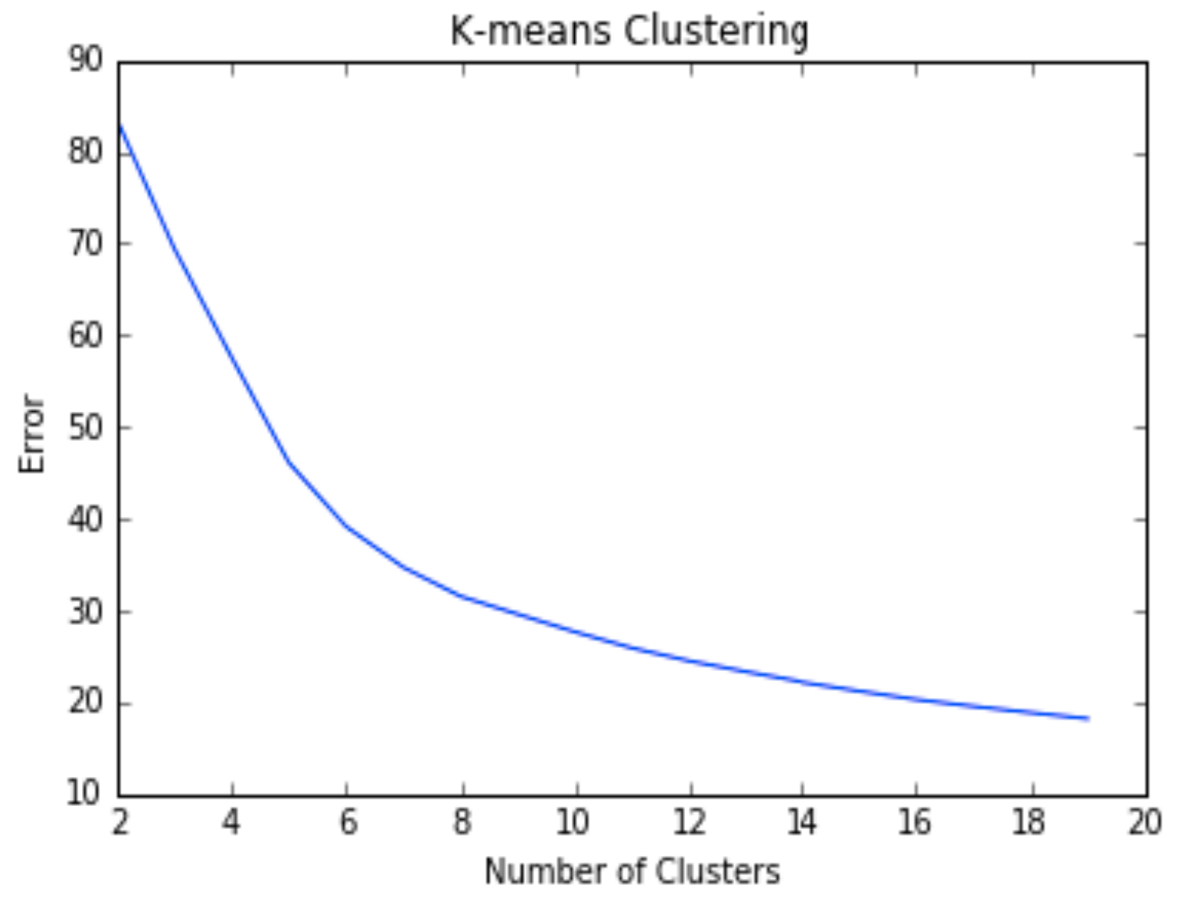
There are high frequency terms in job descriptions, such as "summary" and "requirements", that are not helpful for revealing what the specific job is about. Also, some of the clusters generated were not as informative as others, as they center on locations mentioned in the job descriptions. An example is the clusters of the Medical/Healthcare industry.



Techniques

The TfidfVectorizer is used to encode text data, which are the job descriptions in our dataset, as term-document matrices, in which each job description yields a vector of term frequencies. Then dimensionality reduction is applied to extract the most significant principal components, which represent important concepts in the job descriptions. Using graphs showing the spectrum of Singular Value Decomposition, a reasonable number of principal components for each industry was chosen. The k-means error graphs were used to determine the optimal number of clusters; the value chosen for this project should have an error of less than 70 on the graph.





The second component of the analysis is running k-means clustering on the document encodings with dimensionality reduced. At a high level, the k-means algorithm moves the means of points until the given number of means (k) converge. Below is the standardized k-means algorithm.

