BIA660 - Text Mining of Job Description

Instructor : Rong Liu

Fanshu Li, Junxin Xia, Yuyi Yan

Contents

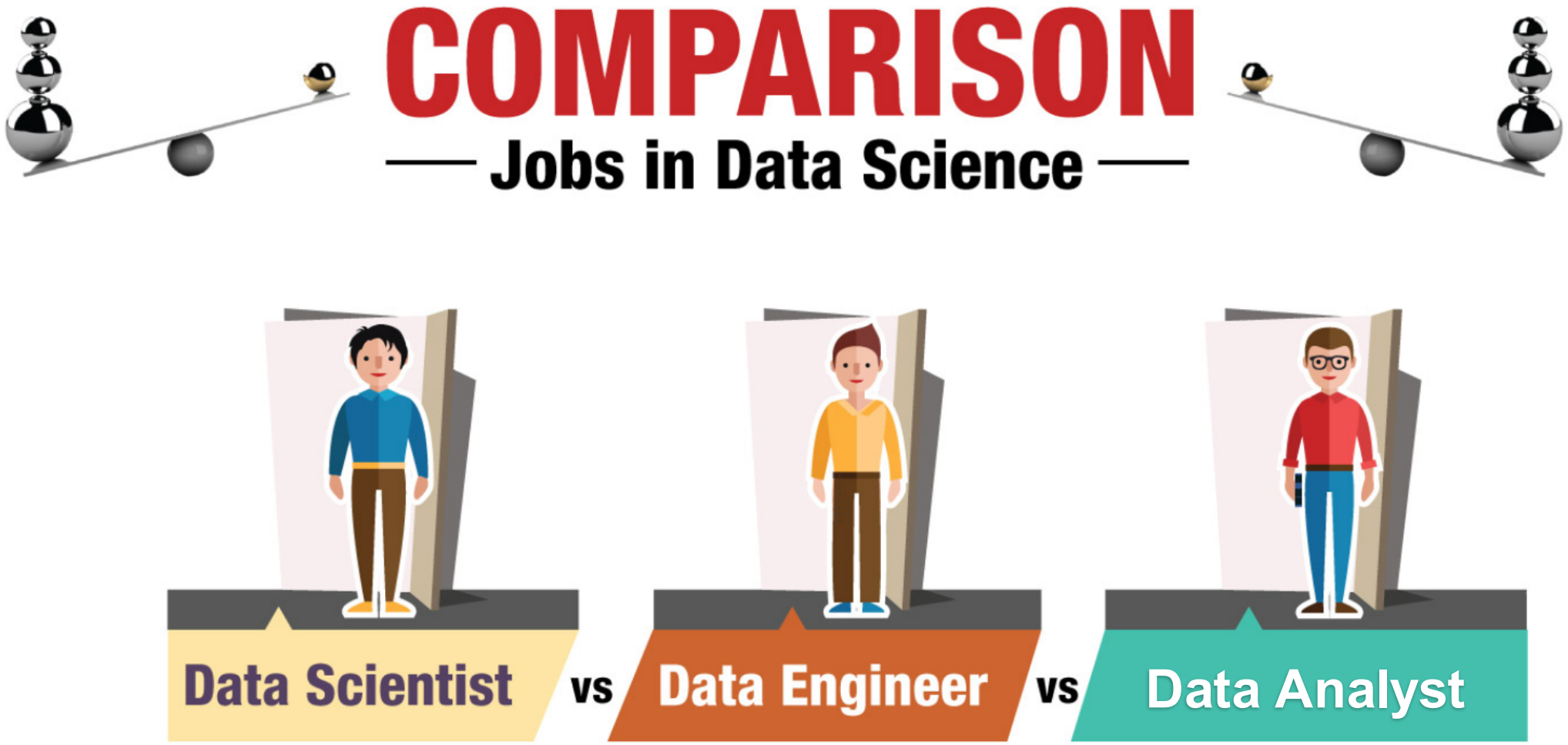
NO.1

Introduction

Introduction

1. **World has entered the big data era.**
2. **Data Analytics is one of the most popular majors in our society.**
3. **Depressed market make it difficult for students to find jobs**

