ComputerScience

Boston University

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 details such as programming languages, frameworks, etc. which will highly influence a candidate's employability and eligibilåity as well as the position's attractiveness to a candidate.

Using document clustering and analysis, this project will identify clusters 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, analysis of which cities have a demand for certain sub-industries, and perhaps even identifying trends within industries to further understand the job-market.

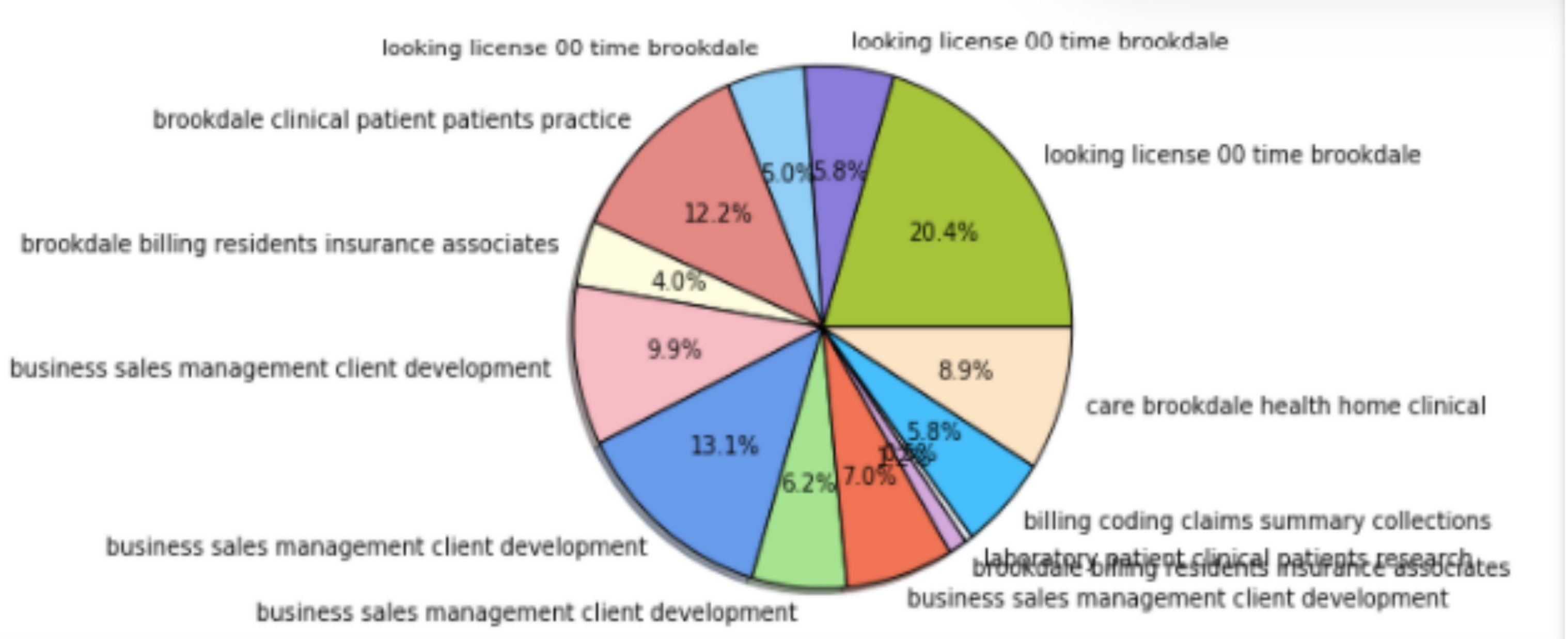
Datasets

The dataset is generated as a result of information from Glassdoor's API, and also includes 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. This information, of the top ten cities per industry, was stored in a CSV (comma separated values) file. 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, a search was run in indeed.com, and using web-scraping techniques, job descriptions from each city and industry combination were generated. These were also stored in CSV files. The Python script used to generate these files is entitled indeed-scraper.py.

