Resume Screening

Group 2: Ruiyuan Luan, Sandhya Nallarkkiniyan, Shuang Peng, Chenyi Zhang

TABLE OF CONTENTS

01

OVERVIEW

Why Important?

Resume is the first impression

What are the Goals?

- What job category fits your resume?
- How your resume matches your desired job?

What are the Tools?

TextHero
WordCloud
Lexical diversity
Summarization
Auto Gluon Modeling

Recommendations

ML Models
Text similarity (Cosine)
Dashboard

02

EXPLORATORY DATA ANALYSIS

ABOUT OUR DATASET

2 Columns

Category and Resume

25 Categories

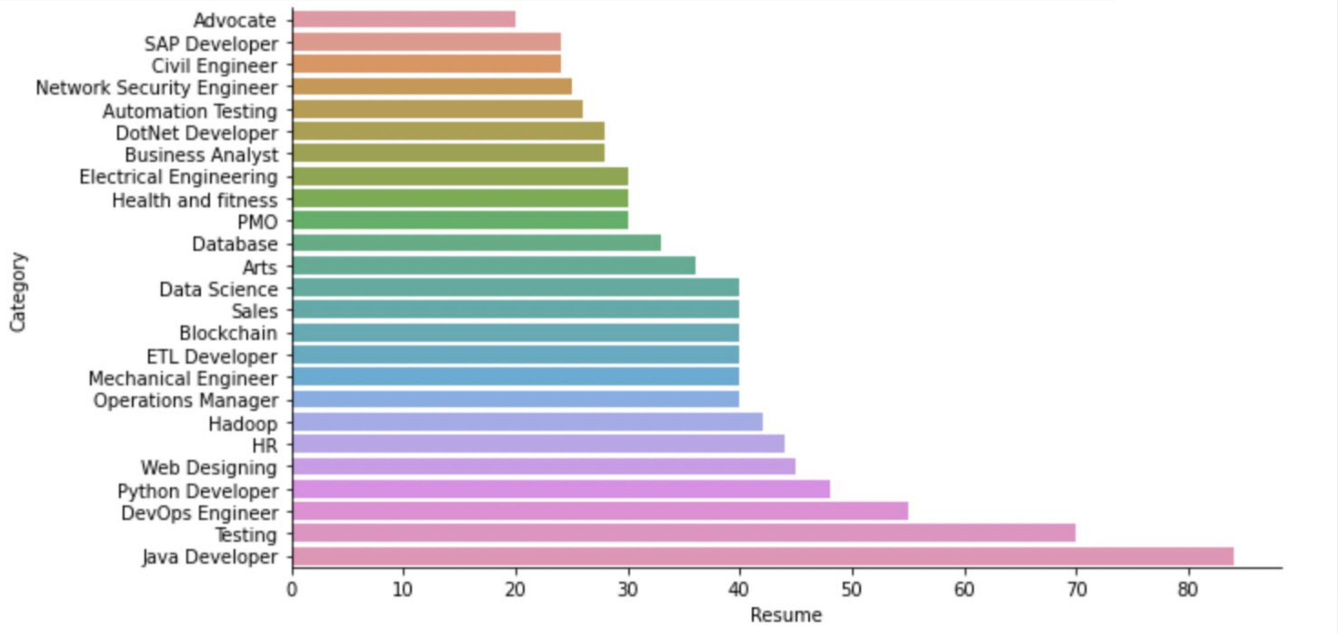
'Data Science', 'HR', 'Web Designing', 'Mechanical Engineer', 'Civil Engineer', 'Java Developer', 'Business Analyst', 'Operations Manager', 'Python Developer'.......

