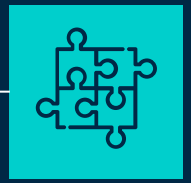
Job-Resume Matching

Final Project
Machine Learning Course

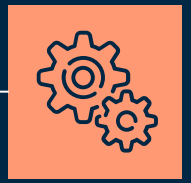
Majdal Hindi

TABLE OF CONTENTS



01

Motivation & Project Objective



02

EDA & Models

03

Limitations & Conclusion

Understanding The Motivation

The biggest issue facing the companies today is how to match the best candidates with job postings by using minimum resources, and in a short amount of time.

There are two major challenges:

•Separating the right candidates from the pack.

•Knowing that candidates can do the job before you hire them.

Project Objective

Finding the best short list of candidates to interview by matching their resumes with the job descriptions.

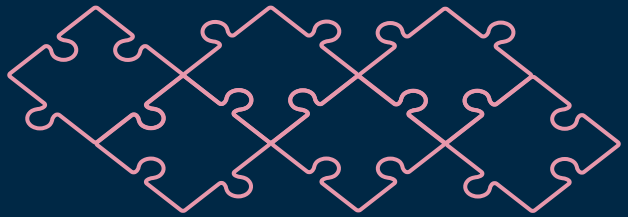
Using NLP Techniques to build Applicant Tracking Systems that could help to make the recruitment process more cost-effective.

ATS prevent filtering out resumes and bias, it ensures that the recruiters can make informed decisions based on subjective data.

Datasets

Resume.csv
From
Kaggle

Jobs.csv
Extracted
from indeed
jobs



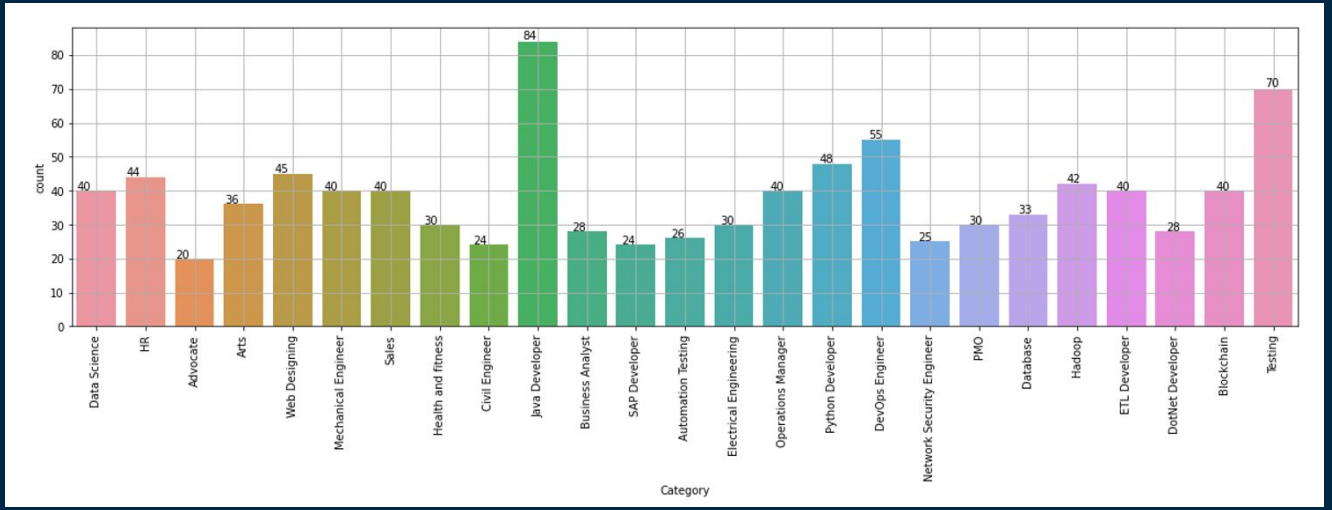
OUR PROCESS


Explore the data
Clean the data
Build Models
Train Job descriptions
Test on Resumes
Cosine Similarity
Limitations & Conclusions

Resume Dataset Exploration

	Applicantion_ID	Category	Resume
0	0	Data Science	Skills * Programming Languages: Python (pandas, numpy, scipy, scikit-learn, matplotlib), Sql, Ja...
1	1	Data Science	Education Details \r\nMay 2013 to May 2017 B.E UIT-RGPV\r\nData Scientist \r\n\r\nData Scienti...
2	2	Data Science	Areas of Interest Deep Learning, Control System Design, Programming in-Python, Electric Machiner...
3	3	Data Science	Skills â□¢ R â□¢ Python â□¢ SAP HANA â□¢ Tableau â□¢ SAP HANA SQL â□¢ SAP HANA PAL â□¢ MS SQL â□...
4	4	Data Science	Education Details \r\n MCA YMCAUST, Faridabad, Haryana\r\nData Science internship \r\n\r\n\r...

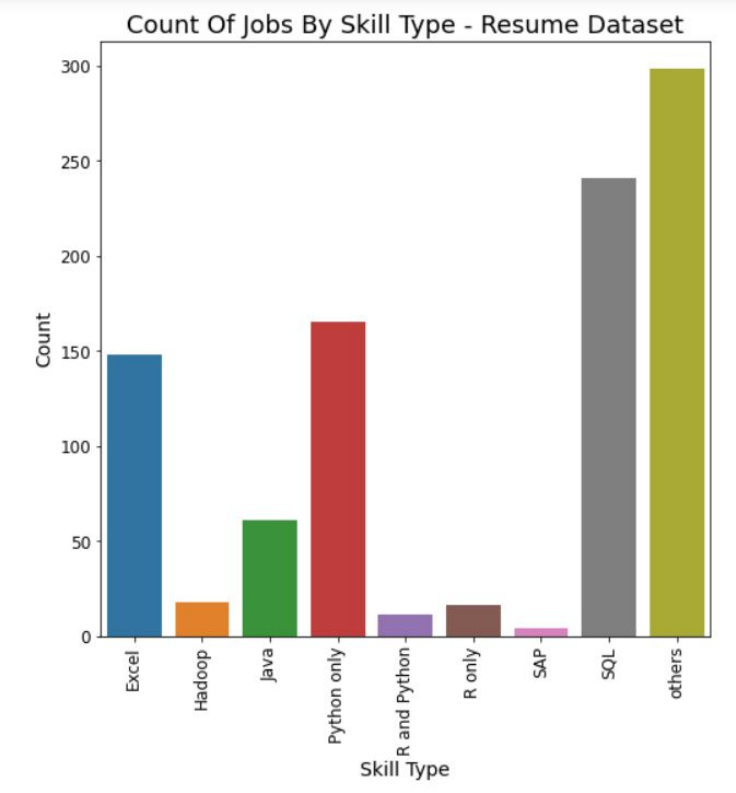
Categories




```
Number of the distinct categories: 25

['Data Science' 'HR' 'Advocate' 'Arts' 'Web Designing'
 'Mechanical Engineer' 'Sales' 'Health and fitness' 'Civil Engineer'
 'Java Developer' 'Business Analyst' 'SAP Developer' 'Automation Testing'
 'Electrical Engineering' 'Operations Manager' 'Python Developer'
 'DevOps Engineer' 'Network Security Engineer' 'PMO' 'Database' 'Hadoop'
 'ETL Developer' 'DotNet Developer' 'Blockchain' 'Testing']
```