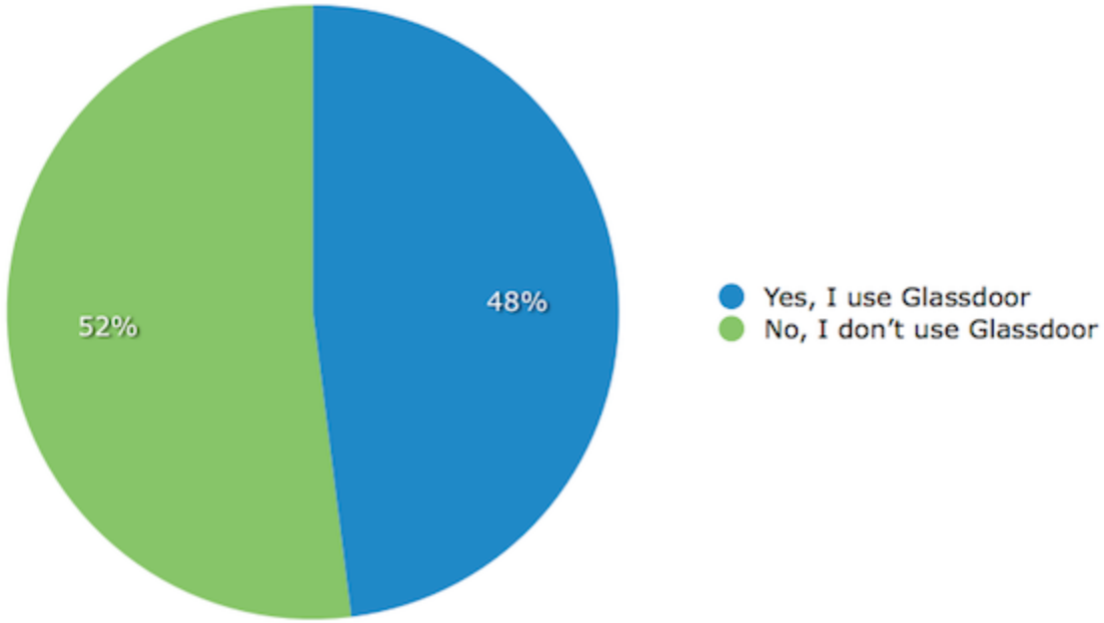
Introduction



Data Source

glassdoor

Why?



Project Goals

Something to descripte your ideas.

Clustering

1

Compare different requirents for applying the three position

2

Classification

Insights about job description

NO.2

Data Scraping & Preprocessing

Data Scraping

Job Type | Date Posted | Salary Range | Distance | More | Create Job Alert

Data Scientist Jobs in New York, NY — 2,416 Jobs

Data Scientist
Raise Marketplace Inc. – New York, NY
$71k-$108k (Glassdoor est.)
EASY APPLY
21 days ago

Data Scientist
Kensho – New York, NY
$115k-$169k (Glassdoor est.)
10 days ago

Data Scientist
HotelTonight – New York, NY
$111k-$163k (Glassdoor est.)
6 days ago

Data Scientist
Teachers Pay Teachers – New York, NY
7 days ago

Data Scientist/ Sr. Data Scientist

Data Scientist

Easy Apply | Save

2.9 ★ **Raise Marketplace Inc.** – New York, NY
Glassdoor Estimated Salary: $71k-$108k