962 Resume

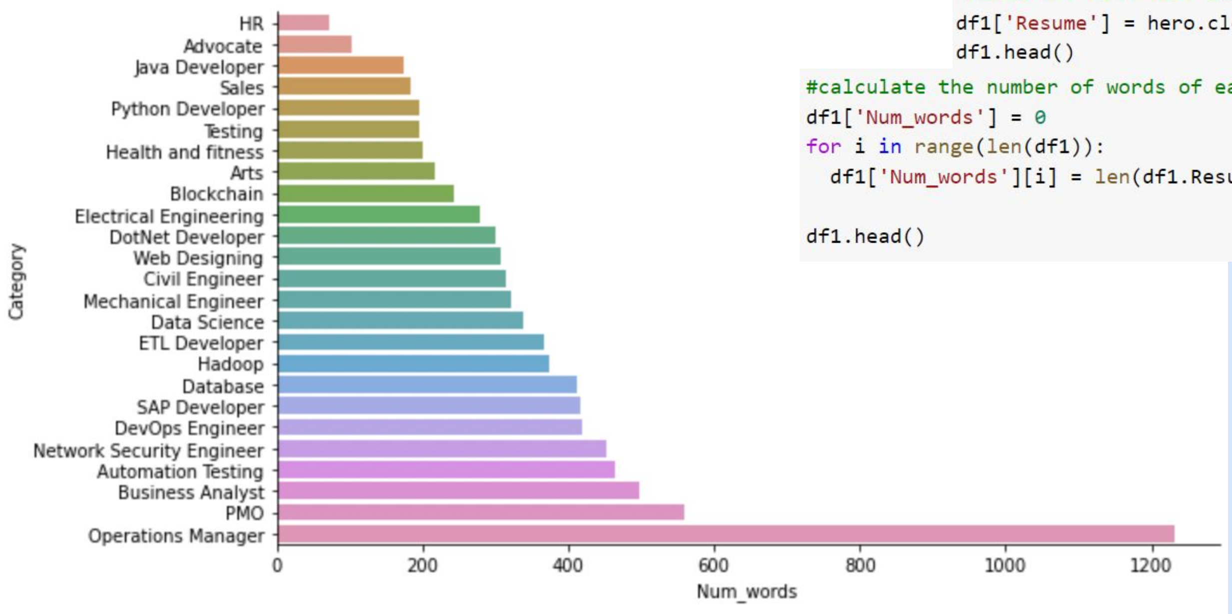
Source: https://www.kaggle.com/code/jillanisofttech/resume-screening-with-knn-ml-99

Category vs. Resume

```
sns.catplot(x='Resume',y='Category',data=df_group,aspect=2, kind = 'bar')
```



Category vs. No. of Words (TextHero)

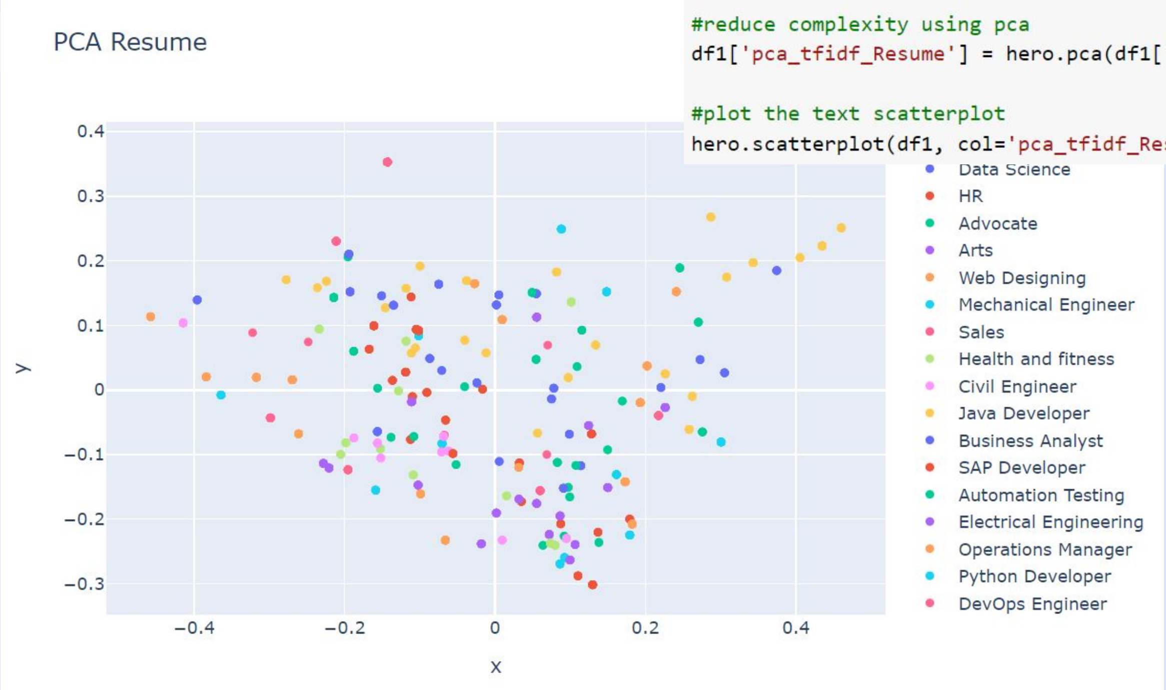



```
#clean the text file using texthero
df1['Resume'] = hero.clean(df1['Resume'])
df1.head()
```

```
#calculate the number of words of each resume
df1['Num_words'] = 0
for i in range(len(df1)):
  df1['Num_words'][i] = len(df1.Resume[i].split(" "))

df1.head()
```