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2$$

Results

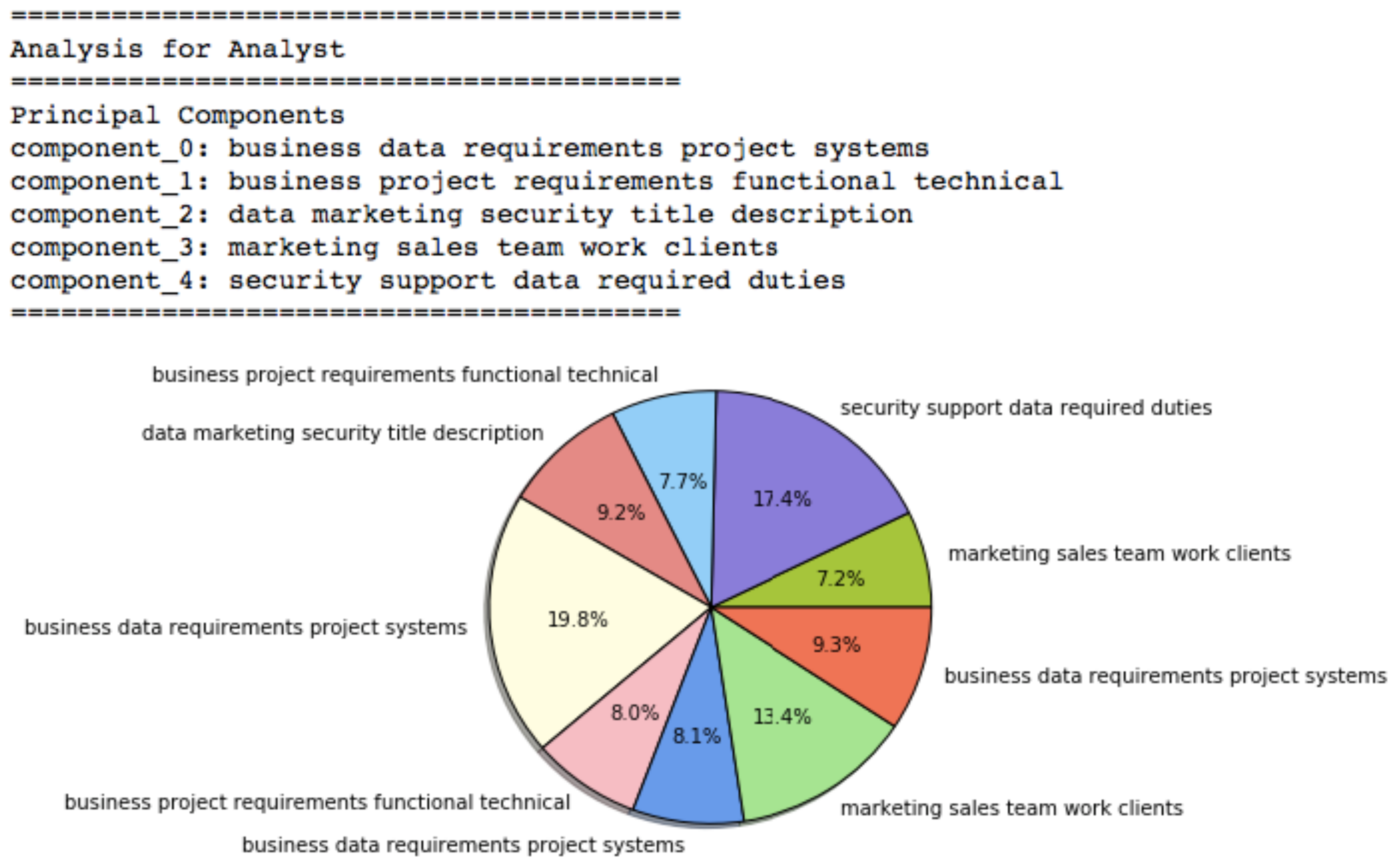
For visualization, pie-charts were generated showing how the clusters make up the industry as a whole.

For each industry, k-means analysis produces cluster assignments for every job description, and by looking at the weights of principal components for all jobs placed in the same cluster, a text label for the cluster overall was generated. The pie charts indicate that for each percentage, that percentage of jobs belong to a cluster that is dominated by jobs with that label.

