

Career Map

Yuxin Zhao
Jinyi Luo

Career Map

Content

1. Background
 2. Intro to Amazon Web Services
 3. Standardization
 - a. Preprocessing
 - b. String Similarity
 - i. Cosine Similarity
 - ii. Jaccard's Coefficient
Similarity
 - iii. Word2Vec
 4. Career Map - Network Analysis
-

Background

Applicants:

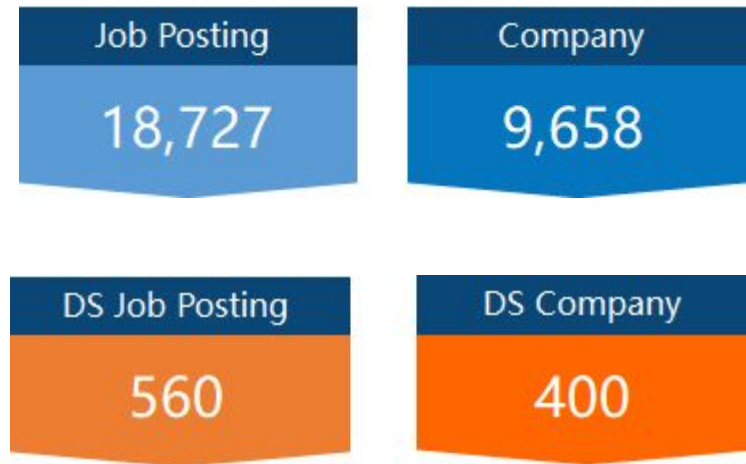
- What can I do?
- Where am I in the labor market?
- How far is it to my dream position?
- Next Step?

Employers:

- Skills
- Experience



Job Posting - 18,727




Resume - 3,564,157

	All	DC
Accountant	586,003	35,384
Marketing Manager	418,090	
Financial Manager	400,268	25,340
Financial Analyst	341,882	21,058
Management Analyst	281,921	27,637
IT Manager	257,602	19,181
Business Operations Manager	186,300	11,491
Construction Manager	160,471	8,488
Loan Officer	146,084	8,532
Laboratory Technician	116,424	7,547
Financial Advisor	101,380	5,379
Interpreter	96,441	7,133
HR Specialist	91,969	6,413
High School Teacher	60,590	3,741
Compliance Officer	56,259	3,517
Database Administrator	53,613	5,503
Computer Support Specialist	38,760	3,310
Civil Engineer	27,964	1,712
Fundraiser	26,972	1,446
Computer Systems Analyst	26,736	3,048
Information Security Analyst	22,968	3,963
Computer Systems Administrator	20,550	2,171
Data Scientist	18,467	1,298
Lawyer	14,032	769
Cost Estimator	6,567	374
Actuary	2,553	117
Cartographer	2,011	244
Computer Network Architect	1,280	154
Grand Total	3,564,157	214,950

Amazon Web Services (AWS)



AWS Management Console

EC2 (Virtual Server in the Cloud) 

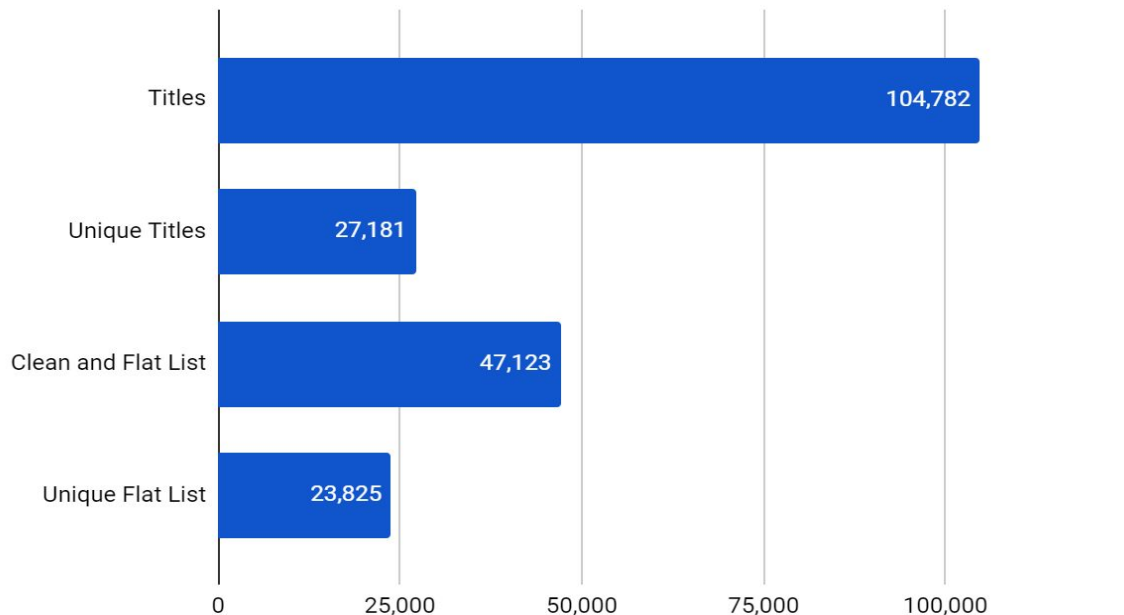
T2 Instances (Ubuntu Server 18.04) **ubuntu** 