PCA

```
#vectorize the resume column
df1['tfidf_Resume'] = hero.tfidf(df1['Resume'])

#reduce complexity using pca
df1['pca_tfidf_Resume'] = hero.pca(df1['tfidf_Resume'])

#plot the text scatterplot
hero.scatterplot(df1, col='pca_tfidf_Resume', color='Category', title="PCA Resume")
```



Top Words

```python
#most common words of wach category
NUM_TOP_WORDS = 6
df1.groupby('Category')['Resume'].apply(lambda x: hero.top_words(x)[:NUM_TOP_WORDS])
```

```
Data Science          data           396
                      exprience      248
                      months         240
                      aC             236
                      learning       204
                      python         176
```

```
Business Analyst      aC             482
                      business       252
                      test           214
                      description    168
                      company        168
                      |              134
```

```
Java Developer        java           636
                      exprience      408
                      months         372
                      developer      288
                      details        276
                      aC             270
```

```
DevOps Engineer       aC             784
                      exprience      424
                      project        420
                      months         386
                      shell          257
                      team           215
```

Top Words

```
common_words = ['aC/','university','+-','bo','|','test']

for i in range(len(df1)):
  df1['Resume'][i] = df1['Resume'][i].replace('exprience', 'experience')
  for j in common_words:
    df1['Resume'][i] = df1['Resume'][i].replace(j, '')
```

```
DevOps Engineer        experience    444
                       project       420
                       months        386
                       shell         257
                       team          215
                       company       213
```

```
Business Analyst       business      252
                       description   168
                       company       168
                       management    128
                       ing           128
                       project       124
```

```
Data Science           data          396
                       experience    300
                       months        240
                       learning      204
                       python        176
                       year          168
```

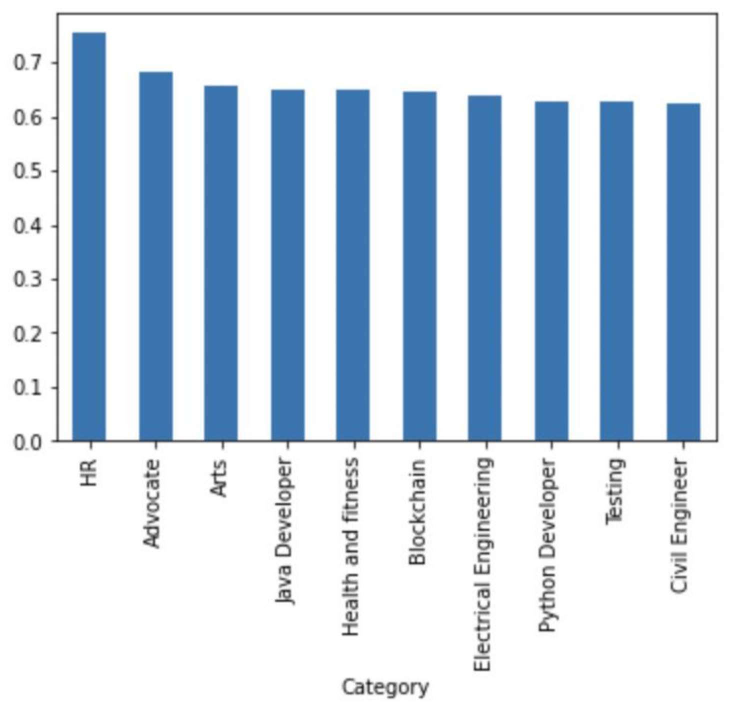
```
Java Developer         java          636
                       experience    426
                       months        372
                       developer     288
                       details       276
                       description   198
```

Word Cloud

```python
text = ''
for i in range(len(df1)):
    text = text + ' ' + df1['Resume'][i]
#number of words in total
len(text.split())
```


```python
from wordcloud import WordCloud
wordcloud = WordCloud().generate(text)

#Use pyplot from matplotlib
figure(figsize=(20,10))
pyplot.imshow(wordcloud, interpolation='bilinear')
pyplot.axis("off")
```