Conclusion

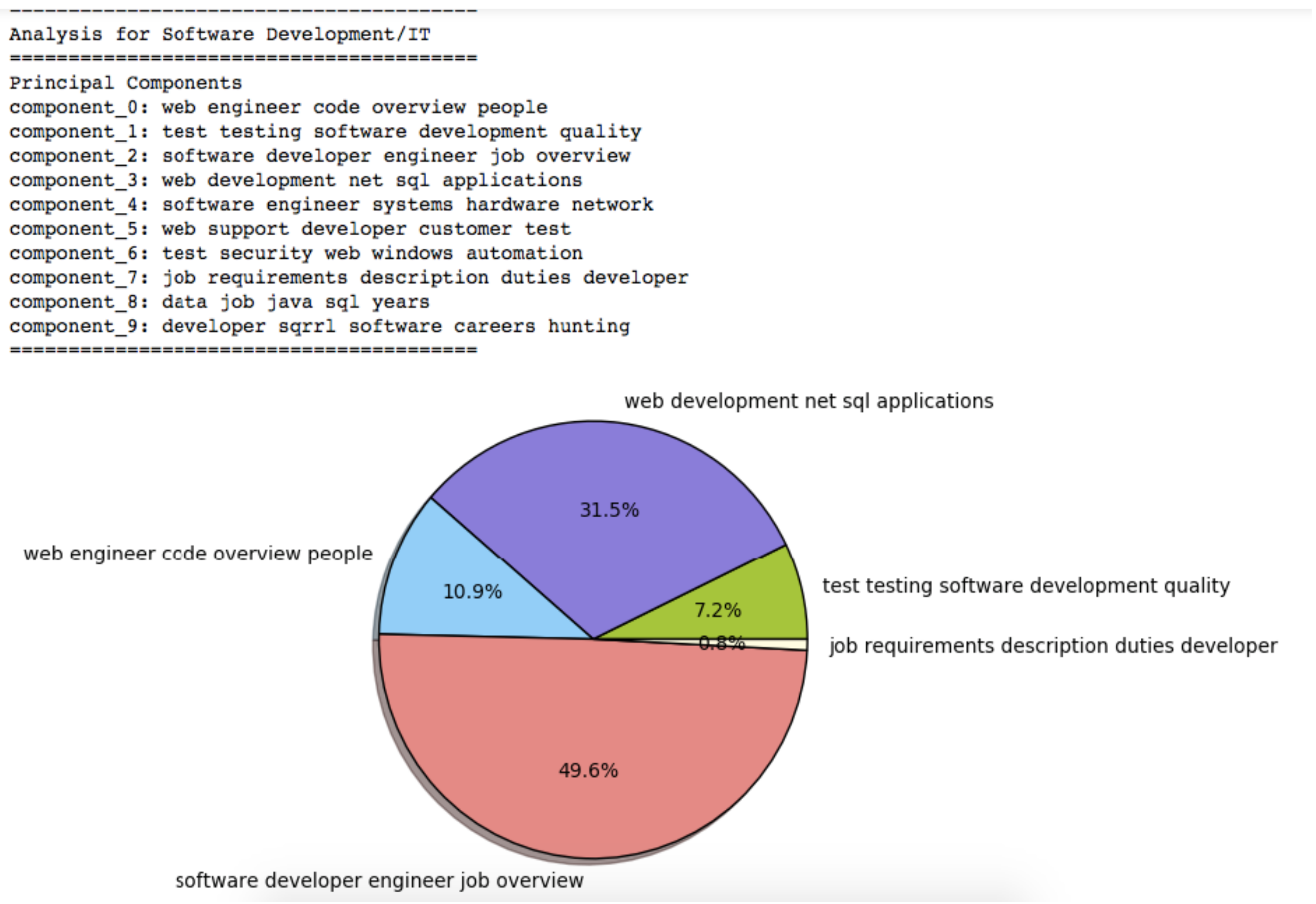
Most of the industries (examples shown on the right) generated useful principal components, which seem to relevantly describe the industry as a whole. However, some of the industries (such as Medical/Healthcare, pictured on the left) center on certain streets or locations, which may not be particularly useful for the objective of this project.