PuTTY 

FileZilla 

Preprocessing

	All Resumes	All Titles	All Unique Titles	DC Unique Titles
Database Administrator	53,613	351,697	100,970	18,348
Computer Systems Administrator	20,550	165,934	53,418	10,942
Data Scientist	18,467	104,782	27,181	5,220



Purpose



Detect different
expressions of same
meaning

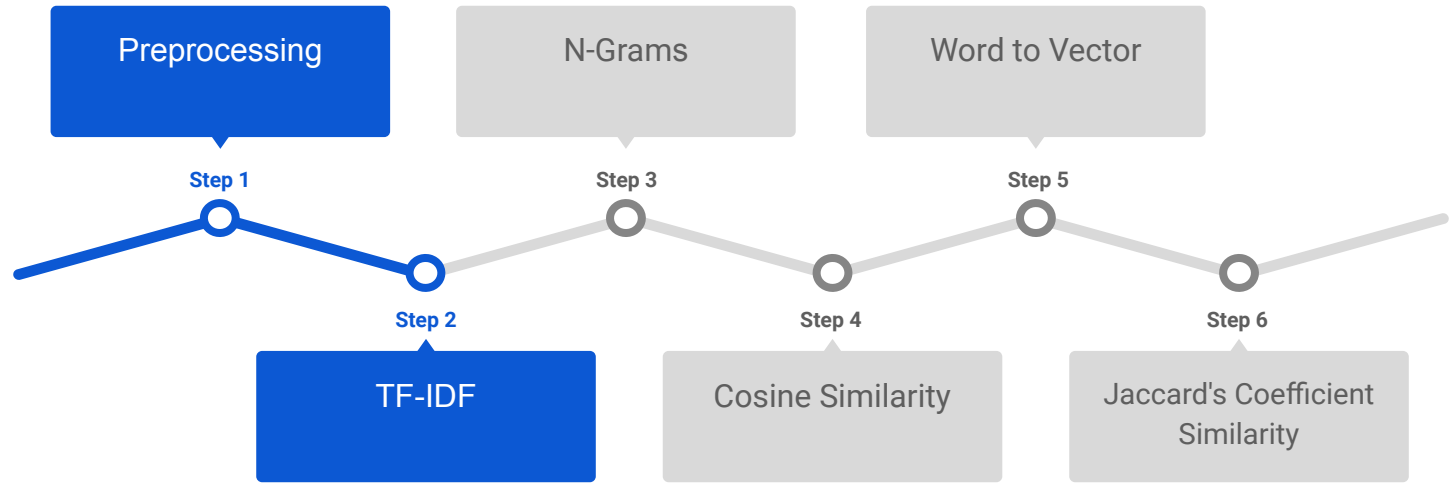


Build a career map



Provide suggestions to the
applicants

String Similarity



Term Frequency-Inverse Document Frequency

$$TF(t) = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}}$$

$$IDF(t) = \log_e \left(\frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}} \right)$$



$$TF(t) = 3/100 = 0.03$$

$$IDF(t) = \log(20,000 / 200) = 2$$

$$TF\text{-}IDF \text{ weight} = 0.03 * 2 = 0.06$$

N-Grams

scientist

```
igrams('scientist',n=3)  
Out[1]: ['sci', 'cie', 'ien', 'ent', 'nti', 'tis', 'ist']
```

```
print(tf_idf_matrix[245])
```



(0, 5053)	0.4009
(0, 1469)	0.3999
(0, 2869)	0.3912
(0, 2076)	0.3043
(0, 4014)	0.4103
(0, 5441)	0.4108
(0, 3045)	0.3108

science

```
igrams('science',n=3)  
Out[2]: ['sci', 'cie', 'ien', 'enc', 'nce']
```