Lexical diversity




```
df1['lex_div'] = 0

for i in range(len(df1)):
  df1['lex_div'][i] = df['Resume'][i].split()
  df1['lex_div'][i] = len(set(df1['lex_div'][i])) / len(df1['lex_div'][i])
```

```
df2 = df1.groupby('Category')['lex_div'].mean()
df2.nlargest(10).plot(kind='bar')
```

Summarization

```
from transformers import pipeline
summarizer = pipeline("summarization")
```

'summary_text': ' Python pandas numpy scipy scikit and matplotlib are programming languages . Java jquery machine learning regression svm naa-ve bayes knn random forest decision trees osting techniques cluster analysis.'

```
txt = df2['Resume'][0]
```

```
if len(txt)>1024:
  txt = txt[:1024]
print(summarizer(txt, max_length=int(len(txt)/4), min_length=25))
```

03

Job Category Recommendations

Auto Gluon Text Classification Models

MODEL 1

Input feature: ***"Resume"***

Train test split

Train Set: 70%

Test Set: 30%

MODEL 2

Additional features:

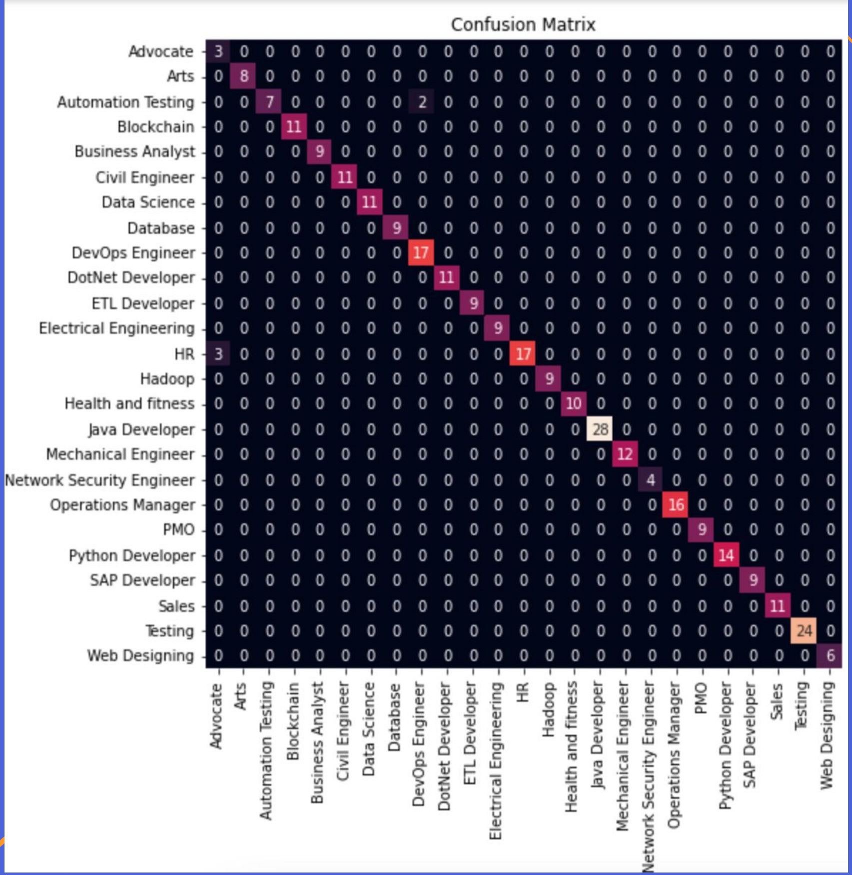
"Num_words", 'lex_div'

Category	Resume
Data Science	skills programming languages python pandas num...
Data Science	education details may may b e uit rgpv data sc...
Data Science	areas interest deep learning control system de...
Data Science	skills aC/ r aC/ python aC/ sap hana aC/ table...
Data Science	education details mca ymcaust faridabad haryan...