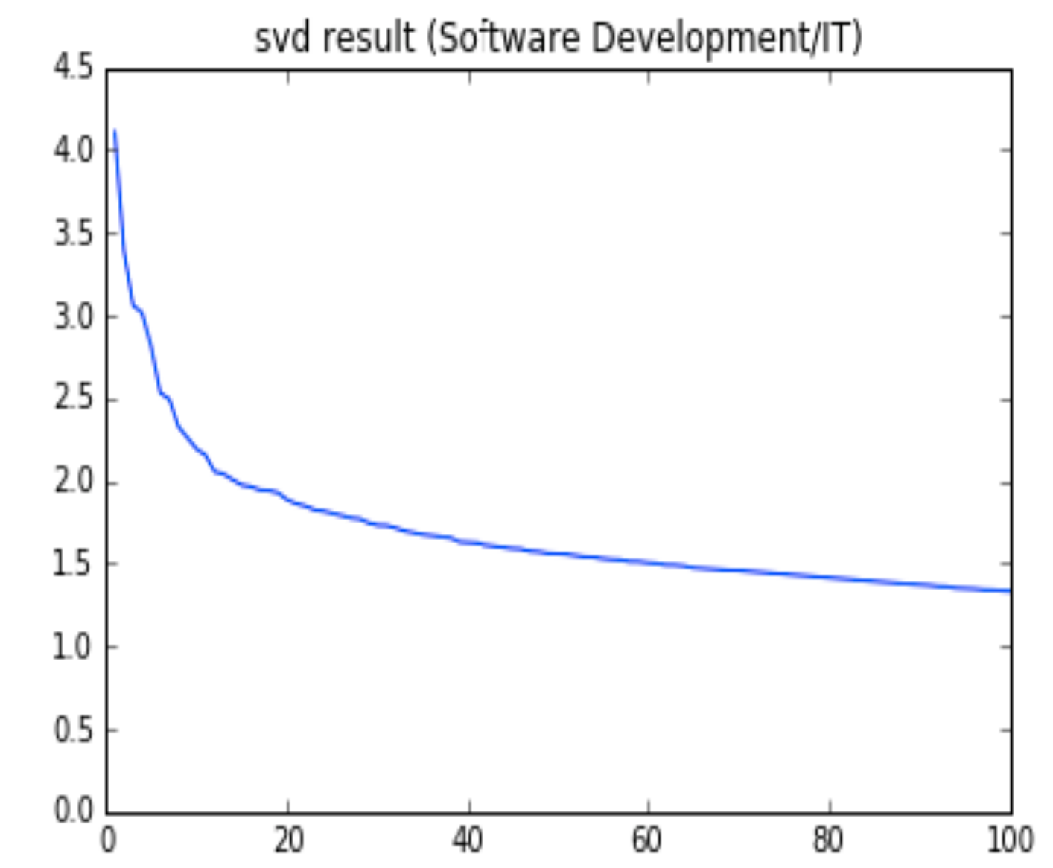
Limitations

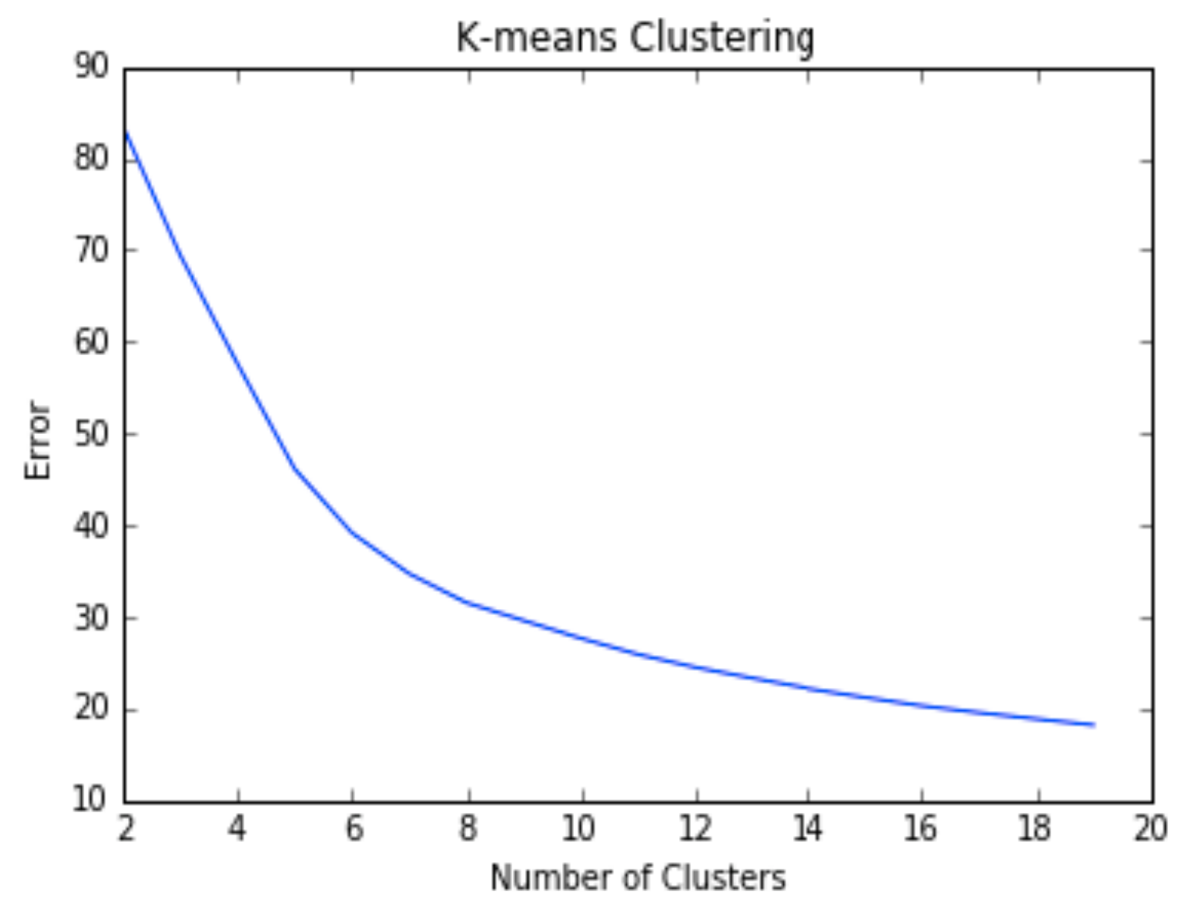
Some of the clusters generated were not as informative as others, as they centered on locations mentioned in the job descriptions. An example is of the Medical/Healthcare industry.



Techniques

The TfidfVectorizer which is used is based on term frequency, in conjunction with inverse document-frequency. This technique essentially transforms the document or text into a vector, and based on the vector clusters the words. The code used can be seen in the ipython notebook file Dataset-Analysis.ipynb, where the first component generates the Singular Value Decomposition (SVD) graphs for each of the industries. Through the SVD graphs, the number of principal components for each cluster is determined. The error graphs generated 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 runs the k-means algorithms on the text vector. 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