Skills Extracted from Resume Docs



Preprocessing Resumes

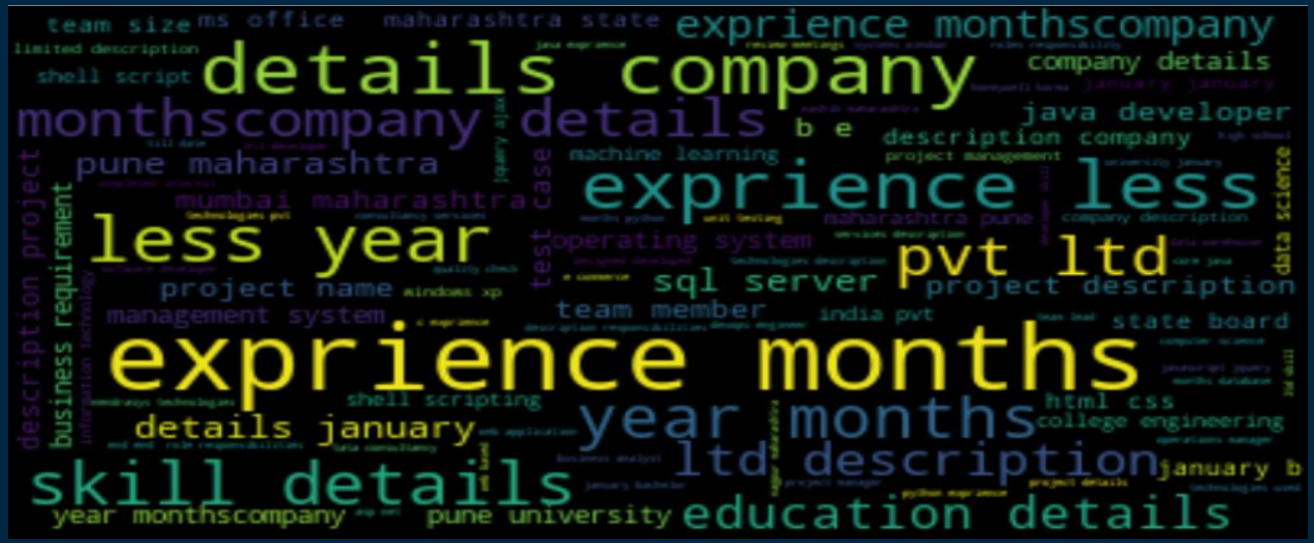
```
# TEXT CLENAING
TEXT_CLEANING_RE = "@\S+|https?:\S+|http?:\S|[^A-Za-z0-9]+"

def preprocess(text, stem=False):
    # Remove link,user and special characters,stop words
    text = re.sub(TEXT_CLEANING_RE, ' ', str(text).lower()).strip()
    tokens = []
    for token in text.split():
        if token not in stop_words:
            if stem:
                tokens.append(stemmer.stem(token))
            else:
                tokens.append(token)
    return " ".join(tokens)
```

	Applicantion_ID	Category	Resume	Cleaned_Resume
0	0	Data Science	Skills * Programming Languages: Python (pandas, numpy, scipy, scikit-learn, matplotlib), Sql, Ja...	skills programming languages python pandas numpy scipy scikit learn matplotlib sql java javascri...
1	1	Data Science	Education Details \r\nMay 2013 to May 2017 B.E UIT-RGPV\r\nData Scientist \r\n\r\nData Scienti...	education details may 2013 may 2017 b e uit rgpv data scientist data scientist matelabs skill de...
2	2	Data Science	Areas of Interest Deep Learning, Control System Design, Programming in-Python, Electric Machiner...	areas interest deep learning control system design programming python electric machinery web dev...
3	3	Data Science	Skills â□¢ R â□¢ Python â□¢ SAP HANA â□¢ Tableau â□¢ SAP HANA SQL â□¢ SAP HANA PAL â□¢ MS SQL â□...	skills r python sap hana tableau sap hana sql sap hana pal ms sql sap lumira c linear programmin...
4	4	Data Science	Education Details \r\n MCA YMCAUST, Faridabad, Haryana\r\nData Science internship \r\n\r\n\r\n\r...	education details mca ymcaust faridabad haryana data science internship skill details data struc...

Most Common Words in Cleaned Resume Docs

NN	140799
NNS	47617
JJ	45825
VBG	20493
CD	15443
VBP	14109
VBD	7332
RB	6869
VBN	5674
VBZ	4350



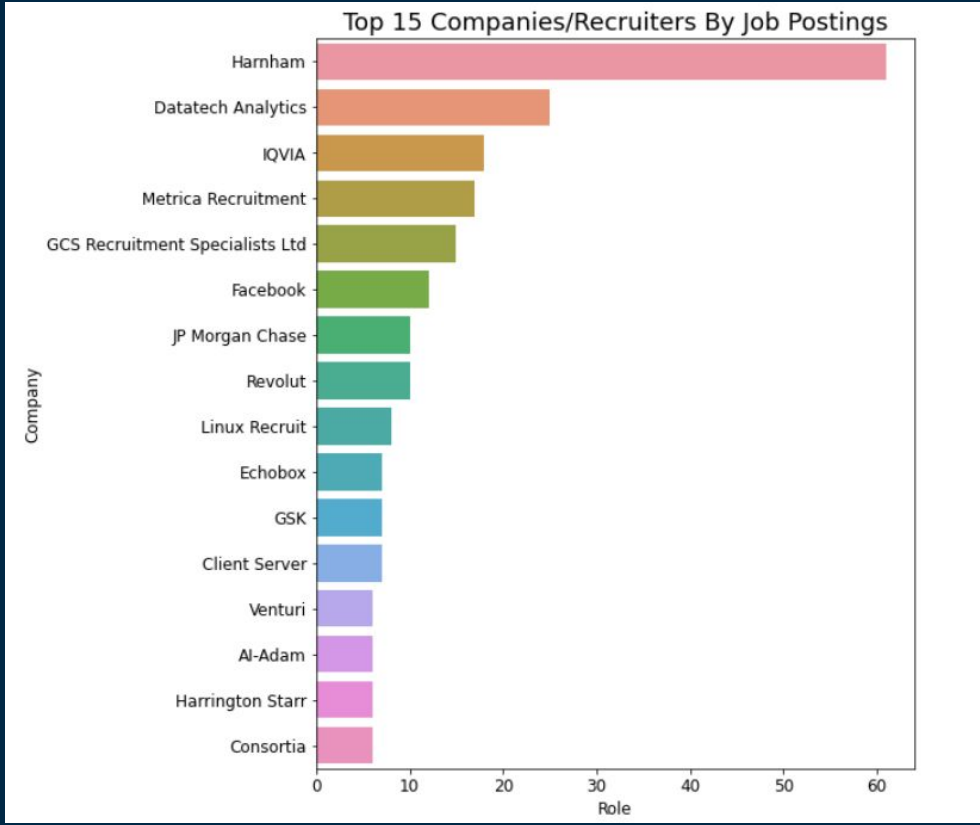
[('exprience', 3829), ('company', 3578), ('project', 3498), ('months', 3288), ('description', 3122), ('details', 3096), ('1', 2162), ('data', 2156), ('management', 1999), ('team', 1950), ('6', 1499), ('maharashtra', 1449), ('system', 1425), ('testing', 1349), ('year', 1336), ('database', 1280), ('development', 1203), ('business', 1196), ('ltd', 1177), ('test', 1174), ('less', 1145), ('using', 1124), ('sql', 1120), ('skill', 1117), ('january', 1090), ('client', 1085), ('java', 1076), ('developer', 1069), ('engineering', 1055), ('application', 1046), ('pune', 1026), ('work', 987), ('services', 956), ('skills', 950), ('c', 910), ('software', 887), ('pvt', 879), ('education', 857), ('responsibilities', 856), ('sales', 825), ('reports', 814), ('process', 813), ('operations', 791), ('requirements', 790), ('2', 776), ('customer', 775), ('server', 773), ('technical', 767), ('technologies', 764), ('india', 762)]

Jobs Dataset Exploration

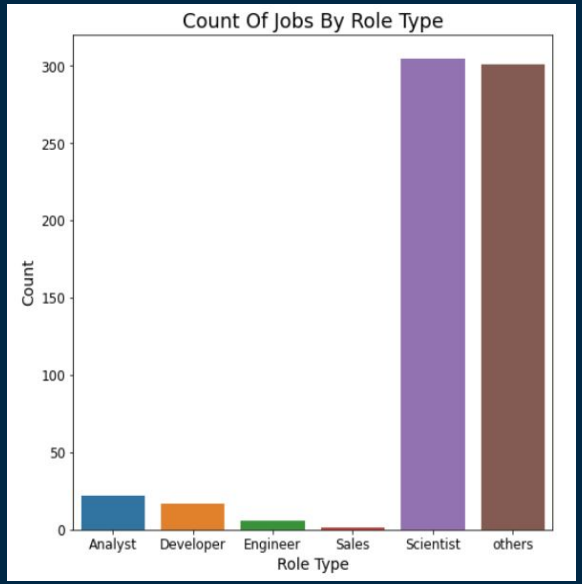
	Job_ID	Role	Company	Location	Contract	Salary	Job_description
0	0	Data Science Consultant - Associate - Artifici...	PwC	London	NaN	NaN	Background PwC are a trusted adviser to some o...
1	1	Data Science Manager	Isharat	London NW1	NaN	NaN	We are looking for a Data Science Manager who ...
2	2	Fraud Analytics Manager	Harnham	London	Permanent	£50,000 - £70,000 a year	FRAUD ANALYTICS MANAGERLONDON - CURRENTLY FLEX...
3	3	CIB Applied AI & Machine Learning □ Economic a...	JP Morgan Chase	London E14	NaN	NaN	CIB Applied AI & Machine Learning □ Economic a...
4	4	Consultant, Data Scientist, Ventures, Tax, London	Deloitte	London	NaN	NaN	Your opportunity Our Tax Ventures team is expa...

