Industry Clusters

Final Project for CS 591

Shreya Ramesh and Jia Yao

12/14/2016

Abstract

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 eligibility 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.

Introduction

The motivation behind this project is to delineate areas within industries of high demand, in the cities that have the greatest demand for each industry. 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.

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.

The results show concrete areas of focus which definitely can be deemed useful and appropriate. For example, within the main industry title of "Accounting/Finance," we found categories of "fin ancial reporting analysis management monthly business finance budget internal annual" which is a specific area within this 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 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. 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.

For the final component, seen in the ipython notebook file Cluster-Analysis.ipynb, 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. These results can be viewed in the appendix.

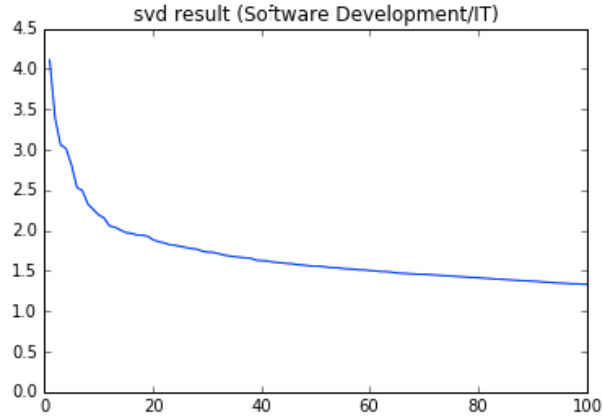
Datasets

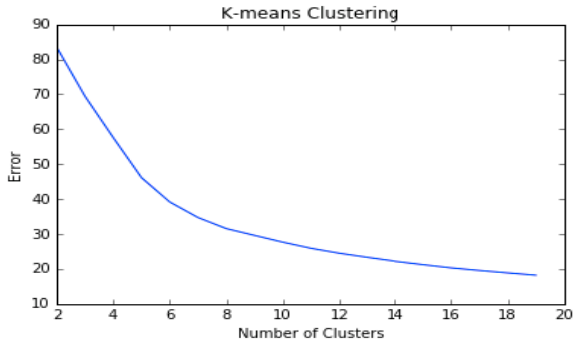
The dataset is generated as a result of information from Glassdoor's API, and also includes more detailed content from indeed.com. In order to determine which cities' data to scrape, the Glassdoor API provides the ability to pull the cities with the most jobs per industry. This information, of the top 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. A CSV file of all the industries' top ten cities can be found in the appendix. The Python script, entitled cities.py, used to generate these CSV files can also be found in the appendix.

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.

Results

The SVD graphs generated determine the number of principal components per cluster. The value of k for each industry is determined based on the error and silhouette score. For standardization, the number of clusters has to have an error less than 70. Below is an example of the error graphs used for the determination of k-values.



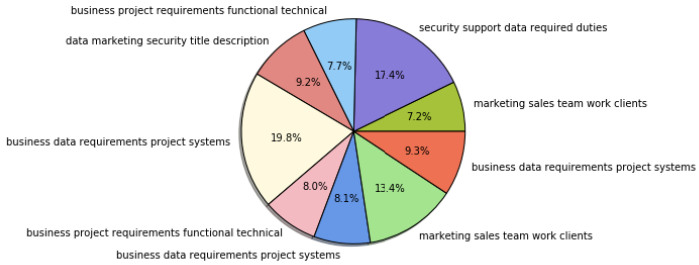


Based on this analysis, below are the values for number of principal components and number of clusters.

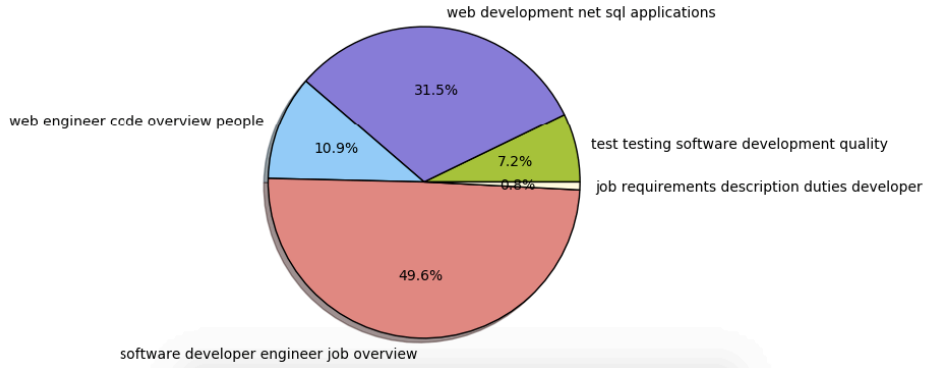
Industry	Number of Clusters	Number of Principal Components
Accounting/Finance	18	15
Administrative	18	20
Analyst	9	5
Architecture/Drafting	13	15
Art/Design/Entertainment	18	10
Banking/Loan/Insurance	12	10
Beauty/Wellness	10	13
Business Development/Consulting	18	7
Education	18	5
Facilities/General Labor	10	5
Hospitality	18	10
Human Resources	18	7
Installation/Maintenance Repairs	18	5
Legal	14	5
Manufacturing/Production/Construction	18	10
Marketing/Advertising/PR	18	10
Medical/Healthcare	13	10
Product/Project Management	18	10
Real Estate	16	20
Restaurant/Food Services	18	15
Retail	18	10
Science/Research	10	15
Security/Law Enforcement	6	10
Senior Management	2	7
Skilled Trade	2	7
Software Development	5	10
Sports/Fitness	7	7
Travel/Transportation	6	7
Writing/Editing/Publishing	6	7

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. Below are a few examples of results; all of the results can be viewed in the Appendix file, categorized by whether the results were informative or not.