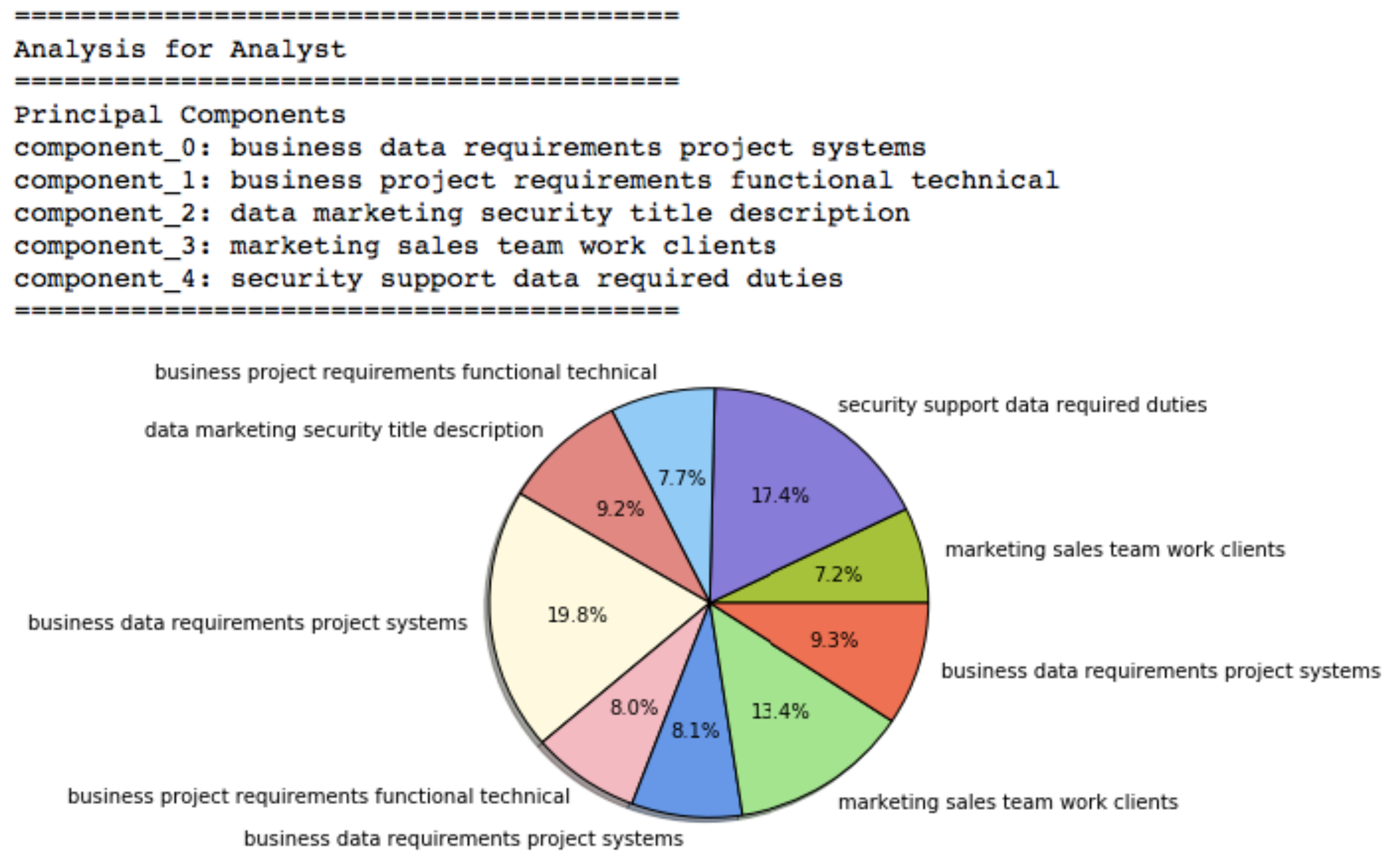
For the final component the principal components for each dataset are generated along with a visualization of a pie-char showing how the clusters make up the industry as a whole.