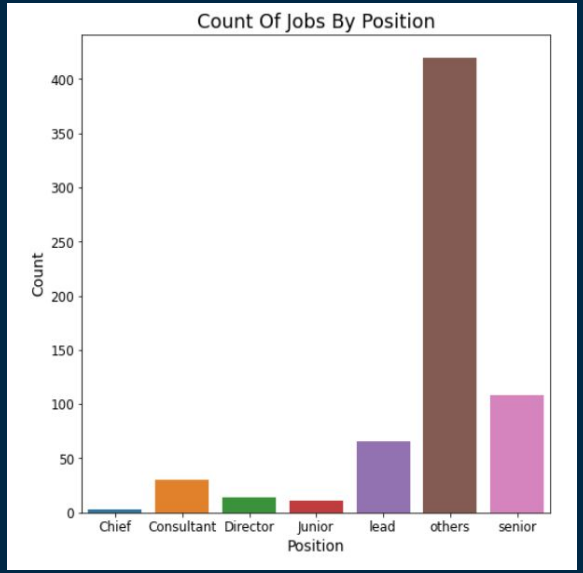
The number of the distinct roles is 455 posted by 58 Companies at 58 Locations in London
The number of (Job_description) documents is 652 consist of 292437 words

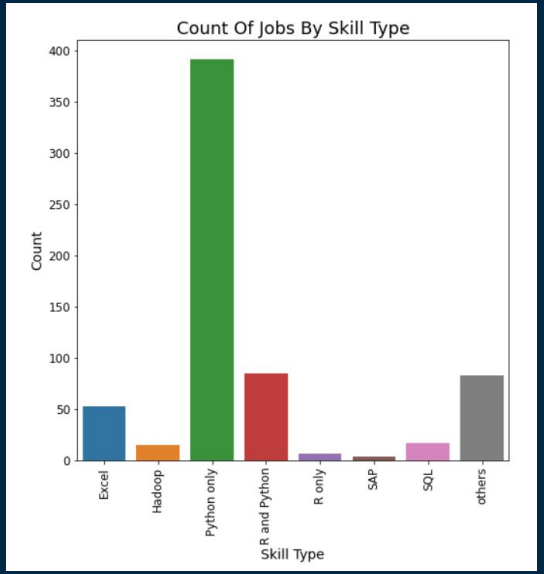
Companies


Top 15 Companies/Recruiters By Job Postings
Harnham
Datatech Analytics
IQVIA
Metrica Recruitment
GCS Recruitment Specialists Ltd
Facebook
JP Morgan Chase
Revolut
Linux Recruit
Echobox
GSK
Client Server
Venturi
Al-Adam
Harrington Starr
Consortia
Company
0
10
20
30
40
50
60
Role

Role Type, Position and Skills Extracted from Job Description docs







Preprocessing Job Description Docs

```
#Before preprocessing:
j_df.loc[2, 'Job_description']
```

```
'FRAUD ANALYTICS MANAGERLONDON - CURRENTLY FLEXIBLE/REMOTE/HOME WORKING UP TO £70,000 + BENEFITS + BONUS An exciting oppor
tunity to join an industry leading team working with data and advanced analytics to develop solutions for some of the hard
est fraud challenges across multiple industries. THE COMPANY A global leading professional services firm with one of the l
argest forensic practices. Having recently been ranked top for using sophisticated analytical and technology driven method
s, this company is leading the way on developing solutions for industry fraud challenges. THE ROLE Design, implement and o
ptimise fraud software solutions Lead multiple workstreams across Fraud Analytics and Data Science Develop business and en
courage the adoption of big data technology and machine learning within clients Support junior team members with ETL proce
sses, data mining and data analytics using SQL Deliver reporting and analytics to senior client stakeholders Respond to RF
Ps and Tenders, as required YOUR SKILLS & EXPERIENCE Strong domain experience within Fraud, Trust & Safety or similar Adva
nced technical skills including hands-on experience with SQL and a programming language such as Python Previous consulting
experience, ideally from Professional Services but can be internal consulting Proven commercial aptitude and ideas on how
develop incremental business within the role Significant experience with third party vendors for fraud detection and preve
ntion BENEFITS Up to £70,000 Competitive bonus Health insurance  Life assurance Pension Flexible working Technical, profes
sional & management training HOW TO APPLY Please submit your CV to Rosalind Madge at Harnham via the Apply Now button. Ple
ase note that our client is currently running a fully remote interview process during the Covid-19 situation.'
```

```
#After preprocessing:
j_df.loc[2, 'Clean_Job_Descriptions']
```

```
'fraud analytics managerlondon currently flexible remote home working 70 000 benefits bonus exciting opportunity join indu
stry leading team working data advanced analytics develop solutions hardest fraud challenges across multiple industries co
mpany global leading professional services firm one largest forensic practices recently ranked top using sophisticated ana
lytical technology driven methods company leading way developing solutions industry fraud challenges role design implement
optimise fraud software solutions lead multiple workstreams across fraud analytics data science develop business encourage
adoption big data technology machine learning within clients support junior team members etl processes data mining data an
alytics using sql deliver reporting analytics senior client stakeholders respond rfps tenders required skills experience s
trong domain experience within fraud trust safety similar advanced technical skills including hands experience sql program
ming language python previous consulting experience ideally professional services internal consulting proven commercial ap
titude ideas develop incremental business within role significant experience third party vendors fraud detection preventio
n benefits 70 000 competitive bonus health insurance life assurance pension flexible working technical professional manage
ment training apply please submit cv rosalind madge harnham via apply button please note client currently running fully re
mote interview process covid 19 situation'
```

Most Common Words in Clean Job Descriptions docs

NN	70675
JJ	34181
NNS	32874
VBG	14033
VBP	11389
RB	6050
CD	3340
VBD	3108
VB	3098
VBN	2933
dtype:	int64


```
[('data', 5357), ('experience', 2401), ('learning', 1836), ('team', 1783), ('science', 1557), ('work', 1475), ('business',
1432), ('machine', 1413), ('working', 1196), ('skills', 1033), ('role', 907), ('analytics', 889), ('solutions', 738), ('de
velopment', 729), ('python', 714), ('new', 693), ('scientist', 650), ('knowledge', 647), ('company', 628), ('strong', 60
9), ('world', 608), ('technical', 600), ('people', 576), ('models', 559), ('clients', 558), ('analysis', 555), ('researc
h', 552), ('product', 551), ('using', 549), ('technology', 545), ('looking', 545), ('ai', 545), ('us', 533), ('help', 53
2), ('across', 523), ('build', 517), ('ability', 506), ('opportunity', 502), ('building', 500), ('projects', 484), ('techn
iques', 483), ('teams', 478), ('including', 469), ('london', 468), ('within', 462), ('engineering', 451), ('software', 45
1), ('tools', 441), ('problems', 427), ('leading', 419)]
```

