

```
1 import numpy as np
2 import pandas as pd
3 import seaborn as sns
4 import matplotlib.pyplot as plt
5 import nltk
6 nltk.download('punkt')
7 from nltk.tokenize import word_tokenize
8 from nltk.tag import pos_tag
9 from sklearn.metrics.pairwise import linear_kernel
10 from sklearn.feature_extraction.text import TfidfVectorizer
11 from sklearn.metrics.pairwise import cosine_similarity

[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Package punkt is already up-to-date!
```

```
1 data = pd.read_csv('/content/job_descriptions.csv')
```

```
1 data.head(3)
```

	Job Id	Experience	Qualifications	Salary Range	location	Country	lat:
0	1089843540111562	5 to 15 Years	M.Tech	\$59K-\$99K	Douglas	Isle of Man	54
1	398454096642776	2 to 12 Years	BCA	\$56K-\$116K	Ashgabat	Turkmenistan	38
2	481640072963533	0 to 12 Years	PhD	\$61K-\$104K	Macao	Macao SAR, China	22

3 rows × 23 columns

```
1 data.shape

(19434, 23)
```

```
1 data.columns

Index(['Job Id', 'Experience', 'Qualifications', 'Salary Range', 'location',
      'Country', 'latitude', 'longitude', 'Work Type', 'Company Size',
      'Job Posting Date', 'Preference', 'Contact Person', 'Contact',
      'Job Title', 'Role', 'Job Portal', 'Job Description', 'Benefits',
      'skills', 'Responsibilities', 'Company', 'Company Profile'],
      dtype='object')
```

```
1 data.isnull().sum()
```

Job Id	0
Experience	0
Qualifications	0
Salary Range	0
location	0
Country	0
latitude	0
longitude	0
Work Type	0
Company Size	0
Job Posting Date	0
Preference	0
Contact Person	0
Contact	0
Job Title	0
Role	0
Job Portal	0

```

Job Description      0
Benefits             0
skills              0
Responsibilities     1
Company             1
Company Profile      65
dtype: int64