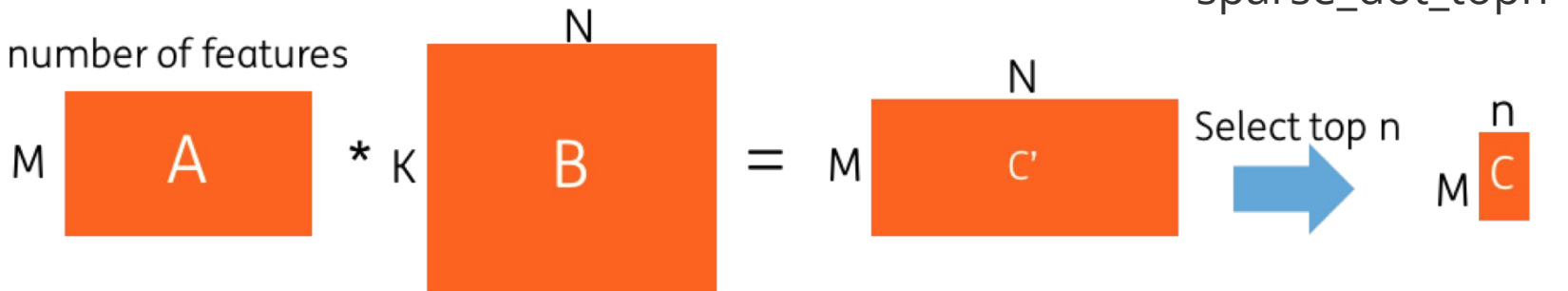
```
print(tf_idf_matrix[2501])
```



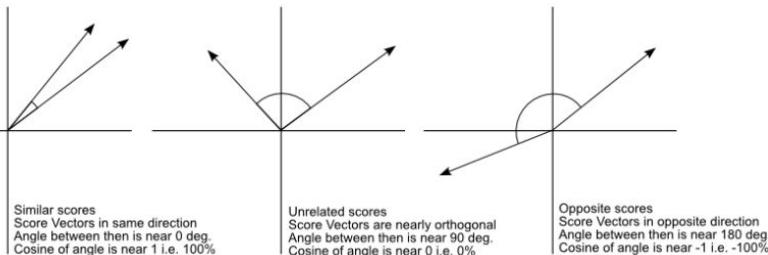
(0, 1469)	0.4157
(0, 2869)	0.4067
(0, 2064)	0.5236
(0, 3805)	0.4624

Cosine Similarity

K: number of features



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$



Algorithm

Sparse (MATRIX)

Step 1: Set M to number of rows in MATRIX

Step 2: Set N to number of columns in MATRIX

Step 3: $I = 0$, $NNZ = 0$. Declare A, JA, and IA.

Set $IA[0]$ to 0

Step 4: for $I = 0 \dots N-1$

Step 5: for $J = 0 \dots N-1$

Step 5: If MATRIX $[I][J]$ is not zero

Add MATRIX $[I][J]$ to A

Add J to JA

$NNZ = NNZ + 1$

[End of IF]

Step 6: Add NNZ to IA

[End of J loop]

[End of I loop]

Step 7: Print vectors A, IA, JA

Step 8: END

Input :
$$\begin{bmatrix} 10 & 20 & 0 & 0 & 0 & 0 \\ 0 & 30 & 0 & 4 & 0 & 0 \\ 0 & 0 & 50 & 60 & 70 & 0 \\ 0 & 0 & 0 & 0 & 0 & 80 \end{bmatrix}$$

Output : $A = [10 \quad 20 \quad 30 \quad 4 \quad 50 \quad 60 \quad 70 \quad 80]$
 $IA = [0 \quad 2 \quad 4 \quad 7 \quad 8]$
 $JA = [0 \quad 1 \quad 1 \quad 3 \quad 2 \quad 3 \quad 4 \quad 5]$

Data Scientist

5	data scientist	data scientist i	0.896795
6	data scientist	data scientist in	0.870202
7	data scientist	data scientist r	0.863371
8	data scientist	d data scientist	0.86019
9	data scientist	data scientist co	0.853945

Data Scientist Intern

43	data scientist intern	d data scientist intern	0.910315
44	data scientist intern	data scientist in	0.88135
45	data scientist intern	data scientist intel	0.877297
46	data scientist intern	scientist intern	0.861003
47	data scientist intern	data scientist i	0.855215

Senior Data Scientist

499	senior data scientist	senior data scientist ii	0.899819
500	senior data scientist	sr senior data scientist	0.858319
501	senior data scientist	senior data scienist	0.856755

Senior Software Engineer

281	senior software engineer	senior software engineer lead	0.889968
282	senior software engineer	sr software engineer	0.855782