Models

CountVectorizer Model

01

CountVectotizer Model

By counting words and matching resumes with job descriptions

1. Train the model on job description docs

```
'background pwc trusted adviser largest organisations around world spanning public private financial sectors increasing av
ailability use data requires organisations adapt quickly face irrelevance different technology strategy recognises aims ke
ep us frontier change part strategy currently looking consultants join artificial intelligence ai team london looking self
starters quick learners technology enthusiasts develop build current artificial intelligence data science propositions rol
e business increasingly looking towards data science solution consultant ai team work business alongside data scientists e
ngineers use techniques machine learning natural language processing realise authentic data driven change solutions team r
eports board works clients senior leadership across four business units enhance performance create valuable impact specifi
c responsibilities include limited responsibilities involved design development data science projects pwc clients support
senior management project proposals developing new business opportunities participating constant learning training skills
development grow understanding data science ai concepts contributing strategy growth fast developing data science capabili
ty craft communicate compelling business stories presenting findings senior internal external stakeholders establish effec
tive working relationships directly clients internal pwc teams ambition build data science ai capabilities deliver lasting
impact essential skills able articulate data science concepts business audience work data scientists engineers translate b
usiness requirements solutions possession outstanding oral written communication skills empathic listener persuasive speak
er planning managing executing delivery projects demonstrable understanding data science concepts particularly focused bus
iness need demonstrable understanding statistics excellent understanding machine learning techniques algorithms proficient
understanding data business intelligence tools alteryx vba sql tableau etc understanding agile delivery methods recognise
takes team deliver ai seek individuals demonstrable skills selection number roles consulting experience strategy house big
4 firm house strategy consulting function data driven company industry experience one target sectors e g financial service
s strong academic excellence business analytics related degree e g economics business intelligence analytics exposure clou
d environments azure gcp aws etc exposure programming desirable essential exposure business financial modelling pwc one wo
rld leading professional services organisations 158 countries help clients successful organisations globe well dynamic ent
repreneurs thriving private businesses create value want help measure protect enhance things matter skills look future emp
loyees people need demonstrate skills behaviours support us delivering business strategy important work business clients s
kills behaviours make global leadership framework pwc professional made five core attributes whole leadership technical ca
pabilities business acumen global acumen relationships learn www pwc com uk careers experienced applydiversity work changi
ng world offers great opportunities people diverse backgrounds experiences seek attract employ best people widest talent p
ool well reflect diverse nature society aim encourage culture people valued strengths creating value diversity makes us st
rong business organisation increasingly agile workforce open flexible working arrangements appropriate learn www pwc com u
k diversity'
```

```
matrix([[ 1,  1,  1,  2,  1,  1,  2,  5,  1,  1,  1,  1,  1,  1,  2,  1,
          1,  1,  1,  2,  1,  1,  1,  1,  1,  1,  1,  1,  1,  2,  1,  1,
          1,  2, 16,  1,  2,  1,  1,  2,  1,  5,  1,  2,  1,  1,  1,  1,
          3,  1,  1,  1,  2,  1,  1,  1,  1,  2,  1,  1,  1,  1, 14,  1,
          2,  1,  2,  3,  1,  1,  1,  1,  2,  2,  1,  1,  2,  2,  2,  1,
          1,  1,  1,  1,  1,  1,  2,  2,  1,  1,  1,  2,  1,  1,  1,  1,
          2,  1,  1,  3,  1,  1,  1,  3,  1,  1,  1,  1,  1,  1,  1,  1,
          1,  2,  1,  1,  1,  1,  2,  2,  2,  1,  1,  1,  2,  1,  1,  4,
          2,  1,  1,  1,  1,  1,  1,  3,  1,  2,  1,  3,  1,  1,  1,  1,
          3,  2,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  2,  1,  1,  1,
          1,  2,  1,  1,  4,  1,  1,  1,  4,  1,  1,  1,  1,  1,  1,  2,
          1,  2,  1,  1,  1,  2,  1,  1,  1,  1,  7,  1,  1,  1,  1,  1,
          1,  1,  2,  1,  1,  1,  2,  1,  1,  8,  2,  2,  2,  1,  1,  3,
          2,  7,  1,  1,  2,  1,  1,  1,  1,  1,  1,  1,  1,  6,  1,  2,
          1,  2,  1,  1,  1,  1,  4,  1,  1,  2,  2,  1,  1,  1,  1,  1,
          1,  2,  6,  1,  2,  1,  2,  1,  1,  1,  1,  4,  1,  2,  1,  3,
          1,  2]], dtype=int64)
```

Data Science Consultant - Associate - Artificial Intelligence

2. Convert search resume into a vector using the trained model

'technical skills sql oracle v10 v11 v12 r programming python linear regression machine learning statistical modelling tec hniques obtained certification edvancer eduventures training institute key skills multitasking working meet client sla hig h pressure scenarios handling sensitive clients along improved skills team player excellent communication skills quick lea rner leadership qualities team networking courage take problems proactively education details june 2012 sadvidya pre unive rsity college application database administrator dbms oracle application database administrator dbms oracle ibm india pvt ltd skill details clients exprience 30 months machine learning exprience 30 months oracle exprience 30 months sql exprienc e 30 months excellent communication skills exprience 6 monthscompany details company ibm india pvt ltd description client blue cross blue shield massachusetts health insurance used oracle sql store organize data includes capacity planning insta llation configuration database design migration security troubleshooting backup data recovery worked client databases inst alled oracle v10 v11 v12 linux platform proficient communication clients across locations facilitating data elicitation ha ndling numerous business requests solving diligently within given time frame responding quickly effectively production iss ues within sla leading team co ordination business conduct weekly checkouts database servers systems ibm certifications st atistics 101 applied data science r big data foundations data science foundations business analytics certification pune wo rked retail banking projects design predictive business model using machine learning techniques r programming efficient bu siness marketing strategy'

```
matrix([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 5, 0, 0, 0, 0, 0, 0, 3,
         0, 0, 0, 3, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 6,
         0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 3, 0, 0, 0, 0, 0, 3, 0,
         0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 0, 0, 0, 1, 1,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 5,
         0, 0, 0, 0, 0, 0, 3, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3,
         0, 1, 2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 1, 0, 0, 0, 0]], dtype=int64)
```

```
resume= r_df.loc[719, 'Cleaned_Resume']
```

3. To get a judgment of distances for each Job Description, cosine distances from the resume was calculated

```
similarities = cosine_similarity(matrix,phrases_matrix).flatten()
print("Resume - Job Similarity =",similarities,"by using CountVectorizer Model")
```

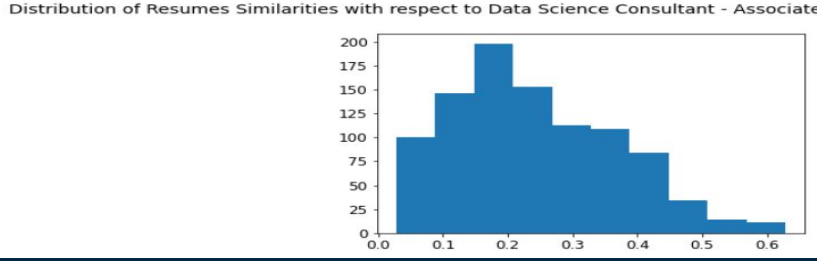
```
Resume - Job Similarity = [0.62872871] by using CountVectorizer Model
```

4. Build Function to find the similarities between specific job description over all the resumes

```
The best 10 applicantions of ' Data Science Consultant - Associate - Artificial Intelligence ' Role at  PwC Company:

      Applicantion_ID  Similarity       Category
719               719    0.628729       Database
730               730    0.628729       Database
741               741    0.628729       Database
9                   9    0.596647   Data Science
19                 19    0.596647   Data Science
29                 29    0.596647   Data Science
39                 39    0.596647   Data Science
6                   6    0.575731   Data Science
16                 16    0.575731   Data Science
26                 26    0.575731   Data Science

Text(0.5, 1.0, 'Distribution of Resumes Similarities with respect to Data Science Consultant - Associate - Artificial Inte
lligence job description\n')
```



TF-IDF Model

02

TF-IDF Model

By calculating how relevant a word in a corpus is to a text. The meaning increases proportionally to the number of times in the text a word appears.