Category	Resume	Num_words	lex_div
Data Science	skills programming languages python pandas num...	503	0.637313
Data Science	education details may may b e uit rgpv data sc...	129	0.619632
Data Science	areas interest deep learning control system de...	190	0.641509
Data Science	skills aC/ r aC/ python aC/ sap hana aC/ table...	763	0.476334
Data Science	education details mca ymcaust faridabad haryan...	49	0.42029

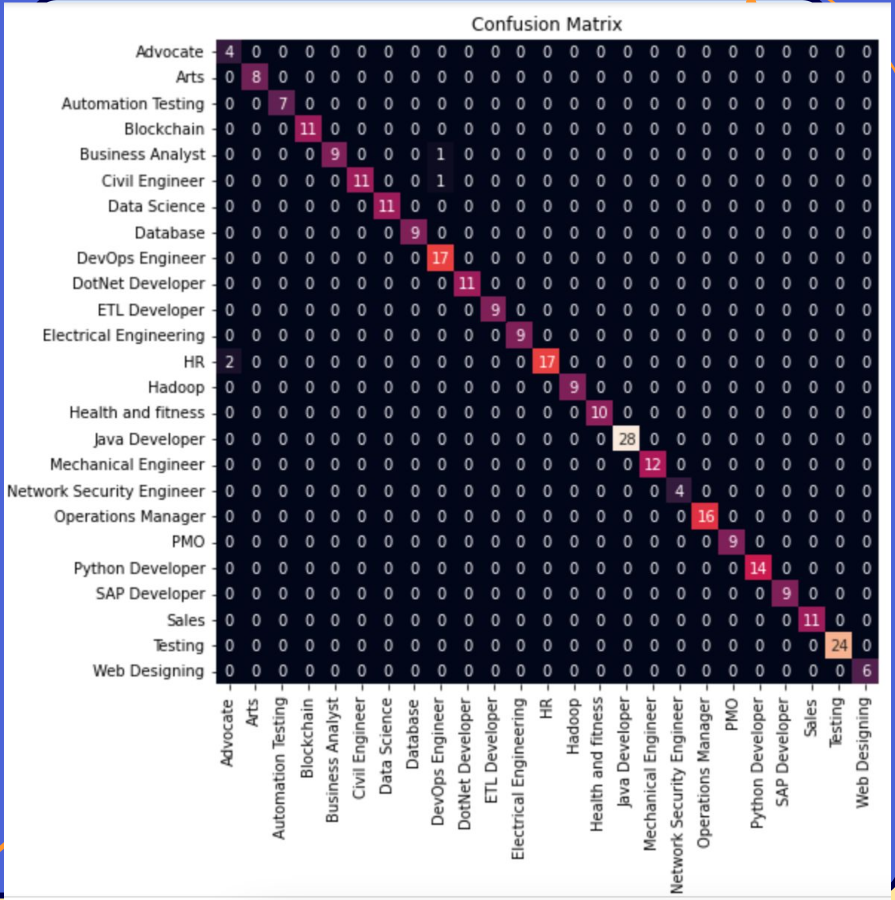
MODEL 1

	precision	recall	f1-score	support
Advocate	1.00	0.50	0.67	6
Arts	1.00	1.00	1.00	8
Automation Testing	0.78	1.00	0.88	7
Blockchain	1.00	1.00	1.00	11
Business Analyst	1.00	1.00	1.00	9
Civil Engineer	1.00	1.00	1.00	11
Data Science	1.00	1.00	1.00	11
Database	1.00	1.00	1.00	9
DevOps Engineer	1.00	0.89	0.94	19
DotNet Developer	1.00	1.00	1.00	11
ETL Developer	1.00	1.00	1.00	9
Electrical Engineering	1.00	1.00	1.00	9
HR	0.85	1.00	0.92	17
Hadoop	1.00	1.00	1.00	9
Health and fitness	1.00	1.00	1.00	10
Java Developer	1.00	1.00	1.00	28
Mechanical Engineer	1.00	1.00	1.00	12
Network Security Engineer	1.00	1.00	1.00	4
Operations Manager	1.00	1.00	1.00	16
PMO	1.00	1.00	1.00	9
Python Developer	1.00	1.00	1.00	14
SAP Developer	1.00	1.00	1.00	9
Sales	1.00	1.00	1.00	11
Testing	1.00	1.00	1.00	24
Web Designing	1.00	1.00	1.00	6
accuracy			0.98	289
macro avg	0.99	0.98	0.98	289
weighted avg	0.99	0.98	0.98	289