```

```
1 data.duplicated().sum()
```

```
0
```

```
1 data['Job Title'].unique()
```

```

array(['Digital Marketing Specialist', 'Web Developer',
      'Operations Manager', 'Network Engineer', 'Event Manager',
      'Software Tester', 'Teacher', 'UX/UI Designer', 'Wedding Planner',
      'QA Analyst', 'Litigation Attorney', 'Mechanical Engineer',
      'Network Administrator', 'Account Manager', 'Brand Manager',
      'Social Worker', 'Social Media Coordinator',
      'Email Marketing Specialist', 'HR Generalist', 'Legal Assistant',
      'Nurse Practitioner', 'Account Director', 'Software Engineer',
      'Purchasing Agent', 'Sales Consultant', 'Civil Engineer',
      'Network Security Specialist', 'UI Developer', 'Financial Planner',
      'Event Planner', 'Psychologist', 'Electrical Designer',
      'Data Analyst', 'Technical Writer', 'Tax Consultant',
      'Account Executive', 'Systems Administrator',
      'Database Administrator', 'Research Analyst', 'Data Entry Clerk',
      'Registered Nurse', 'Investment Analyst', 'Speech Therapist',
      'Sales Manager', 'Landscape Architect', 'Key Account Manager',
      'UX Researcher', 'Investment Banker', 'IT Support Specialist',
      'Art Director', 'Software Developer', 'Project Manager',
      'Customer Service Manager', 'Procurement Manager',
      'Substance Abuse Counselor', 'Supply Chain Analyst',
      'Data Engineer', 'Accountant', 'Sales Representative',
      'Environmental Consultant', 'Electrical Engineer',
      'Systems Engineer', 'Art Teacher', 'Human Resources Manager',
      'Inventory Analyst', 'Legal Counsel', 'Database Developer',
      'Procurement Specialist', 'Systems Analyst', 'Copywriter',
      'Content Writer', 'HR Coordinator', 'Business Development Manager',
      'Java Developer', 'Supply Chain Manager', 'Event Coordinator',
      'Family Nurse Practitioner', 'Front-End Engineer',
      'Customer Success Manager', 'Procurement Coordinator',
      'Urban Planner', 'Architectural Designer', 'Financial Analyst',
      'Environmental Engineer', 'Back-End Developer',
      'Structural Engineer', 'Market Research Analyst',
      'Customer Service Representative', 'Customer Support Specialist',
      'Business Analyst', 'Social Media Manager', 'Family Lawyer',
      'Chemical Analyst', 'Network Technician', 'Interior Designer',
      'Software Architect', 'Nurse Manager', 'Veterinarian',
      'Process Engineer', 'IT Manager', 'Quality Assurance Analyst',
      'Pharmaceutical Sales Representative', 'Office Manager',
      'Architect', 'Physician Assistant', 'Marketing Director',
      'Front-End Developer', 'Research Scientist', 'Executive Assistant',
      'HR Manager', 'Marketing Manager', 'Public Relations Specialist',
      'Financial Controller', 'Investment Advisor', 'Aerospace Engineer',
      'Marketing Analyst', 'Paralegal', 'Landscape Designer',
      'Web Designer', 'Occupational Therapist', 'Legal Advisor',
      'Marketing Coordinator', 'Dental Hygienist', 'SEM Specialist',
      'SEO Specialist', 'Pediatrician', 'QA Engineer', 'Data Scientist',
      'Financial Advisor', 'Personal Assistant', 'SEO Analyst',
      'Network Analyst', 'Mechanical Designer', 'Marketing Specialist',
      'Graphic Designer', 'Finance Manager', 'Physical Therapist',
      'Product Designer', 'Administrative Assistant', 'Brand Ambassador',
      'Project Coordinator', 'Product Manager', 'IT Administrator',
      'Sales Associate', 'Chemical Engineer', 'Legal Secretary',
      'Market Analyst'], dtype=object)

```

```
1 data['skills'].unique()
```

```

'Environmental assessments Regulatory compliance Environmental impact analysis',
'Agile product development Scrum or Kanban frameworks Backlog management User story writing Prioritization techniques
Collaboration with development teams Product vision and strategy',
'Architectural drafting AutoCAD 2D and 3D modeling Blueprint reading Building codes Collaboration with architects Detail-
oriented',
'Clinical nursing Patient assessment Care planning Staff development Healthcare protocols Communication skills Problem-
solving abilities',
'Web design concepts Graphic design software (e.g., Adobe Creative Suite) HTML and CSS Responsive design User interface (UI)
principles Visual communication Typography Color theory Creativity Attention to detail Collaboration',
'Customer experience strategy Customer journey mapping Feedback analysis User research Metrics and KPIs Communication
skills',
'Frontend and backend development Database integration Problem-solving API knowledge',
'Network troubleshooting IT support Network configuration Technical knowledge Problem-solving',
'Avionics systems Electronic warfare systems Radar systems Aircraft communication systems System integration Testing and
validation',
'System administration Network administration Troubleshooting Security management Server maintenance Backup and recovery',
'Hospitality interior design Hotel and restaurant design Branding and theme integration FF&E (furniture, fixtures, and
equipment) selection Customer experience design',
'Project planning and scheduling Budget management Risk assessment and mitigation Stakeholder communication Agile or
Waterfall methodologies',
'Onboarding process management New employee orientation Training and development HR policies and procedures Employee
engagement strategies',
'Search engine optimization Keyword research SEO tools (e.g., Google Analytics, SEMrush) On-page and off-page optimization',
'Budget planning Financial forecasting Excel proficiency Data analysis',
'Business systems analysis Requirements gathering Process improvement',
'Pediatric specialization Advanced pediatric care Pediatric subspecialty (e.g., pediatric cardiology, pediatric oncology)
Pediatric diagnostics Pediatric treatment',
'Marketing communication Public relations Content strategy Crisis management Media relations',
'Addiction recovery support Motivational interviewing Relapse prevention strategies',
'Mental health assessment and evaluation Psychotherapy techniques Cognitive-behavioral therapy (CBT) Trauma-informed care
Crisis intervention Substance abuse treatment Group therapy Cultural competence Ethical standards and boundaries Active listening
skills',
'Customer service Sales techniques Product knowledge Cash handling Communication skills',
'Knowledge of HVAC systems and equipment Mechanical engineering skills Problem-solving abilities AutoCAD proficiency HVAC
design and installation expertise',
'PCB design Electronics PCB layout software proficiency Circuit design Quality control',
'Creative leadership Art direction Concept development Visual design Advertising and marketing expertise',
'Nursing education Curriculum development Clinical instruction Assessment and evaluation Educational technology
Communication skills',
'Data integration ETL (Extract, Transform, Load) Big data technologies (e.g., Hadoop, Spark) Database management Data
warehousing',
'Search engine marketing PPC advertising SEO strategy Data analysis Digital marketing skills',
'Healthcare industry knowledge Health data analysis HIPAA regulations EMR systems',
'User interface (UI) design User experience (UX) design Wireframing and prototyping tools (e.g., Sketch, Figma) Usability
testing Information architecture Interaction design Mobile and responsive design',
'Retirement planning Social security Investment products Tax planning Estate planning',
'School psychology Special education knowledge Assessment and testing Individualized education plans Collaboration skills',