1. Build the model tf-idf pipline

```
#Build The Model
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.preprocessing import Normalizer
from sklearn.pipeline import Pipeline
tf_idf_pipeline = Pipeline([
    ('vect',CountVectorizer(stop_words='english')), # compute bag of words
    ('tfidf',TfidfTransformer()), # use bag of words to compute tf-idf
    ('norm',Normalizer(norm='l2')), # normalize tf-dif vectors for every doc
    ])
#Train The Model
tf_idf = tf_idf_pipeline.fit_transform([jobdesc])
```

2. Train the model on job description docs

'background pwc trusted adviser largest organisations around world spanning public private financial sectors increasing av ailability use data requires organisations adapt quickly face irrelevance different technology strategy recognises aims ke ep us frontier change part strategy currently looking consultants join artificial intelligence ai team london looking self starters quick learners technology enthusiasts develop build current artificial intelligence data science propositions rol e business increasingly looking towards data science solution consultant ai team work business alongside data scientists e ngineers use techniques machine learning natural language processing realise authentic data driven change solutions team r eports board works clients senior leadership across four business units enhance performance create valuable impact specifi c responsibilities include limited responsibilities involved design development data science projects pwc clients support senior management project proposals developing new business opportunities participating constant learning training skills development grow understanding data science ai concepts contributing strategy growth fast developing data science capabili ty craft communicate compelling business stories presenting findings senior internal external stakeholders establish effec tive working relationships directly clients internal pwc teams ambition build data science ai capabilities deliver lasting impact essential skills able articulate data science concepts business audience work data scientists engineers translate b usiness requirements solutions possession outstanding oral written communication skills empathic listener persuasive speak er planning managing executing delivery projects demonstrable understanding data science concepts particularly focused bus iness need demonstrable understanding statistics excellent understanding machine learning techniques algorithms proficient understanding data business intelligence tools alteryx vba sql tableau etc understanding agile delivery methods recognise takes team deliver ai seek individuals demonstrable skills selection number roles consulting experience strategy house big 4 firm house strategy consulting function data driven company industry experience one target sectors e g financial service s strong academic excellence business analytics related degree e g economics business intelligence analytics exposure clou d environments azure gcp aws etc exposure programming desirable essential exposure business financial modelling pwc one wo rld leading professional services organisations 158 countries help clients successful organisations globe well dynamic ent repreneurs thriving private businesses create value want help measure protect enhance things matter skills look future emp loyees people need demonstrate skills behaviours support us delivering business strategy important work business clients s kills behaviours make global leadership framework pwc professional made five core attributes whole leadership technical ca pabilities business acumen global acumen relationships learn www pwc com uk careers experienced applydiversity work changi ng world offers great opportunities people diverse backgrounds experiences seek attract employ best people widest talent p ool well reflect diverse nature society aim encourage culture people valued strengths creating value diversity makes us st rong business organisation increasingly agile workforce open flexible working arrangements appropriate learn www pwc com u k diversity'

```
matrix([[ 1,  1,  1,  2,  1,  1,  2,  5,  1,  1,  1,  1,  1,  1,  2,  1,
          1,  1,  1,  2,  1,  1,  1,  1,  1,  1,  1,  1,  1,  2,  1,  1,
          1,  2, 16,  1,  2,  1,  1,  2,  1,  5,  1,  2,  1,  1,  1,  1,
          3,  1,  1,  1,  2,  1,  1,  1,  1,  2,  1,  1,  1,  1, 14,  1,
          2,  1,  2,  3,  1,  1,  1,  1,  2,  2,  1,  1,  2,  2,  2,  1,
          1,  1,  1,  1,  1,  1,  2,  2,  1,  1,  1,  2,  1,  1,  1,  1,
          2,  1,  1,  3,  1,  1,  1,  3,  1,  1,  1,  1,  1,  1,  1,  1,
          1,  2,  1,  1,  1,  1,  2,  2,  2,  1,  1,  1,  2,  1,  1,  4,
          2,  1,  1,  1,  1,  1,  1,  3,  1,  2,  1,  3,  1,  1,  1,  1,
          3,  2,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  2,  1,  1,  1,
          1,  2,  1,  1,  4,  1,  1,  1,  4,  1,  1,  1,  1,  1,  1,  2,
          1,  2,  1,  1,  1,  2,  1,  1,  1,  1,  7,  1,  1,  1,  1,  1,
          1,  1,  2,  1,  1,  1,  2,  1,  1,  8,  2,  2,  2,  1,  1,  3,
          2,  7,  1,  1,  2,  1,  1,  1,  1,  1,  1,  1,  1,  6,  1,  2,
          1,  2,  1,  1,  1,  1,  4,  1,  1,  2,  2,  1,  1,  1,  1,  1,
          1,  2,  6,  1,  2,  1,  2,  1,  1,  1,  1,  4,  1,  2,  1,  3,
          1,  2]], dtype=int64)
```

Data Science Consultant - Associate - Artificial Intelligence

Vocab of job description No.0

	vocab	frequency	document_frequency
0	158	1	1
1	able	1	1
2	academic	1	1
3	acumen	2	1
4	adapt	1	1
...	...	...	...
253	working	2	1
254	works	1	1
255	world	3	1
256	written	1	1
257	www	2	1

258 rows × 3 columns

3. Convert search resume into a vector using the trained model

'technical skills sql oracle v10 v11 v12 r programming python linear regression machine learning statistical modelling tec hniques obtained certification edvancer eduventures training institute key skills multitasking working meet client sla hig h pressure scenarios handling sensitive clients along improved skills team player excellent communication skills quick lea rner leadership qualities team networking courage take problems proactively education details june 2012 sadvidya pre unive rsity college application database administrator dbms oracle application database administrator dbms oracle ibm india pvt ltd skill details clients exprience 30 months machine learning exprience 30 months oracle exprience 30 months sql exprienc e 30 months excellent communication skills exprience 6 monthscompany details company ibm india pvt ltd description client blue cross blue shield massachusetts health insurance used oracle sql store organize data includes capacity planning insta llation configuration database design migration security troubleshooting backup data recovery worked client databases inst alled oracle v10 v11 v12 linux platform proficient communication clients across locations facilitating data elicitation ha ndling numerous business requests solving diligently within given time frame responding quickly effectively production iss ues within sla leading team co ordination business conduct weekly checkouts database servers systems ibm certifications st atistics 101 applied data science r big data foundations data science foundations business analytics certification pune wo rked retail banking projects design predictive business model using machine learning techniques r programming efficient bu siness marketing strategy'

```
resume= r_df.loc[719, 'Cleaned_Resume']
```

```
matrix([[0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.07537784,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.07537784, 0.        , 0.        , 0.37688918,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.22613351, 0.        , 0.        , 0.        ,
         0.22613351, 0.07537784, 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.45226702, 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.15075567,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.15075567,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.07537784, 0.07537784, 0.        , 0.        , 0.22613351,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.22613351, 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.07537784, 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
         0.        , 0.        , 0.        , 0.        , 0.        ,
```

4. Find the similarity between job description No.0 and Resume No.719

```
similarities = cosine_similarity(tf_idf,phrases_matrix).flatten()
print("Resume - Job Similarity =",similarities,"by using TF_IDF Pipline Model")
```

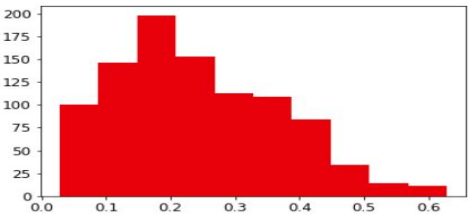
```
Resume - Job Similarity = [0.62872871] by using TF_IDF Pipline Model
```

5. Build Function to find the similarities between specific job description over all the resumes

```
The best 10 applicantions of ' Data Science Consultant - Associate - Artificial Intelligence ' Role at  PwC Company:
```

	Applicantion_ID	Similarity	Category
719	719	0.628729	Database
730	730	0.628729	Database
741	741	0.628729	Database
9	9	0.596647	Data Science
19	19	0.596647	Data Science
29	29	0.596647	Data Science
39	39	0.596647	Data Science
6	6	0.575731	Data Science
16	16	0.575731	Data Science
26	26	0.575731	Data Science

```
Text(0.5, 1.0, 'Distribution of Resumes Similarities with respect to Data Science Consultant - Associate - Artificial Intelligence job description\n')
```