Visualization of Results



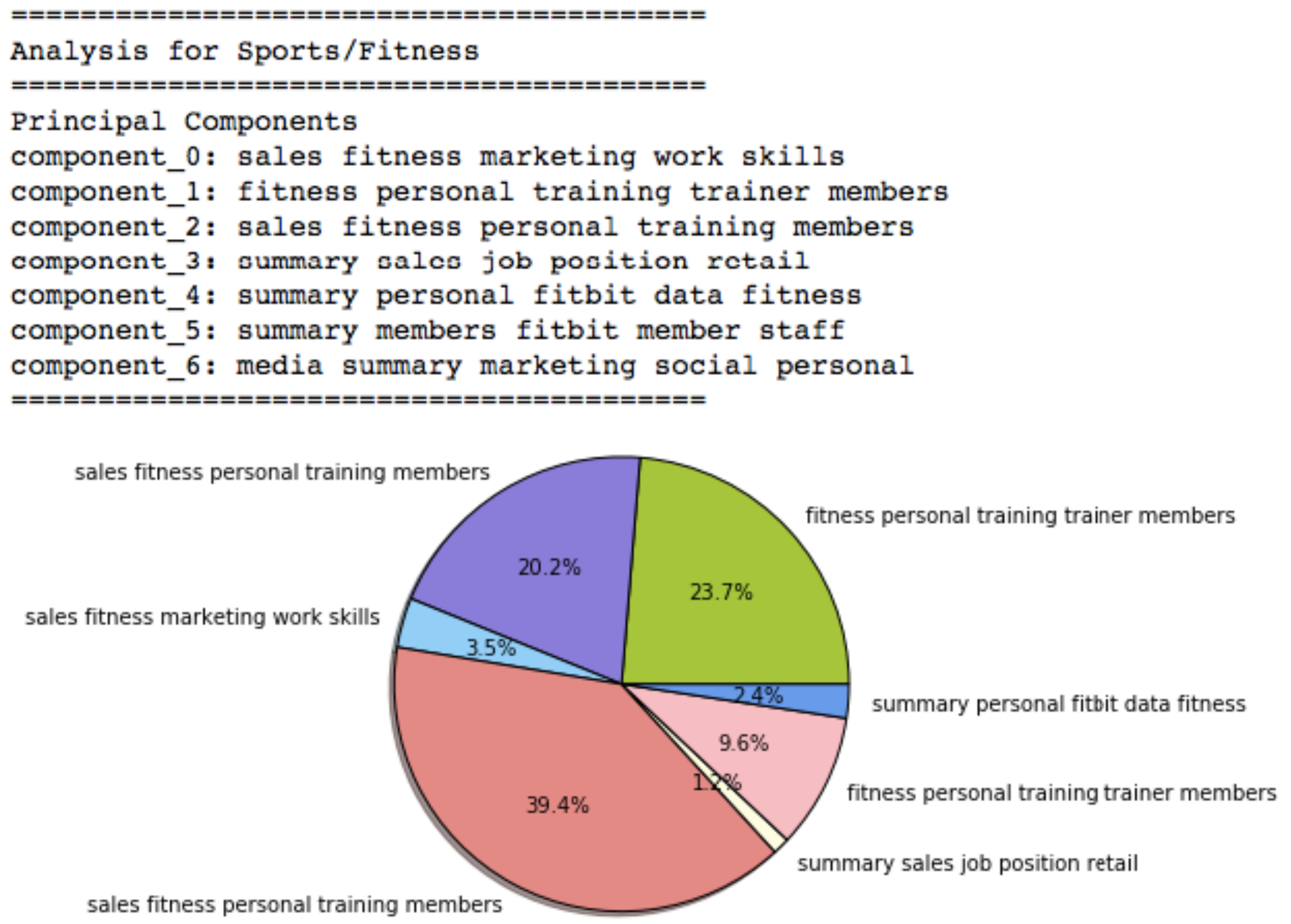
```
========================================
Analysis for Analyst
========================================
Principal Components
component_0: business data requirements project systems
component_1: business project requirements functional technical
component_2: data marketing security title description
component_3: marketing sales team work clients
component_4: security support data required duties
========================================
```

business project requirements functional technical
data marketing security title description
security support data required duties
marketing sales team work clients
business data requirements project systems
business data requirements project systems
business project requirements functional technical
marketing sales team work clients
business data requirements project systems
7.7%
17.4%
9.2%
7.2%
19.8%
9.3%
8.0%
8.1%
13.4%



```
----------------------------------------
Analysis for Software Development/IT
========================================
Principal Components
component_0: web engineer code overview people
component_1: test testing software development quality
component_2: software developer engineer job overview
component_3: web development net sql applications
component_4: software engineer systems hardware network
component_5: web support developer customer test
component_6: test security web windows automation
component_7: job requirements description duties developer
component_8: data job java sql years
component_9: developer sqrrl software careers hunting
========================================
```

web development net sql applications
web engineer ccde overview people
test testing software development quality
job requirements description duties developer
software developer engineer job overview
31.5%
10.9%
7.2%
0.8%
49.6%



```
========================================
Analysis for Sports/Fitness
========================================
Principal Components
component_0: sales fitness marketing work skills
component_1: fitness personal training trainer members
component_2: sales fitness personal training members
component_3: summary salcs job position rctail
component_4: summary personal fitbit data fitness
component_5: summary members fitbit member staff
component_6: media summary marketing social personal
========================================
```

sales fitness personal training members
fitness personal training trainer members
sales fitness marketing work skills
summary personal fitbit data fitness
fitness personal training trainer members
summary sales job position retail
sales fitness personal training members
20.2%
23.7%
3.5%
2.4%
9.6%
1.2%
39.4%