```

```
1 data.head(10)
```

	Job Id	Experience	Qualifications	Salary Range	location	Country	1
0	1089843540111562	5 to 15 Years	M.Tech	\$59K-\$99K	Douglas	Isle of Man	
1	398454096642776	2 to 12 Years	BCA	\$56K-\$116K	Ashgabat	Turkmenistan	
2	481640072963533	0 to 12 Years	PhD	\$61K-\$104K	Macao	Macao SAR, China	
3	688192671473044	4 to 11 Years	PhD	\$65K-\$91K	Porto-Novo	Benin	
4	117057806156508	1 to 12 Years	MBA	\$64K-\$87K	Santiago	Chile	
5	116831420231957	4 to 12 Years	MCA	\$59K-\$93K	Brussels	Belgium	
6	1292168246729889	3 to 15 Years	PhD	\$63K-\$103K	George Town	Cayman Islands	
7	1498778686197107	2 to 8 Years	M.Com	\$65K-\$102K	SÃ£o TomÃ©	Sao Tome and Principe	
8	1680293940995740	2 to 9 Years	BBA	\$65K-\$102K	Male	Maldives	

```
1 data['Qualifications'].unique()

array(['M.Tech', 'BCA', 'PhD', 'MBA', 'MCA', 'M.Com', 'BBA', 'B.Tech',
      'B.Com', 'BA'], dtype=object)

1 data.columns

Index(['Job Id', 'Experience', 'Qualifications', 'Salary Range', 'location',
      'Country', 'latitude', 'longitude', 'Work Type', 'Company Size',
      'Job Posting Date', 'Preference', 'Contact Person', 'Contact',
      'Job Title', 'Role', 'Job Portal', 'Job Description', 'Benefits',
      'skills', 'Responsibilities', 'Company', 'Company Profile'],
      dtype='object')

1 nltk.download('averaged_perceptron_tagger')
2 phrase = 'Network security protocols Intrusion detection and prevention Security incident response Firewall administration Vulnerability'

https://colab.research.google.com/drive/1XKQ_kEgZnL_hRt4QO8k1-eKd4IoYIZPL#scrollTo=Agee9xHGulqG&uniqifier=1&printMode=true
```

```

3 tokens = word_tokenize(phrase)
4 tags = pos_tag(tokens)
5 competence_tags = ['NN', 'NNS', 'NNP', 'JJ']
6 competences = []
7 for token, tag in tags:
8     if tag in competence_tags:
9         competences.append(token)
10
11 print("Compétences extraites :", competences)
12
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data] /root/nltk_data...
[nltk_data] Unzipping taggers/averaged_perceptron_tagger.zip.
Compétences extraites : ['Network', 'security', 'protocols', 'Intrusion', 'detection', 'prevention', 'Security', 'incident', 'response']