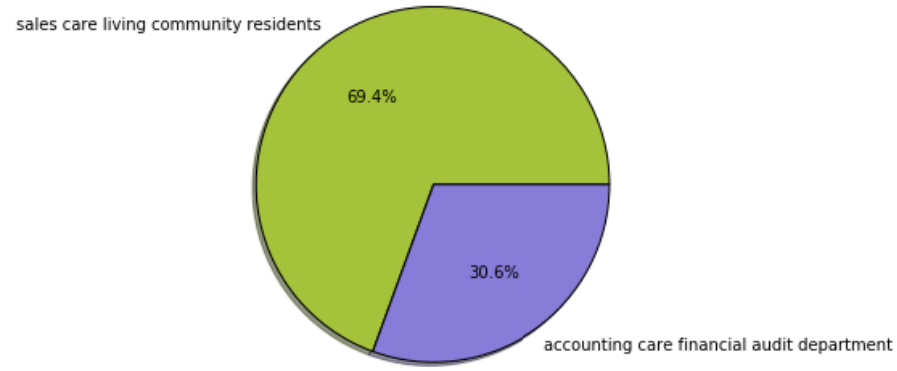
```
========================================
Analysis for Analyst
========================================
Principal Components
component_0: business data requirements project systems
component_1: business project requirements functional technical
component_2: data marketing security title description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



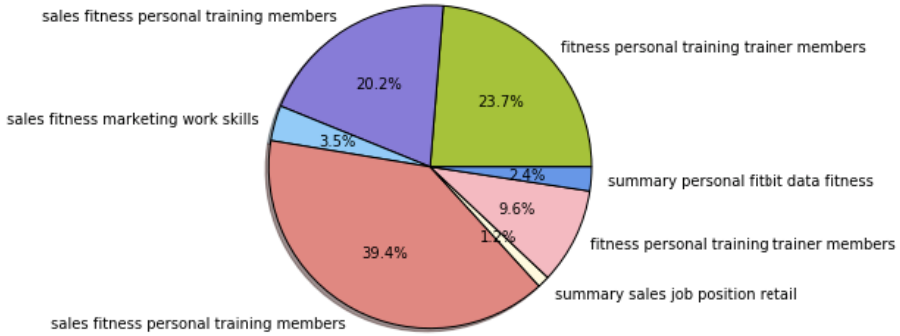
```
----------------------------------------
Analysis for Software Development/IT
========================================
Principal Components
component_0: web engineer code overview people
component_1: test testing software development quality
component_2: software developer engineer job overview
component_3: 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
====================================
```



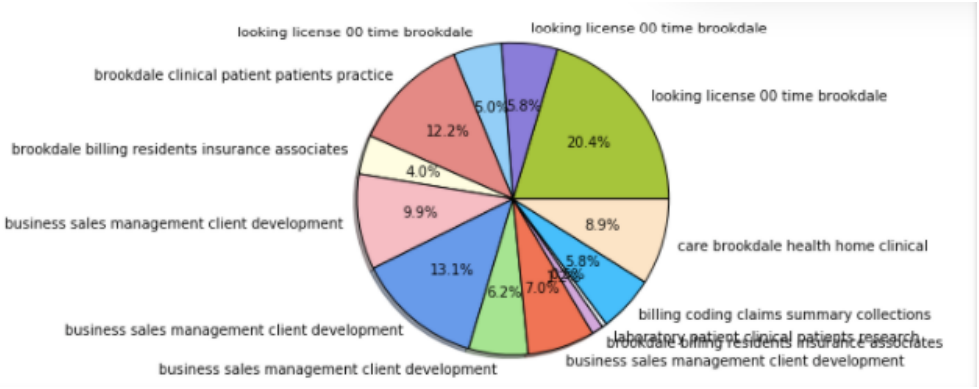

```
========================================
Analysis for Senior Management
========================================
Principal Components
component_0: accounting care financial audit department
component_1: sales care living community residents
component_2: media marketing care living community
component_3: accounting media marketing financial audit
component_4: marketing accounting financial media care
component_5: client project accounting care living
component_6: project program mit sales manager
========================================
```



```
========================================
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
========================================
```




```
========================================
Analysis for Medical/Healthcare
========================================
Principal Components
component_0: business sales management client development
component_1: looking license 00 time brookdale
component_2: care brookdale health home clinical
component_3: billing coding claims summary collections
component_4: brookdale billing residents insurance associates
component_5: care health billing home healthcare
component_6: summary patient patients position equipment
component_7: brookdale clinical patient patients practice
component_8: laboratory patient clinical patients research
component_9: office clinical data research study
========================================
```



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 above) center on certain streets or locations, which may not be particularly useful for the objective of this project. Furthermore, 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.