LDA

03

Doc2bow to create the corpus

LDA

By building topics per document

1. Creating Corpus by CountVectorizer

'background pwc trusted adviser largest organisations around world spanning public private financial sectors increasing av ailability use data requires organisations adapt quickly face irrelevance different technology strategy recognises aims ke ep us frontier change part strategy currently looking consultants join artificial intelligence ai team london looking self starters quick learners technology enthusiasts develop build current artificial intelligence data science propositions rol e business increasingly looking towards data science solution consultant ai team work business alongside data scientists e ngineers use techniques machine learning natural language processing realise authentic data driven change solutions team r eports board works clients senior leadership across four business units enhance performance create valuable impact specifi c responsibilities include limited responsibilities involved design development data science projects pwc clients support senior management project proposals developing new business opportunities participating constant learning training skills development grow understanding data science ai concepts contributing strategy growth fast developing data science capabili ty craft communicate compelling business stories presenting findings senior internal external stakeholders establish effec tive working relationships directly clients internal pwc teams ambition build data science ai capabilities deliver lasting impact essential skills able articulate data science concepts business audience work data scientists engineers translate b usiness requirements solutions possession outstanding oral written communication skills empathic listener persuasive speak er planning managing executing delivery projects demonstrable understanding data science concepts particularly focused bus iness need demonstrable understanding statistics excellent understanding machine learning techniques algorithms proficient understanding data business intelligence tools alteryx vba sql tableau etc understanding agile delivery methods recognise takes team deliver ai seek individuals demonstrable skills selection number roles consulting experience strategy house big 4 firm house strategy consulting function data driven company industry experience one target sectors e g financial service s strong academic excellence business analytics related degree e g economics business intelligence analytics exposure clou d environments azure gcp aws etc exposure programming desirable essential exposure business financial modelling pwc one wo rld leading professional services organisations 158 countries help clients successful organisations globe well dynamic ent repreneurs thriving private businesses create value want help measure protect enhance things matter skills look future emp loyees people need demonstrate skills behaviours support us delivering business strategy important work business clients s kills behaviours make global leadership framework pwc professional made five core attributes whole leadership technical ca pabilities business acumen global acumen relationships learn www pwc com uk careers experienced applydiversity work changi ng world offers great opportunities people diverse backgrounds experiences seek attract employ best people widest talent p ool well reflect diverse nature society aim encourage culture people valued strengths creating value diversity makes us st rong business organisation increasingly agile workforce open flexible working arrangements appropriate learn www pwc com u k diversity'

Data Science Consultant - Associate - Artificial Intelligence

```
matrix([[ 1,  1,  1,  2,  1,  1,  2,  5,  1,  1,  1,  1,  1,  1,  2,  1,
          1,  1,  1,  2,  1,  1,  1,  1,  1,  1,  1,  1,  1,  2,  1,  1,
          1,  2, 16,  1,  2,  1,  1,  2,  1,  5,  1,  2,  1,  1,  1,  1,
          3,  1,  1,  1,  2,  1,  1,  1,  1,  2,  1,  1,  1,  1, 14,  1,
          2,  1,  2,  3,  1,  1,  1,  1,  2,  2,  1,  1,  2,  2,  2,  1,
          1,  1,  1,  1,  1,  1,  2,  2,  1,  1,  1,  2,  1,  1,  1,  1,
          2,  1,  1,  3,  1,  1,  1,  3,  1,  1,  1,  1,  1,  1,  1,  1,
          1,  2,  1,  1,  1,  1,  2,  2,  2,  1,  1,  1,  2,  1,  1,  4,
          2,  1,  1,  1,  1,  1,  1,  3,  1,  2,  1,  3,  1,  1,  1,  1,
          3,  2,  1,  1,  1,  1,  1,  1,  1,  1,  1,  1,  2,  1,  1,  1,
          1,  2,  1,  1,  4,  1,  1,  1,  4,  1,  1,  1,  1,  1,  1,  2,
          1,  2,  1,  1,  1,  2,  1,  1,  1,  1,  7,  1,  1,  1,  1,  1,
          1,  1,  2,  1,  1,  1,  2,  1,  1,  8,  2,  2,  2,  1,  1,  3,
          2,  7,  1,  1,  2,  1,  1,  1,  1,  1,  1,  1,  1,  6,  1,  2,
          1,  2,  1,  1,  1,  1,  4,  1,  1,  2,  2,  1,  1,  1,  1,  1,
          1,  2,  6,  1,  2,  1,  2,  1,  1,  1,  1,  4,  1,  2,  1,  3,
          1,  2]], dtype=int64)
```

```
array([[9.98324404e-01, 8.37797979e-04, 8.37797975e-04]])
```

2. Train the model on job description docs

```
# Build and train LDA on job desc doc
lda_model = LatentDirichletAllocation(n_components=3, max_iter=3, n_jobs = -1, verbose=1)
document_topics= lda_model.fit_transform(corpus)
```

```
iteration: 1 of max_iter: 3
iteration: 2 of max_iter: 3
iteration: 3 of max_iter: 3
```

Topics of job description No.0

```
topics = print_LDA_topics(lda_model.components_, vec.get_feature_names())
```

```
Topic 0: possession authentic tableau directly algorithms new written presenting design applydiversity make increasing tec
hnical proposals
Topic 1: possession gcp statistics design directly new managing valued listener largest function propositions number frame
work
Topic 2: business data science pwc skills strategy understanding clients ai intelligence work organisations team people
```

3. Convert search resume into a vector using the trained models

'technical skills sql oracle v10 v11 v12 r programming python linear regression machine learning statistical modelling tec hniques obtained certification advancer eduventures training institute key skills multitasking working meet client sla hig h pressure scenarios handling sensitive clients along improved skills team player excellent communication skills quick lea rner leadership qualities team networking courage take problems proactively education details june 2012 sadvidya pre unive rsity college application database administrator dbms oracle application database administrator dbms oracle ibm india pvt ltd skill details clients exprience 30 months machine learning exprience 30 months oracle exprience 30 months sql exprienc e 30 months excellent communication skills exprience 6 monthscompany details company ibm india pvt ltd description client blue cross blue shield massachusetts health insurance used oracle sql store organize data includes capacity planning insta llation configuration database design migration security troubleshooting backup data recovery worked client databases inst alled oracle v10 v11 v12 linux platform proficient communication clients across locations facilitating data elicitation ha ndling numerous business requests solving diligently within given time frame responding quickly effectively production iss ues within sla leading team co ordination business conduct weekly checkouts database servers systems ibm certifications st atistics 101 applied data science r big data foundations data science foundations business analytics certification pune wo rked retail banking projects design predictive business model using machine learning techniques r programming efficient bu siness marketing strategy'

```
resume= r_df.loc[719, 'Cleaned_Resume']
```

```
matrix([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 5, 0, 0, 0, 0, 0, 0, 3,
         0, 0, 0, 3, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 6,
         0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 3, 0, 0, 0, 0, 0, 3, 0,
         0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 0, 0, 0, 1, 1,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 5,
         0, 0, 0, 0, 0, 0, 3, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 3,
         0, 1, 2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 1, 0, 0, 0, 0]], dtype=int64)
```

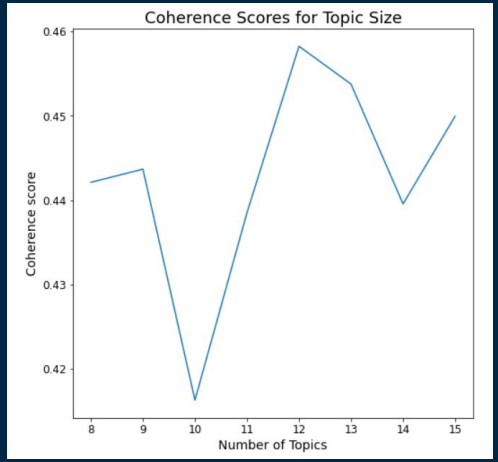
```
array([[0.98884232, 0.00557884, 0.00557884]])
```

4. Find the similarity between job description No.0 and Resume No.719

```
similarities=cosine_similarity(document_topics,current_document_topics ).flatten()
print("Resume - Job Similarity =",similarities,"by LDA Model")
```

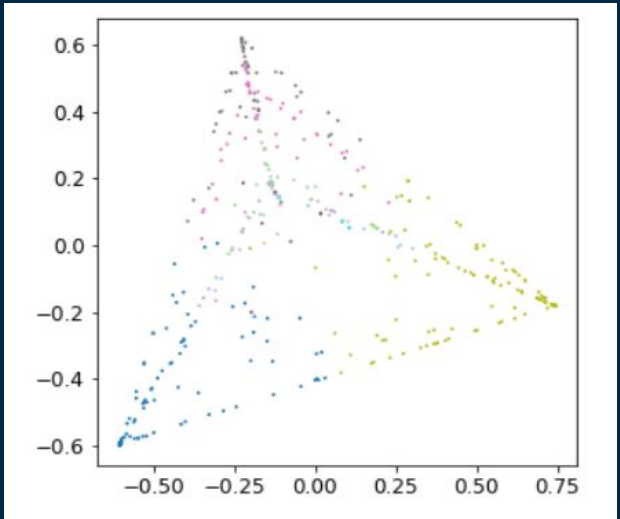
```
Resume - Job Similarity = [0.99997694] by LDA Model
```