Job | Company | Rating | Reviews | Why Work For Us

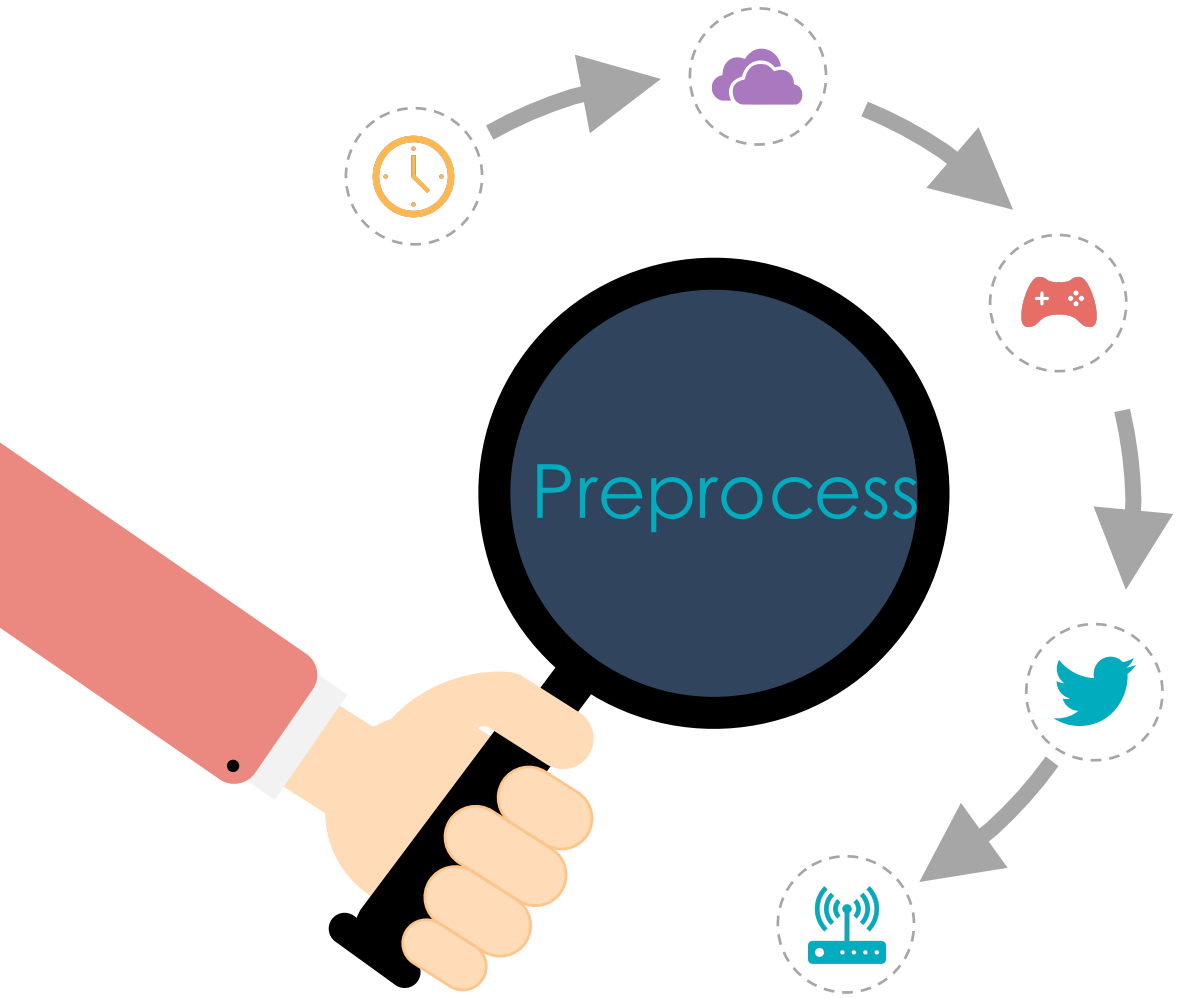
About Raise

Were Raise, a leading retail payments company and the worlds largest gift card marketplace, that connects consumers to buy discount gift cards or sell their unwanted cards for cash. Were partnered with over 400 national brands, offering retailers a new form of digital payment, while incentivizing consumers to make their money worth more.

Since our launch in 2013, weve saved millions of consumers more than $140 million! and have received $147 million in funding from investors including Accel, PayPal, Bessemer Venture Partners and New Enterprise Associates. Raise is headquartered in Chicago with an office in New York.

About the Position

Data Preprocessing



- Dealith with missing values
- Duplicated company names and positions

Data Sample

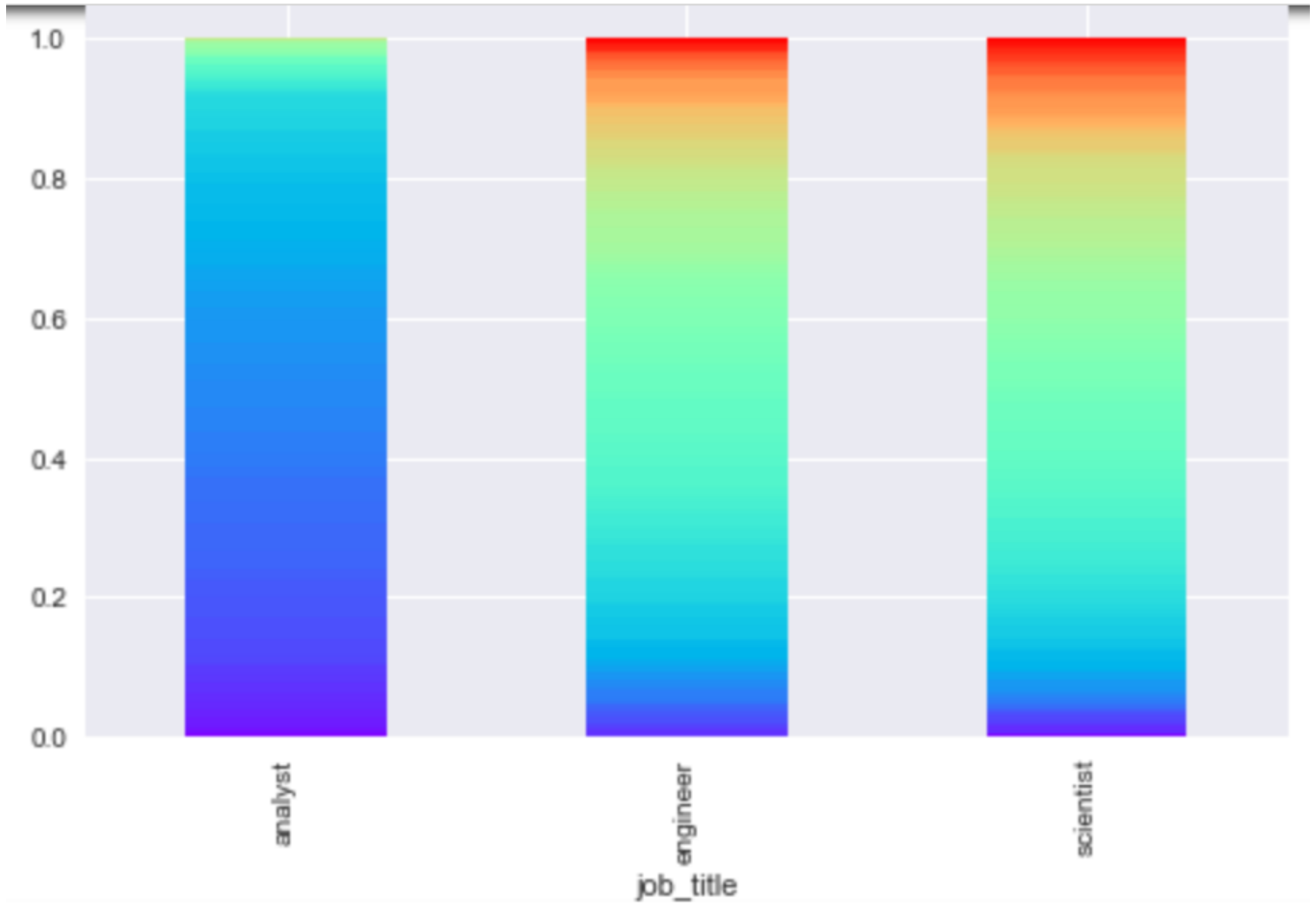
Sample data of Data Scientist:

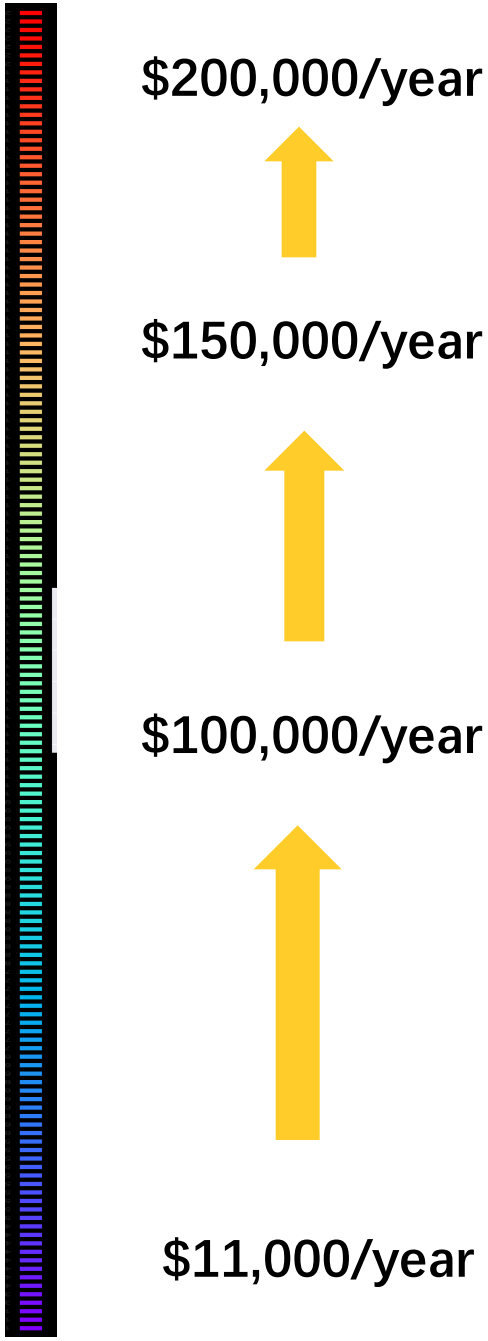
	sci_link	company_name	income	job_desc
0	http://www.glassdoor.com/partner/jobListing.ht...	Grubhub	$121,000/ year	We are looking for a Data Scientist, who will...
1	http://www.glassdoor.com/partner/jobListing.ht...	Newsela	$121,000/ year	Newsela is an Instructional Content Platform ...
2	http://www.glassdoor.com/partner/jobListing.ht...	National Grid USA	$114,000/ year	DescriptionDo you want to change the world? W...
3	http://www.glassdoor.com/partner/jobListing.ht...	Squarespace	$124,000/ year	Marketing Data Scientists formulate the analy...
4	http://www.glassdoor.com/partner/jobListing.ht...	Raise Marketplace Inc.	$84,000/ year	About Raise Were Raise, a leading retail payme...
5	http://www.glassdoor.com/partner/jobListing.ht...	Two Sigma	$164,000/ year	Two Sigma is looking for Data Scientists from...
6	http://www.glassdoor.com/partner/jobListing.ht...	T. Rowe Price	$103,000/ year	Our mission as a leading investment managemen...
7	http://www.glassdoor.com/partner/jobListing.ht...	SeatGeek	$109,000/ year	SeatGeek operates a unique business model in ...
8	http://www.glassdoor.com/partner/jobListing.ht...	None	$134,000/ year	A data scientist at Kensho is passionate about...
9	http://www.glassdoor.com/partner/jobListing.ht...	Intent Media	$153,000/ year	Intent Media isnt your usual company. Our wor...
10	http://www.glassdoor.com/partner/jobListing.ht...	PwC	$122,000/ year	PwC/LOS Overview PwC is a network of firms c...