Using the sklearn.cluster library, KMeans analysis is used to generate the principal components for each industry, and a corresponding pie chart using the matplotlib.pyplot library. 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

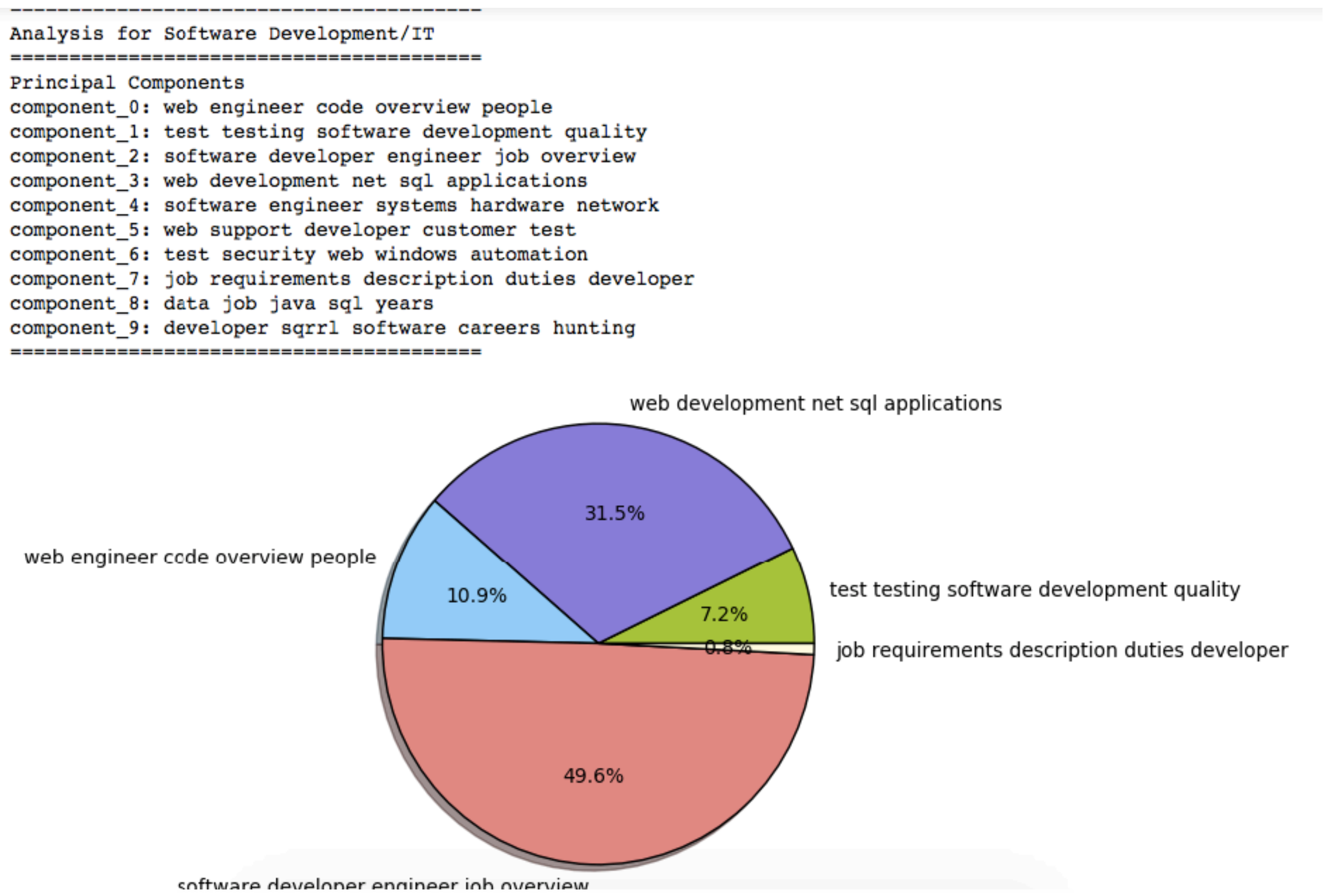
Most of the industries generated useful principal components, which seems to relevantly describe the industry as a whole. However, some of the industries (such as Medical/Healthcare, pictured on the left) center on certain streets or locations, which may not be particularly useful for the objective of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