5. Choosing the optimal number of topics by using coherence value

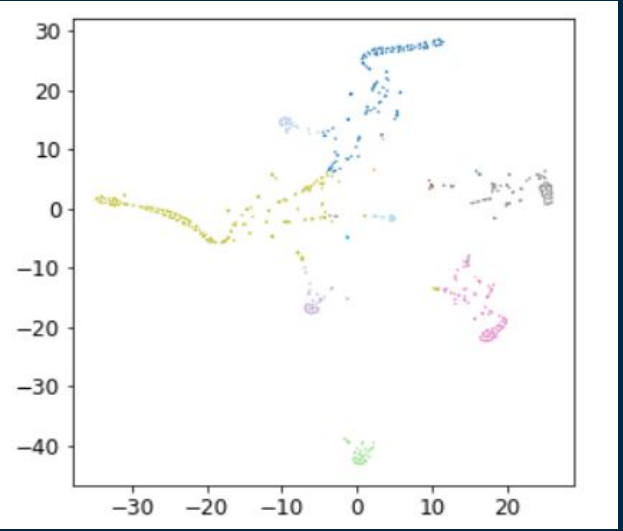


Display Summary of all the Topics of all Job descriptions Postings

PCA



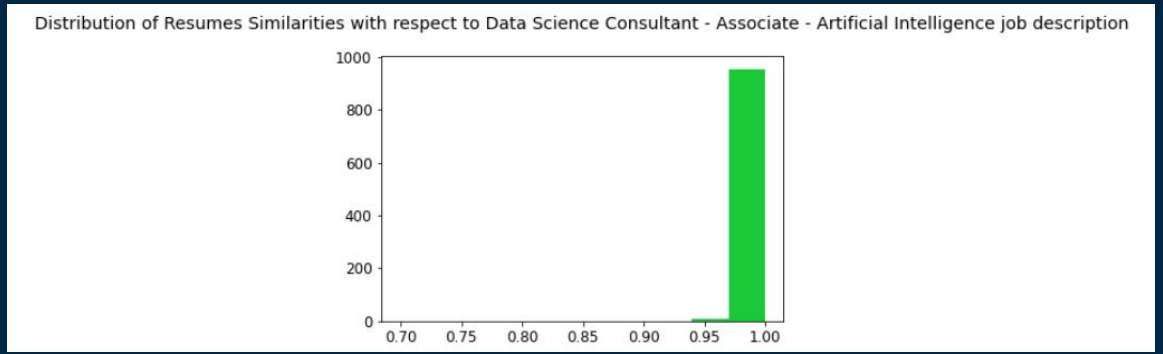
TSNE



6. Build Function to find the similarities between specific job description over all the resumes

The best 10 applicantions of ' Data Science Consultant - Associate - Artificial Intelligence ' Role at PwC Company:

	Applicantion_ID	Similarity	Category
519	519	1.0	Operations Manager
523	523	1.0	Operations Manager
547	547	1.0	Operations Manager
511	511	1.0	Operations Manager
515	515	1.0	Operations Manager
527	527	1.0	Operations Manager
531	531	1.0	Operations Manager
535	535	1.0	Operations Manager
539	539	1.0	Operations Manager
543	543	1.0	Operations Manager

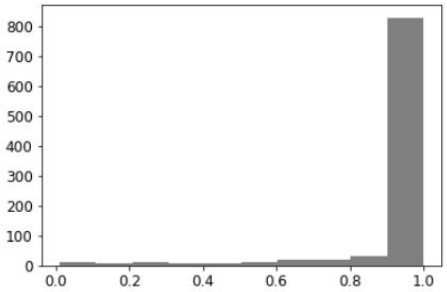


2- Creating Corpus by doc2bow

```
The best 10 applicantions of ' Data Science Consultant - Associate - Artificial Intelligence ' Role at  PwC Company:

     Applicantion_ID  Similarity           Category
511              511         1.0  Operations Manager
519              519         1.0  Operations Manager
547              547         1.0  Operations Manager
535              535         1.0  Operations Manager
515              515         1.0  Operations Manager
523              523         1.0  Operations Manager
530              530         1.0  Operations Manager
518              518         1.0  Operations Manager
550              550         1.0  Operations Manager
522              522         1.0  Operations Manager

Text(0.5, 1.0, 'Distribution of Resumes Similarities with respect to Data Science Consultant - Associate - Artificial Inte
lligence job description\n')
```



DOC2VEC

04

DOC2VEC

By capturing the semantics and the meaning of the input texts. This means that texts which are similar in meaning or context will be closer to each other in vector space than texts which aren't necessarily related.

Top resumes for Data Science Consultant Role

```
query=j_df.loc[0,"Clean_Job_Descriptions"]
```

```
V1_infer [-0.29555482  2.3926542   4.1826797  -4.4016156  -0.54693073 -0.3435357
  0.283935   -4.435324    1.1551149  -0.84808785  0.9646278  -0.3250328
 -1.2072817  -1.6462535  -0.58223784  1.6676261   2.020507    3.2771957
 -5.6735363   1.926741  ]
```

MOST 0.7285773158073425: education details january 1992 january 2003 first year science mumbai maharashtra st micheal high personal fitness trainer level3 personal trainer skill details company details company golds gym fitness solution flora hotel description certification american college sports science golds gym heart saver reps level 3 responsibilities obtain challenging position commensurate qualification experience field health fitness environment accomplishments good skills used fitness

SECOND-MOST 0.724015474319458: education details january 1992 january 2003 first year science mumbai maharashtra st micheal high personal fitness trainer level3 personal trainer skill details company details company golds gym fitness solution flora hotel description certification american college sports science golds gym heart saver reps level 3 responsibilities obtain challenging position commensurate qualification experience field health fitness environment accomplishments good skills used fitness

MEDIAN 0.6643062233924866: skills python tableau data visualization r studio machine learning statistics iabac certified data scientist versatile experience 1 years managing business data science consulting leading innovation projects bringing business ideas working real world solutions strong advocator augmented era human capabilities enhanced machines fahed passionate bringing business concepts area machine learning ai robotics etc real life solutions education details january 2017 b tech computer science engineering mohali punjab indo global college engineering data science consultant data science consultant datamites skill details machine learning exprience 13 months python exprience 24 months solutions exprience 24 months data science exprience 24 months data visualization exprience 24 months tableau exprience 24 monthscompany details company datamites description analyzed processed complex data sets using advanced querying visualization analytics tools responsible loading extracting validation client data worked manipulating cleaning processing data using python used tableau data visualization company heretic solutions pvt ltd description worked closely business identify issues used data propose solutions effective decision making manipulating cleansing processing data using python excel r analyzed raw data drawing conclusions developing recommendations used machine learning tools statistical techniques produce solutions problems

LEAST 0.6360458135604858: skills python tableau data visualization r studio machine learning statistics iabac certified data scientist versatile experience 1 years managing business data science consulting leading innovation projects bringing business ideas working real world solutions strong advocator augmented era human capabilities enhanced machines fahed passionate bringing business concepts area machine learning ai robotics etc real life solutions education details january 2017 b tech computer science engineering mohali punjab indo global college engineering data science consultant data science consultant datamites skill details machine learning exprience 13 months python exprience 24 months solutions exprience 24 months data science exprience 24 months data visualization exprience 24 months tableau exprience 24 monthscompany details company datamites description analyzed processed complex data sets using advanced querying visualization analytics tools responsible loading extracting validation client data worked manipulating cleaning processing data using python used tableau data visualization company heretic solutions pvt ltd description worked closely business identify issues used data propose solutions effective decision making manipulating cleansing processing data using python excel r analyzed raw data drawing conclusions developing recommendations used machine learning tools statistical techniques produce solutions problems

```
[('271', 0.7285773158073425),
 ('265', 0.724015474319458),
 ('277', 0.7107254862785339),
 ('283', 0.7007771730422974),
 ('289', 0.6794911623001099),
 ('26', 0.6643062233924866),
 ('36', 0.6584516763687134),
 ('269', 0.6397997736930847),
 ('16', 0.6366413831710815),
 ('6', 0.6360458135604858)]
```

Another Point of View

Jobs recommendation by Doc2vec

```
query=r_df.loc[0,"Cleaned_Resume"]
```

Infer vector

```
[-1.42833     1.9001637   4.00275     4.350105   -1.7420647   1.1900498
  1.1057764   1.7911824  -1.5740274   0.20123945 -0.6958347  -3.2743049
  2.4414587  -1.046271    0.31195134  0.59389395  1.781105   -0.7970009
 -3.7747104  -0.38275546]
```