MODEL 2

	precision	recall	f1-score	support
Advocate	1.00	0.67	0.80	6
Arts	1.00	1.00	1.00	8
Automation Testing	1.00	1.00	1.00	7
Blockchain	1.00	1.00	1.00	11
Business Analyst	0.90	1.00	0.95	9
Civil Engineer	0.92	1.00	0.96	11
Data Science	1.00	1.00	1.00	11
Database	1.00	1.00	1.00	9
DevOps Engineer	1.00	0.89	0.94	19
DotNet Developer	1.00	1.00	1.00	11
ETL Developer	1.00	1.00	1.00	9
Electrical Engineering	1.00	1.00	1.00	9
HR	0.89	1.00	0.94	17
Hadoop	1.00	1.00	1.00	9
Health and fitness	1.00	1.00	1.00	10
Java Developer	1.00	1.00	1.00	28
Mechanical Engineer	1.00	1.00	1.00	12
Network Security Engineer	1.00	1.00	1.00	4
Operations Manager	1.00	1.00	1.00	16
PMO	1.00	1.00	1.00	9
Python Developer	1.00	1.00	1.00	14
SAP Developer	1.00	1.00	1.00	9
Sales	1.00	1.00	1.00	11
Testing	1.00	1.00	1.00	24
Web Designing	1.00	1.00	1.00	6
accuracy			0.99	289
macro avg	0.99	0.98	0.98	289
weighted avg	0.99	0.99	0.99	289



04

Job Description Matching

Using Text Similarity

MODELING PROCESS

Import Real Job Description for **7 Roles: Data Scientist, Business Analyst, Python Developer, Software Engineer, DevOps Engineer, and ETL Developer**

2 Documents: Job Description and Resume

Import Data — **Clean Data** — **Vectorization** — **Similarity**

Use TextHero

Use **Cosine Similarity** and **Jaccard Similarity**

Working Code

```
job =  df1['Job_Description']
resume = df2['Resume']

columns = list(df_similarity)
data = []

resume_job = []

for r in resume:
    resume_job = []
    resume_job.append(r)
    similarity = []
    for j in job:
        corpus = [r,j]   #Creating corpus of the resume and job to be compared
        trsfm = vectorizer.fit_transform(corpus)  #Vectorizing the text
        cosine = cosine_similarity(trsfm[0:1], trsfm) #Finding Cosine Similarity
        cosine_prcnt = cosine[0,1]*100
        similarity.append(cosine_prcnt)
    resume_job.extend(similarity)
    zipped = zip(columns, resume_job)
    a_dictionary = dict(zipped)
    data.append(a_dictionary)

#Creating the similarity matrix
df_similarity = df_similarity.append(data, True)
#Finding Max Match
df_similarity['Max_Match_Prcnt'] = df_similarity.drop(['Resume'],axis=1).max(axis=1)
df_similarity['Max_Match_Job'] = df_similarity.drop(['Resume'],axis=1).idxmax(axis=1)
```

Output - Cosine Similarity

	Resume	Data Scientist-Google	Data Scientist-Meta	Business Analyst-Amazon	Python Developer-Microsoft	Software Engineer-Apple	DevOps Engineer-Amazon	ETL Developer-Amazon	Max_Match_Prcnt	Max_Match_Job
0	skills programming languages python pandas num...	19.115197	14.651894	12.564862	14.832144	7.887163	5.032347	12.953251	19.115197	Data Scientist-Google
1	education details may may b e uit rgpv data sc...	4.206203	5.376500	4.992131	3.710819	1.329258	1.143470	3.703753	5.376500	Data Scientist-Meta
2	areas interest deep learning control system de...	8.016517	8.978215	7.097594	10.097543	5.893798	3.391123	7.409766	10.097543	Python Developer-Microsoft
3	skills aC/ r aC/ python aC/ sap hana aC/ table...	12.688287	11.118550	14.245513	11.615033	8.014396	5.644845	9.641717	14.245513	Business Analyst-Amazon
4	education details mca ymcaust faridabad haryan...	7.009412	4.833435	5.485900	5.089924	2.928431	0.376121	4.476053	7.009412	Data Scientist-Google

- >15% match makes the resume a very good fit for the job description (61/962 resumes - Top 6%)
- The best resumes in our database had 28% match

05

END USER INTERFACE: DASHBOARD

06

RECOMMENDATIONS & LIMITATIONS

Recommendations

- Perfect tool when finding a specific job title/role
- Useful if you want to know how to edit your resume
- SCU Practicum Program should consider to use this tool
- Can coordinate with employment websites (Indeed/LinkedIn)

Limitations

- Limited dataset
  - Add more publicly available job descriptions and resumes
- Limited advice on resume
- Please share your resumes and the jobs you are interested in to help us improve the model. :-)

Thank You!
Good luck!