Visualization of Results



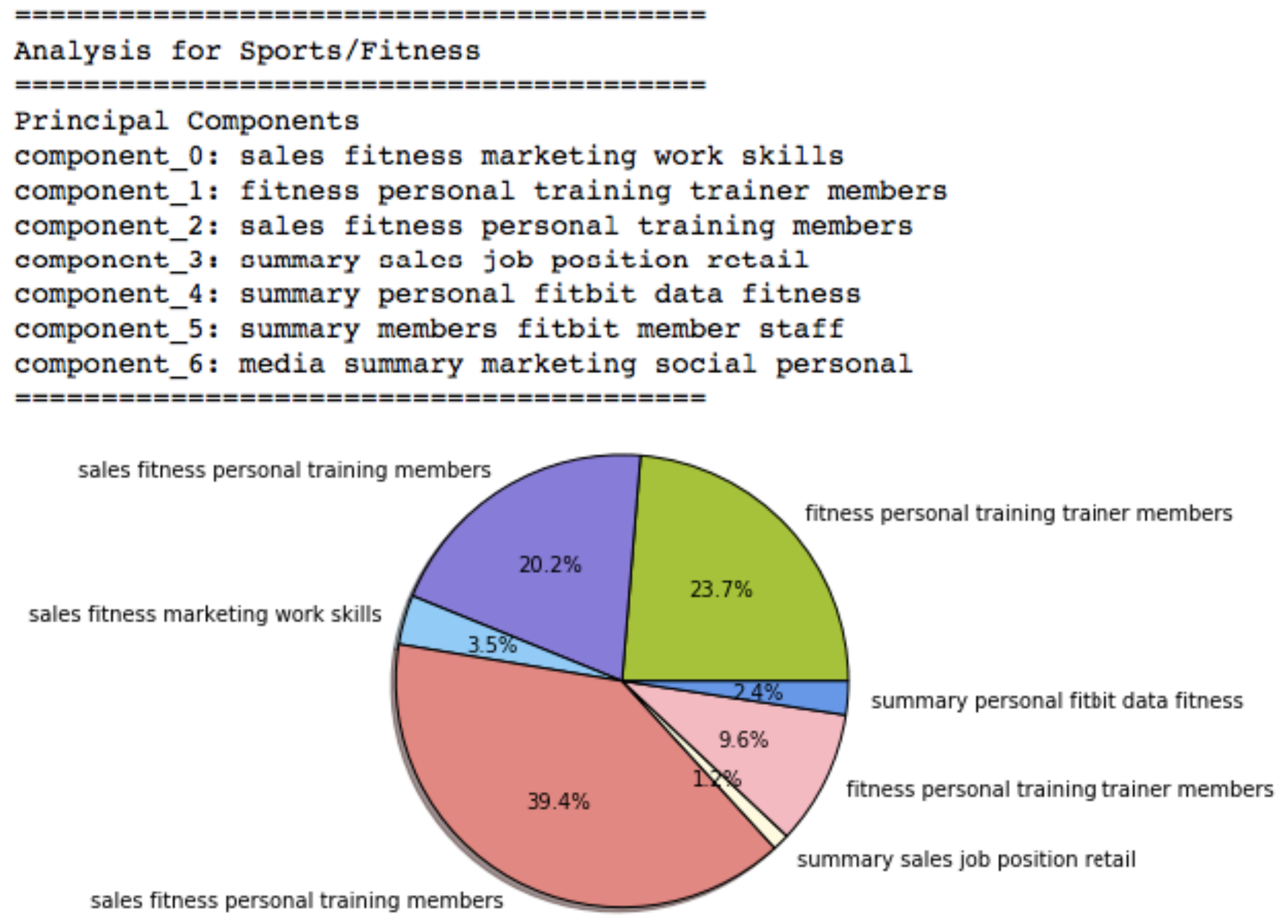
```
========================================
Analysis for Analyst
========================================
Principal Components
component_0: business data requirements project systems
component_1: business project requirements functional technical
component_2: data marketing security title description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

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



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



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