```

```

1 def extract_skills(phrase):
2     tokens = word_tokenize(phrase)
3     tags = pos_tag(tokens)
4     competence_tags = ['NN', 'NNS', 'NNP', 'JJ']
5     competences = []
6     for token, tag in tags:
7         if tag in competence_tags:
8             competences.append(token)
9     return competences

```

```

1 phrase = 'Network security protocols Intrusion detection and prevention Security incident response Firewall administration Vulnerability
2 competences_extraites = extract_skills(phrase)
3 print("Compétences extraites :", competences_extraites)

```

```
Compétences extraites : ['Network', 'security', 'protocols', 'Intrusion', 'detection', 'prevention', 'Security', 'incident', 'response']
```

```

1 data.columns

Index(['Job Id', 'Experience', 'Qualifications', 'Salary Range', 'location',
      'Country', 'latitude', 'longitude', 'Work Type', 'Company Size',
      'Job Posting Date', 'Preference', 'Contact Person', 'Contact',
      'Job Title', 'Role', 'Job Portal', 'Job Description', 'Benefits',
      'skills', 'Responsibilities', 'Company', 'Company Profile'],
      dtype='object')

```

```
1 data.isnull().sum()
```

```

Job Id          0
Experience      0
Qualifications  0
Salary Range    0
location        0
Country         0
latitude        0
longitude       0
Work Type       0
Company Size    0
Job Posting Date 0
Preference      0
Contact Person  0
Contact         0
Job Title       0
Role            0
Job Portal      0
Job Description  0
Benefits        0
skills          0
Responsibilities 1
Company         1
Company Profile 65
dtype: int64

```

```
1 data['extracted_skills'] = data['skills'].apply(extract_skills)
```

```
1 data.columns
```

```

Index(['Job Id', 'Experience', 'Qualifications', 'Salary Range', 'location',
      'Country', 'latitude', 'longitude', 'Work Type', 'Company Size',