Top Ten Job recommendations

```
[('456', 0.7025666832923889),
 ('457', 0.6985500454902649),
 ('323', 0.6783030033111572),
 ('482', 0.6744669675827026),
 ('300', 0.6474058628082275),
 ('219', 0.6400119066238403),
 ('81', 0.6353856921195984),
 ('265', 0.6330556273460388),
 ('330', 0.6309281587600708),
 ('490', 0.630798876285553)]
```

Jobs recommendation by LDA

```
query_string = r_df.loc[10, 'Cleaned_Resume']



from sklearn.metrics.pairwise import cosine_similarity



query_bow = vec.transform([query_string])# create bag of words
topcs_vector = lda_model.transform(query_bow) # generate vector of topics
similarities = cosine_similarity(topcs_vector,document_topics)[0] # similarity to every other document. 1: identical.
sililarity_rank = np.argsort(similarities)[::-1] # rank all topics by their similarity to the query.
# print results:
for rank, item_index in enumerate(sililarity_rank[:3]):
    print(f"{rank}.\t{similarities[item_index]:6.6f}\t{j_df['Clean_Job_Descriptions'].iloc[item_index][:500]}")
```

```
0.      0.948375        looking data scientists interested using data draw insights result policy changes business p
rocess optimisation benefiting public applicant scoping projects stakeholders using data sets across government agen
cies applying business acumen tease relevant impactful insights presenting insights clear concise manner using appro
priate visualisations training working experiences data analytics comfortable hands data manipulation data modelling
data visualisation also comfortable engaging stakeholders s
1.      0.941415        senior manager optimisation analytics negotiable salary plus benefits london entrepreneurial
company operates within leisure arena dealing millions customers thus collecting vast amounts data every minute ever
y single day role presents real exciting challenges extracting commercial value ever growing volume data unique oppo
rtunity utilise real passion data science whilst directly influencing bottom line working brightest minds tech indus
try energetic international work environment role responsib
2.      0.908829        company r grid early stage start streamlines administrative medical research processes using
machine learning artificial intelligence job overview looking machine learning engineer skills natural language proc
essing job responsibilities include working team software engineers building app utilises deep learning based natura
l language processing data science combined user friendly ux strong problem solving ability background mathematical
modeling statistical analysis also able align products busi
```

Conclusion

1. The solution of our problem is not only counting words and matching between resumes and job descriptions, its more complicated!

2. CountVectorizer and TF-IDF models performances are almost the same

3.The problem with TF-IDF approach is that words like ‘Python’ which is a key skill and can appear in multiple Job Description would lose its weights because of the IDF factor.

4. We expected better performance from LDA as it depends on the topics of the documents. If we work on the limitations , we can improve the performance of the model

5. Doc2Vec approach represent only two vectors;. a single vector representation of all the jobs and the resume vector representation

6. Finally, We can not say clearly which model is the best . Our solution needs improvement to increase the accuracy of cosine similarity.

LIMITATIONS

1. The model accepts resumes in CSV format, however in the real world, resumes of candidates are typically in .docx, .pdf, or other formats. So, there is an extra process to do which is convert pdfs and docs to texts and extract the relevant information from them

3.. The data sets are not big enough to train and test models.

4.. There are duplicated job descriptions that are posted by the same companies, and duplicated resumes sent by applicants.

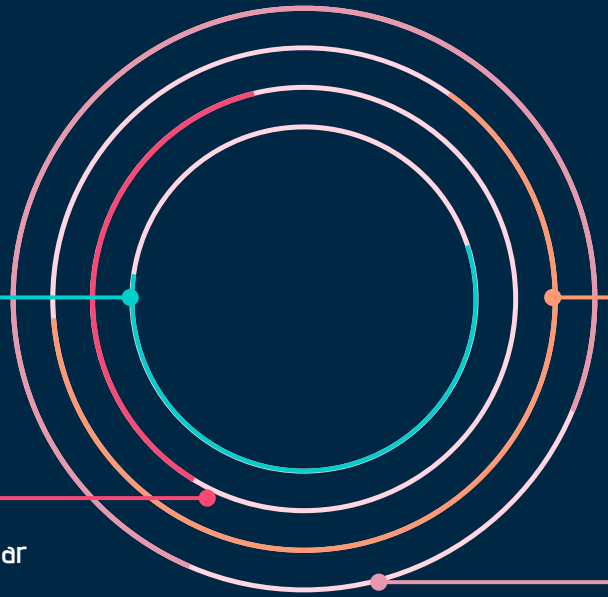
2. We found the two datasets randomly and separately . So not all the resumes are relevant to the job postings.

5. We can improve the models by preparing a keywords dataset, so the matching between job-resume is not limited to count all the words, but to count the relevant keywords.

6. Some great applicants send their resumes in different formats. So maybe our model will filter them.

Ways to Improve Matching:

1. We can improve the models by preparing a keywords dataset, so the matching between job-resume is not limited to count general words, but to count the relevant keywords. So we can have positive and negative cases to find the threshold similarities value



2. By training an LDA model. Similar Jobs can be clustered, and topics can be extracted. Then to apply matching.

3. To improve the preprocess function. This function skip some skills such as: R language skill.

4. To try more models: i.e., to build word2vec and to find the words which are close to the Job Description vector

Word Embedding

05

Word Embedding (works as word2vec)

1. Text summarization for both resumes and job descriptions to help us find the most useful information that we want to use when comparing similarities between the resumes and the job descriptions.

	Applicantion_ID	Cleaned_Resume	Summary	Embeddings
0	0	skills programming languages python pandas num...	skills programming languages python pandas num...	[-0.48579407, 1.1287568, 0.3529383, 0.38239697...
1	1	education details may 2013 may 2017 b e uit rg...	education details may 2013 may 2017 b e uit rg...	[-0.28759128, 1.0566523, 0.28568777, 0.2083249...
2	2	areas interest deep learning control system de...	areas interest deep learning control system de...	[-0.24453898, 1.0482582, 0.6277059, 0.09711098...
3	3	skills r python sap hana tableau sap hana sql ...	skills r python sap hana tableau sap hana sql ...	[-0.3049553, 1.0736828, 0.81079173, 0.11519936...
4	4	education details mca ymcaust faridabad haryan...	education details mca ymcaust faridabad haryan...	[-0.471591, 1.0701611, 0.20403111, 0.41489664,...

2. Classify the text using Word embedding with Bert to convert the words to numbers where words that have the same meaning have a similar representation.

	Job_ID	Clean_Job_Descriptions	Summary	Embeddings
0	0	background pwc trusted adviser largest organis...	background pwc trusted adviser largest organis...	[-0.6036923, 0.78614163, 1.055162, 0.033558644...
1	1	looking data science manager help clients navi...	looking data science manager help clients navi...	[-0.5766502, 0.89438796, 1.2430494, 0.16873676...
2	2	fraud analytics managerlondon currently flexib...	fraud analytics managerlondon currently flexib...	[-0.64401454, 1.1870322, 0.9419006, 0.25586233...
3	3	cib applied ai machine learning economic finan...	cib applied ai machine learning economic finan...	[-0.5888, 1.0973191, 0.101591766, 0.13015941, ...
4	4	opportunity tax ventures team expanding ai dat...	opportunity tax ventures team expanding ai dat...	[-0.42869496, 0.45713067, 0.9302575, -0.250375...

3. Apply cosine similarity to find the best candidate for a job

```
def match_resume_with_job_description_embedding(job_index):

cosines = []
for i in range(len(r_df)):
      cosines.append(cosine(r_df.iloc[i]['Embeddings'], j_df.iloc[job_index]['Embeddings']))
print("Index of best candidate is : ", cosines.index(max(cosines)),'with similarity', max(cosines))

match_resume_with_job_description_embedding(0)
```

```
print("Index of best candidate is : ", cosines.index(max(cosines)),'with similarity', max(cosines))
```

```
Index of best candidate is :  6 with similarity 0.85714775
```

Top 10 Applicants:

```
df=pd.DataFrame (cosines, columns = ['Similarity'])
match=df.nlargest(n=10, columns=['Similarity'], keep='all')
print('ID' ,match.head(10))
```

ID	Similarity
6	0.857148
16	0.857148
26	0.857148
36	0.857148
719	0.833126
730	0.833126
741	0.833126
743	0.826696
750	0.826696
757	0.826696

Thank you!