NO.3

EDA (Exploratory Data Analysis)

Salary Comparison



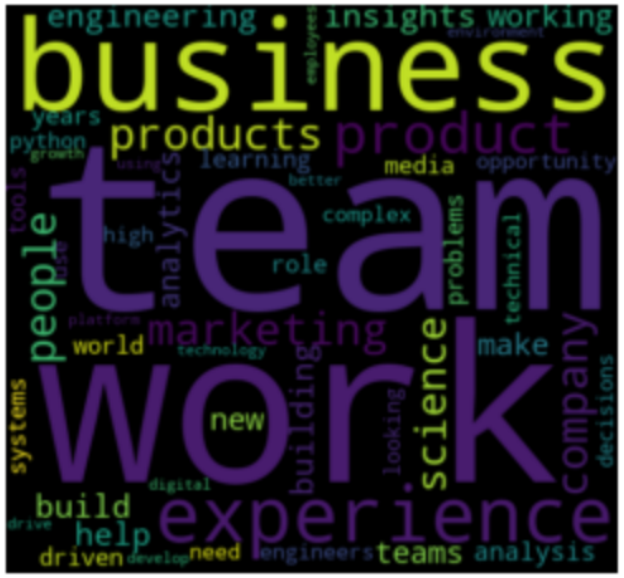


WordCloud for Job Description

Topic: 0

Analytic Skills

Topic: 1



Soft Skills

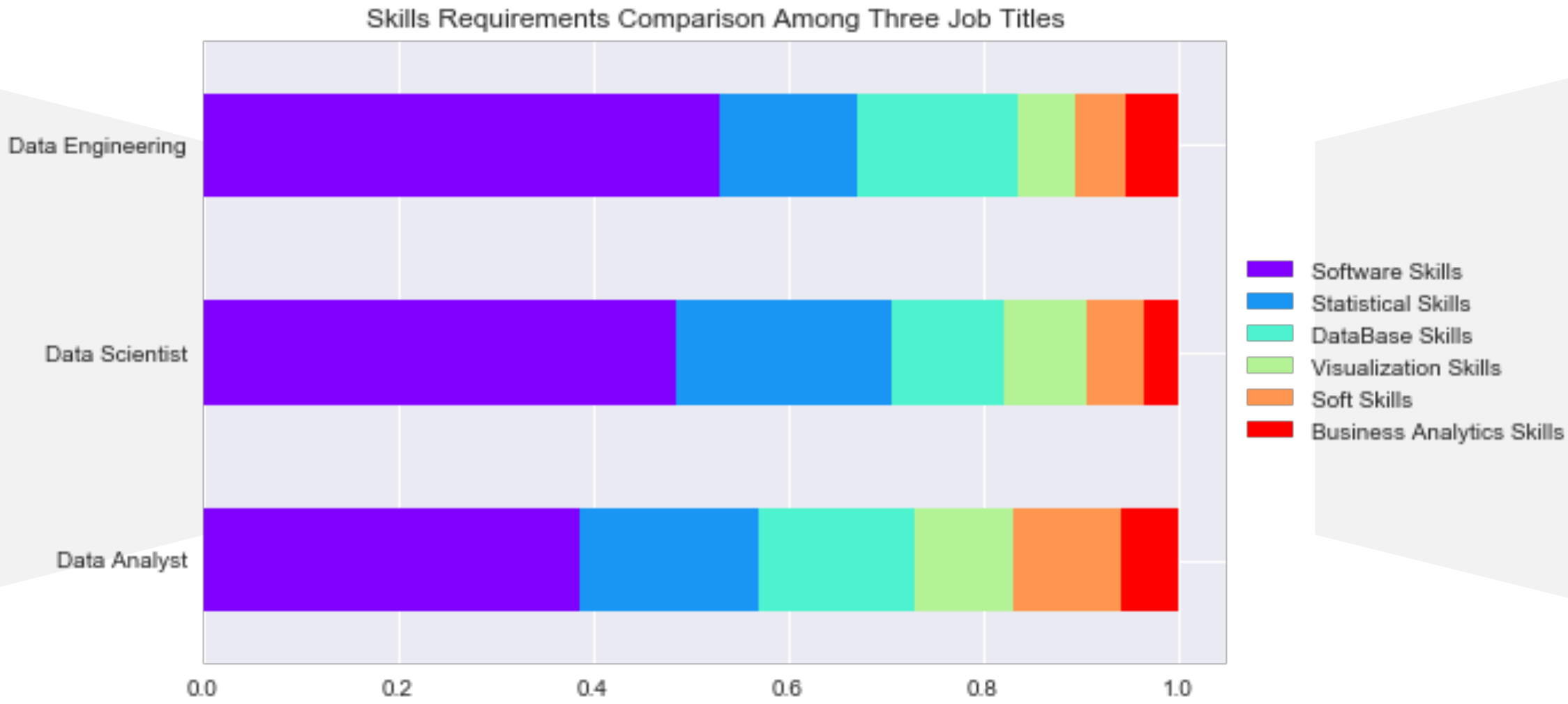
Topic: 2

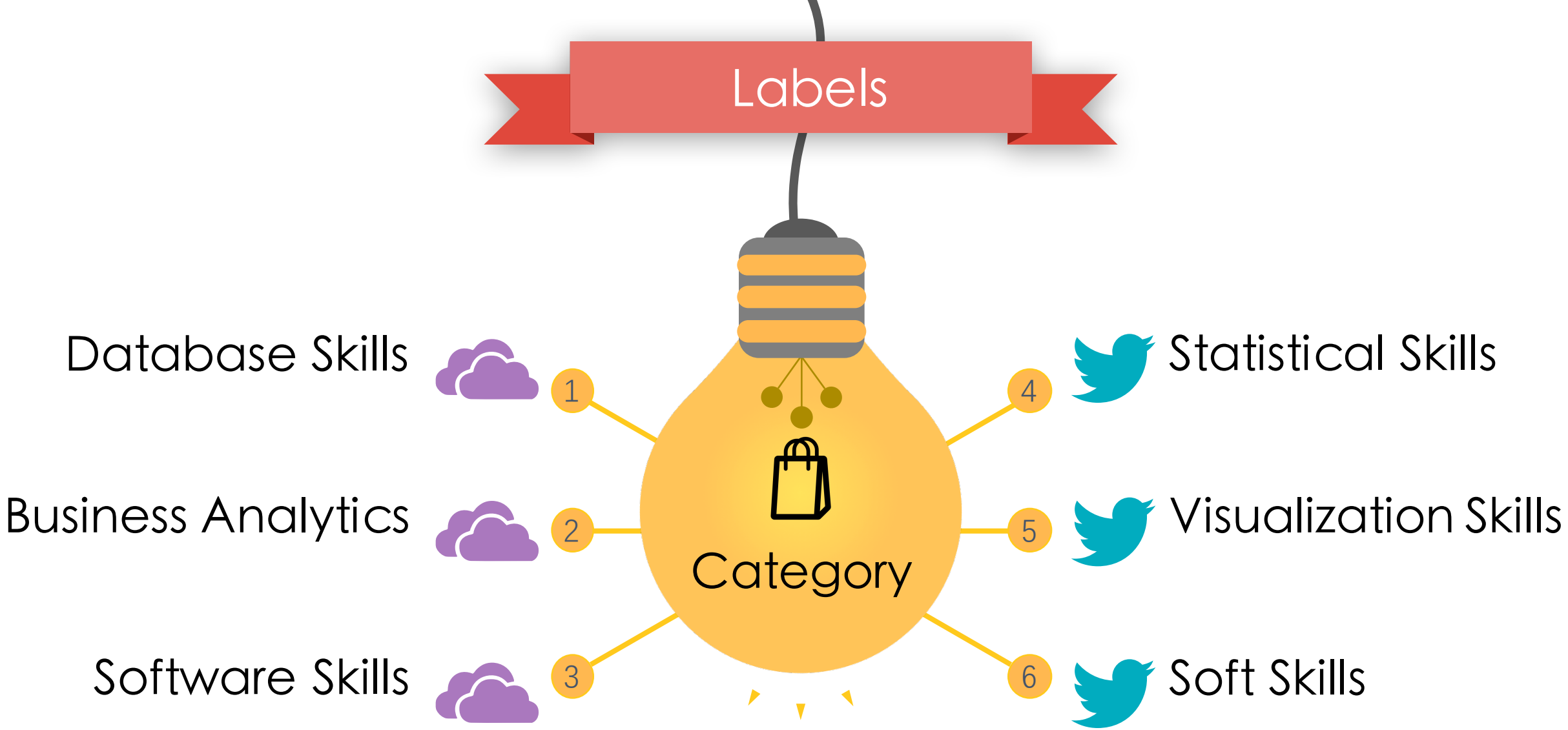
Personal Background

NO.4

Clustering Analysis

Skills Requirements Comparison





The clustering of skills requirement

Labels Sample

Database Skills

- mysql
- microsoft sql server
- oracle essbase
- data warehouse
- data management
- apache cassandra
- apache hadoop
- SAP HANA
- PostgreSQL
- data pipelines

Statistical Skills

- Matlab
- SPSS
- SAS
- JMP
- R
- Linear Regression
- Parameter Estimation
- Hypothesis test
- Bayesian Analysis
- A/B testing
- Bootstrapping

Business Analytics

- Google Analytics
- AWS
- Microsoft Word
- Microsoft Excel
- Google Docs
- IBM Watson Analysis
- Wireframing
- Microsoft Visio
- Microsoft Powerpoint
- Rational Requisite Pro