```

```
'Job Posting Date', 'Preference', 'Contact Person', 'Contact',
'Job Title', 'Role', 'Job Portal', 'Job Description', 'Benefits',
'skills', 'Responsibilities', 'Company', 'Company Profile',
'extracted_skills'],
dtype='object')
```

```
1 data['extracted_skills']
```

```
0      [Social, media, platforms, e.g., Facebook, Twi...
1      [HTML, CSS, JavaScript, Frontend, frameworks, ...
2      [Quality, control, methodologies, Statistical,...
3      [Wireless, network, design, architecture, Wi-F...
4      [Event, Conference, logistics, Budget, managem...
...
19429   [Investment, management, Financial, analysis, ...
19430   [Construction, management, Structural, design,...
19431   [Organizational, time, management, skills, Cal...
19432   [analytics, Data, analysis, tools, e.g., Googl...
19433   [Brand, management, Public, relations, Brand, ...
Name: extracted_skills, Length: 19434, dtype: object
```

```
1
```

```
1 data.columns
```

```
Index(['Job Id', 'Experience', 'Qualifications', 'Salary Range', 'location',
'Country', 'latitude', 'longitude', 'Work Type', 'Company Size',
'Job Posting Date', 'Preference', 'Contact Person', 'Contact',
'Job Title', 'Role', 'Job Portal', 'Job Description', 'Benefits',
'skills', 'Responsibilities', 'Company', 'Company Profile',
'extracted_skills'],
dtype='object')
```

```
1
```

```
1
```

```
1 data.drop(columns= ['latitude' , 'longitude' , 'location' , 'Preference', 'Job Portal' , 'Benefits' , 'Company' , 'Company Profile' , '
```

```
1 data.columns
```

```
Index(['Job Id', 'Experience', 'Qualifications', 'Salary Range', 'Country',
'Work Type', 'Job Posting Date', 'Contact Person', 'Contact',
'Job Title', 'Role', 'Job Description', 'skills', 'Responsibilities',
'extracted_skills'],
dtype='object')
```

```
1 data.drop(columns = ['Experience' , 'Salary Range' , 'Contact Person' , 'Country' , 'Job Posting Date'],inplace = True)
```

```
1 data.columns
```

```
Index(['Job Id', 'Qualifications', 'Work Type', 'Contact', 'Job Title', 'Role',
'Job Description', 'skills', 'Responsibilities', 'extracted_skills'],
dtype='object')
```

```
1 data.drop(columns = ['Work Type', 'Contact'],inplace=True)
```

```
1 selected_columns = ['Job Title','Job Description', 'skills', 'Responsibilities', 'extracted_skills']
```

```
1 print(selected_columns)
```

```
['Job Title', 'Job Description', 'skills', 'Responsibilities', 'extracted_skills']
```

```
1 selected_data = data[selected_columns]
```

```
1 selected_data['text'] = selected_data['Job Title'] + ' ' + selected_data['Job Description'] + ' ' + selected_data['skills']+' '+selected_data['Responsibilities']
```

```
<ipython-input-31-f97bf8c4be6e>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/indexing.html#returning-a-view-versus-a-selected\\_data\['text'\] = selected\\_data\['Job Title'\] + ' ' + selected\\_data\['Job Description'\] + ' ' + selected\\_data\['skills'\]+' '+selected\\_data\['text'\] = selected\\_data\['text'\].fillna\(''\)](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-selected_data['text'] = selected_data['Job Title'] + ' ' + selected_data['Job Description'] + ' ' + selected_data['skills']+' '+selected_data['text'] = selected_data['text'].fillna(''))

```
1 selected_data['text'] = selected_data['text'].fillna('')
```

<ipython-input-33-7a7ad2e9277d>:1: SettingWithCopyWarning:  
A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation: [https://pandas.pydata.org/pandas-docs/stable/user\\_guide/indexing.html#returning-a-view-versus-a-selected\\_data\['text'\] = selected\\_data\['text'\].fillna\(''\)](https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-selected_data['text'] = selected_data['text'].fillna(''))

```
1 tfidf_vectorizer = TfidfVectorizer(stop_words='english')
2 tfidf_matrix = tfidf_vectorizer.fit_transform(selected_data['text'])
```

```
1 def predictions(skills, tfidf_vectorizer=tfidf_vectorizer, tfidf_matrix=tfidf_matrix, top_n=8):
2     skills_vector = tfidf_vectorizer.transform([skills])
3     cosine_similarities = cosine_similarity(skills_vector, tfidf_matrix).flatten()
4     job_scores = {}
5     for skill in skills.split():
6         similar_jobs_indices = cosine_similarities.argsort()[::-1]
7         for index in similar_jobs_indices:
8             job_title = selected_data.iloc[index]['Job Title']
9             if job_title not in job_scores:
10                 job_scores[job_title] = 0
11                 if skill in selected_data.iloc[index]['skills']:
12                     job_scores[job_title] += 1
13     sorted_jobs = sorted(job_scores.items(), key=lambda x: x[1], reverse=True)
14     predicted_job_titles = [job[0] for job in sorted_jobs[:top_n]]
15     return predicted_job_titles
```

```
1 skills = "java php html networking"
2 predicted_jobs = predictions(skills)
3 print("8 Posts predicted : ", skills, ":")
4 for i, job_title in enumerate(predicted_jobs):
5     print(i+1, ".", job_title)
6
```