Labels Sample

Visualization Skills

- Tableau
- Plotly
- Google Charts
- Kartograph
- Timeline
- Data visualization
- JP Graph
- Sigma JS
- Fusion Charts
- Datawrapper

Software Skills

- C
- C++
- java
- Hadoop
- R
- Python
- Spark
- Natural language
- Java script
- Software engineering
- Julia

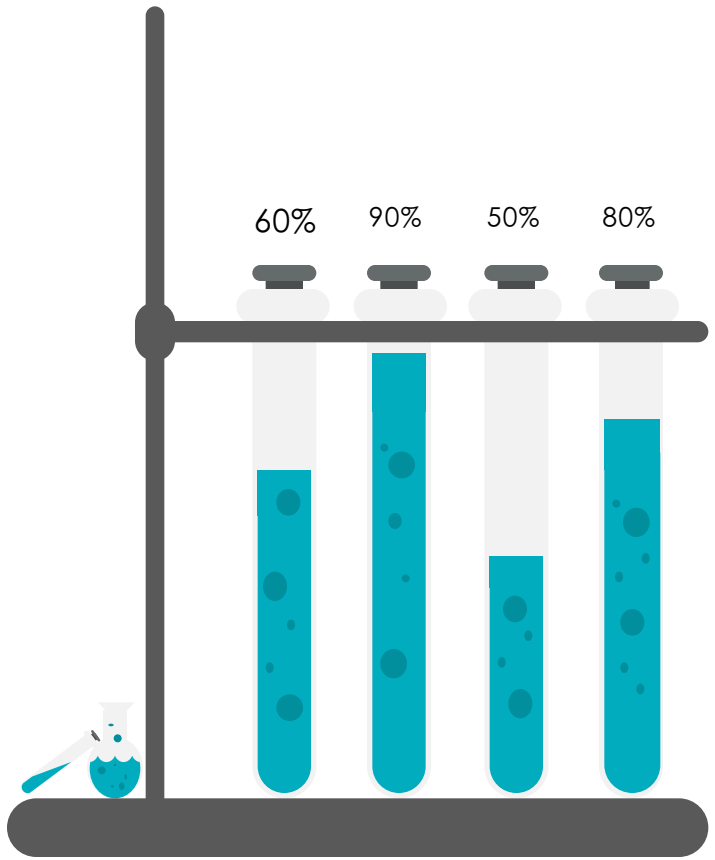
Soft skills

- Communication skills
- Written varbal
- Verbal communication
- Project management
- Problem solving
- Decision msking

NO.5

Classification Analysis

Algorithms



Multi-label Classification

- SVC
- CNN

	precision	recall	f1-score	support
Business	0.70	0.35	0.47	20
DataBase	0.91	0.74	0.82	111
Soft	0.88	0.28	0.42	50
Software	1.00	1.00	1.00	272
Statistical	0.89	0.71	0.79	144
Viz	0.94	0.35	0.51	46
avg / total	0.94	0.77	0.82	643

SVC Performance

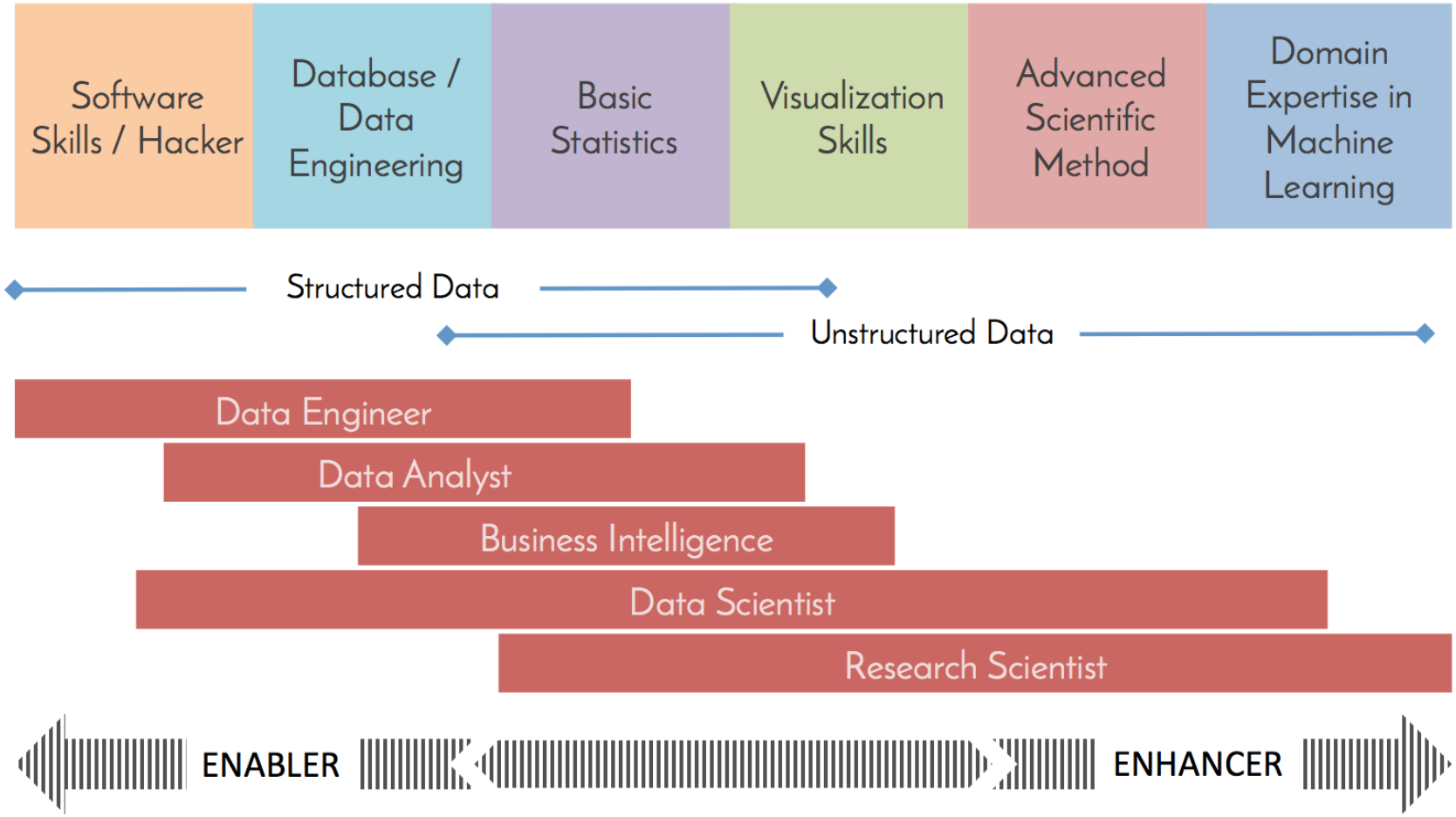
	precision	recall	f1-score	support
Business	0.00	0.00	0.00	14
DataBase	0.85	0.68	0.76	78
Soft	0.75	0.09	0.16	33
Software	1.00	1.00	1.00	181
Statistical	0.99	0.87	0.93	101
Viz	1.00	0.15	0.26	33
avg / total	0.92	0.75	0.79	440

CNN Performance

NO.6

Conclusion & Discussion

Summary


Software Skills / Hacker
Database / Data Engineering
Basic Statistics
Visualization Skills
Advanced Scientific Method
Domain Expertise in Machine Learning
Structured Data
Unstructured Data
Data Engineer
Data Analyst
Business Intelligence
Data Scientist
Research Scientist
ENABLER
ENHANCER

Limitation

More features

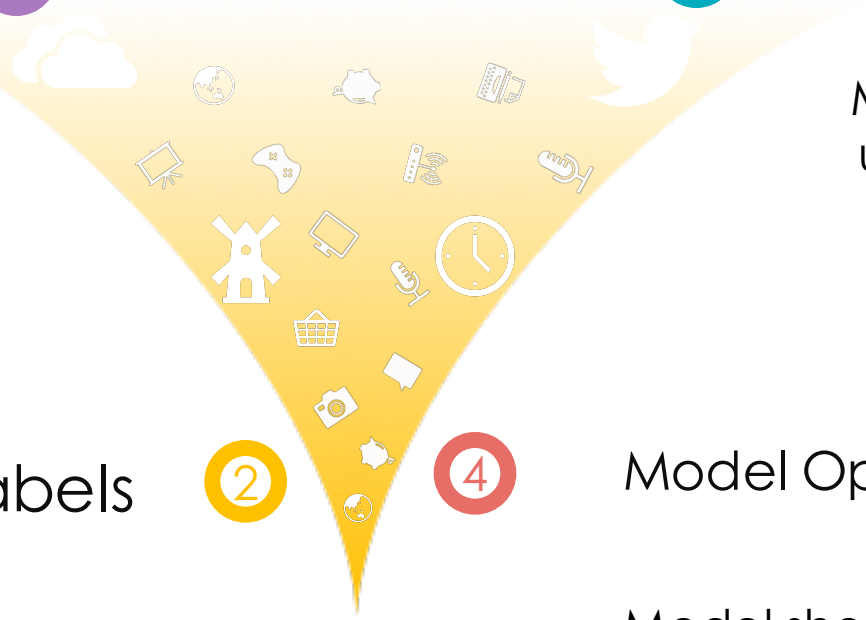
1

Not only analyze job description, company size and other related information should also be considered

Data Set

3

More data can be used for clustering



More detailed labels

2

More detailed skills categories and labels

Model Optimization

4

Model should be optimizaed in order to make a better performance and accuracy

Future Prospects

- Based on the analysis, when someone hope to adjust job, it is easily to explore the skills and knowledge gap among positions
- It can be applied to predict salaries based on the skills you have
- Employer can select resumes that match their reuirements easily

Questions?