Senior Software Architect

505	senior software architect	sr software architect	0.869285
-----	---------------------------	-----------------------	----------

Lead Data Scientist

1943	lead data scientist	head data scientist	0.916895
1944	lead data scientist	d data scientist	0.875953

Senior Network Design Engineer

2605	senior network design engineer	senior network design engineer iv	0.945961
2603	senior network design engineer iv	senior network design engineer	0.945961

Senior Operations Research Analyst

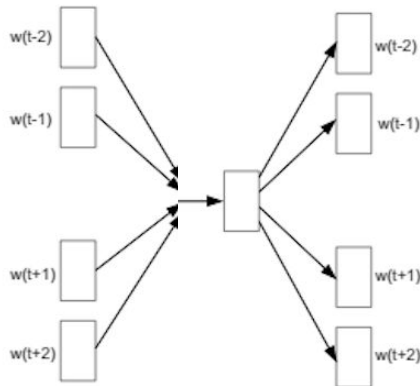
4828	senior operations research analyst	sr operations research analyst	0.876618
------	------------------------------------	--------------------------------	----------

Word2Vector

"You shall know a word by the company it keeps." – J.R.Firth



Input:
senior



label:
investment

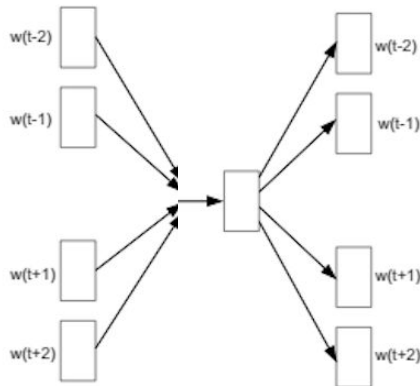
Skip-gram

Word2Vector

"You shall know a word by the company it keeps." – J.R.Firth



Input:
senior
investment
investment

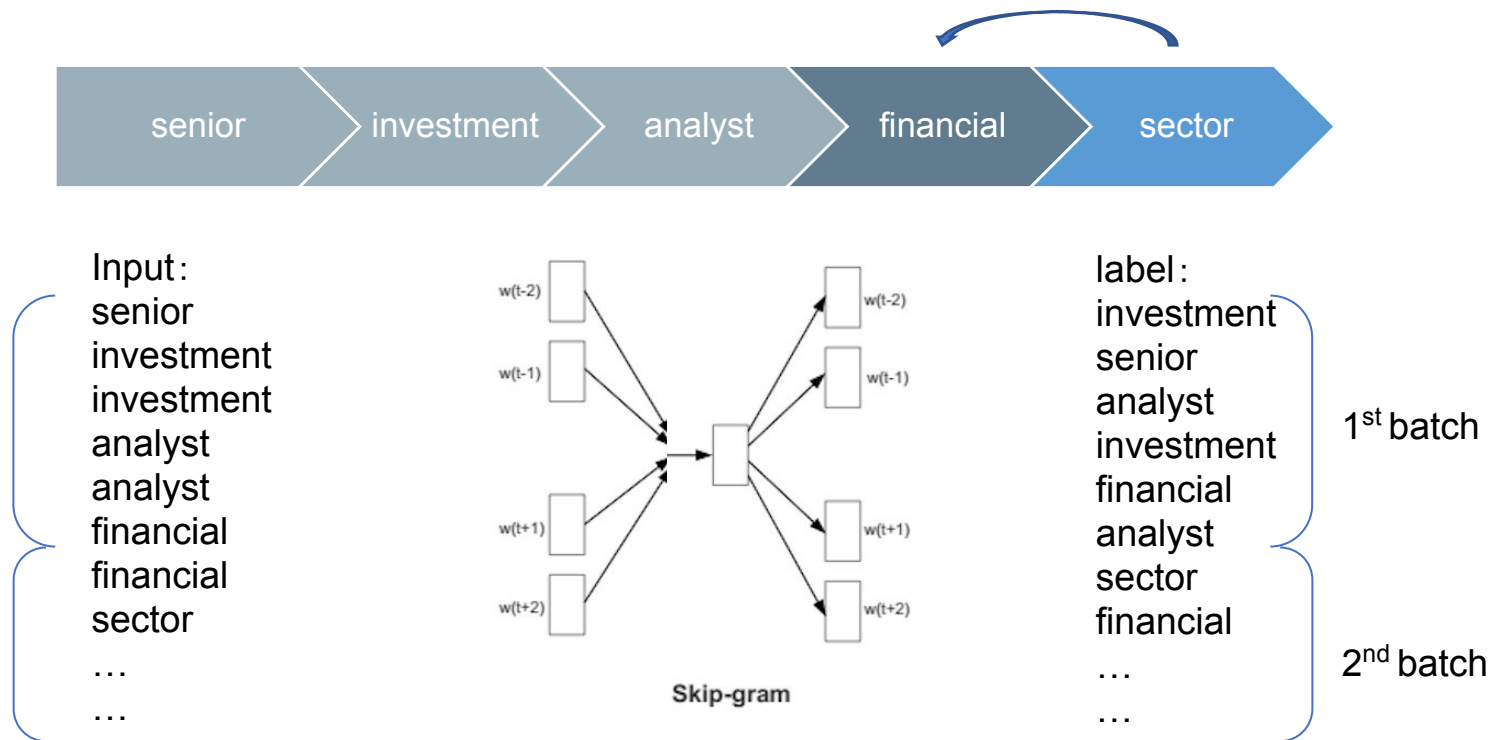


label:
investment
senior
analyst

Skip-gram

Word2Vector

"You shall know a word by the company it keeps." – J.R.Firth



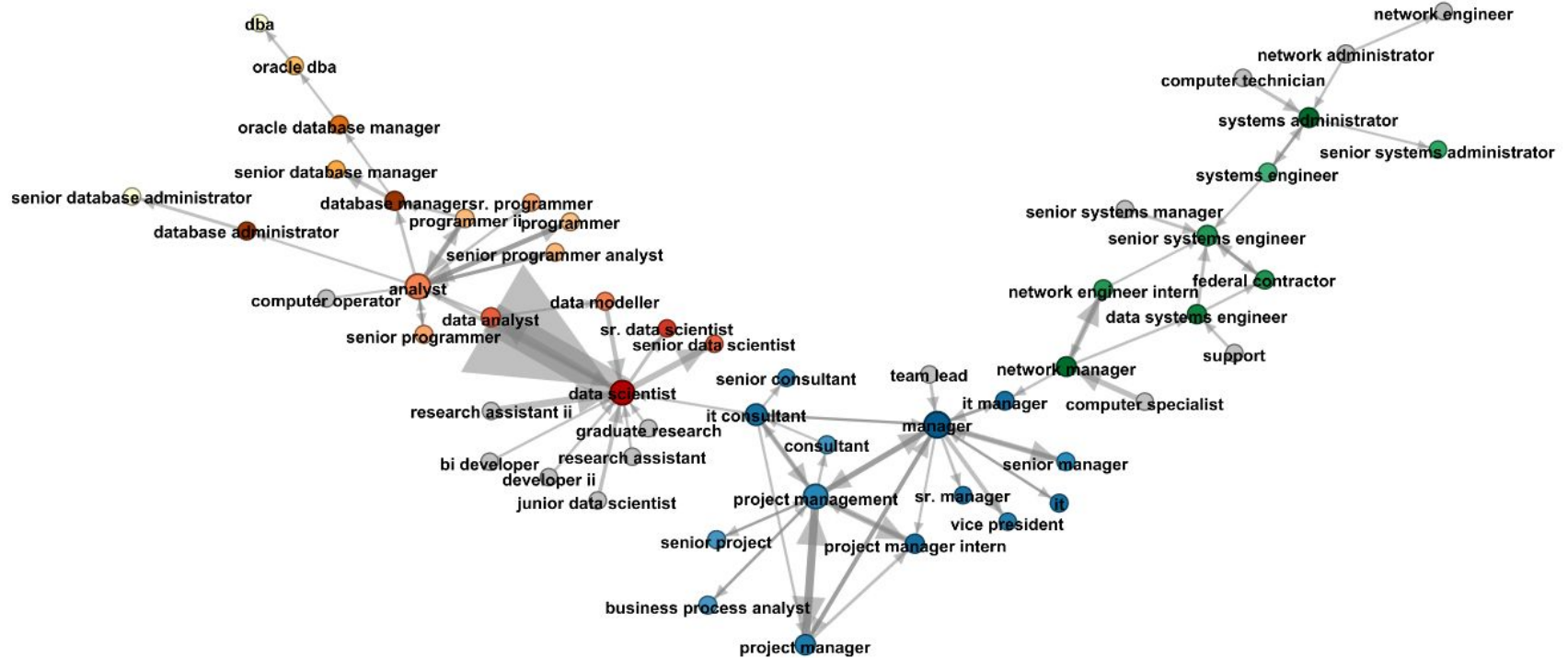
Word2ector

“You shall know a word by the company it keeps.” – J.R.Firth

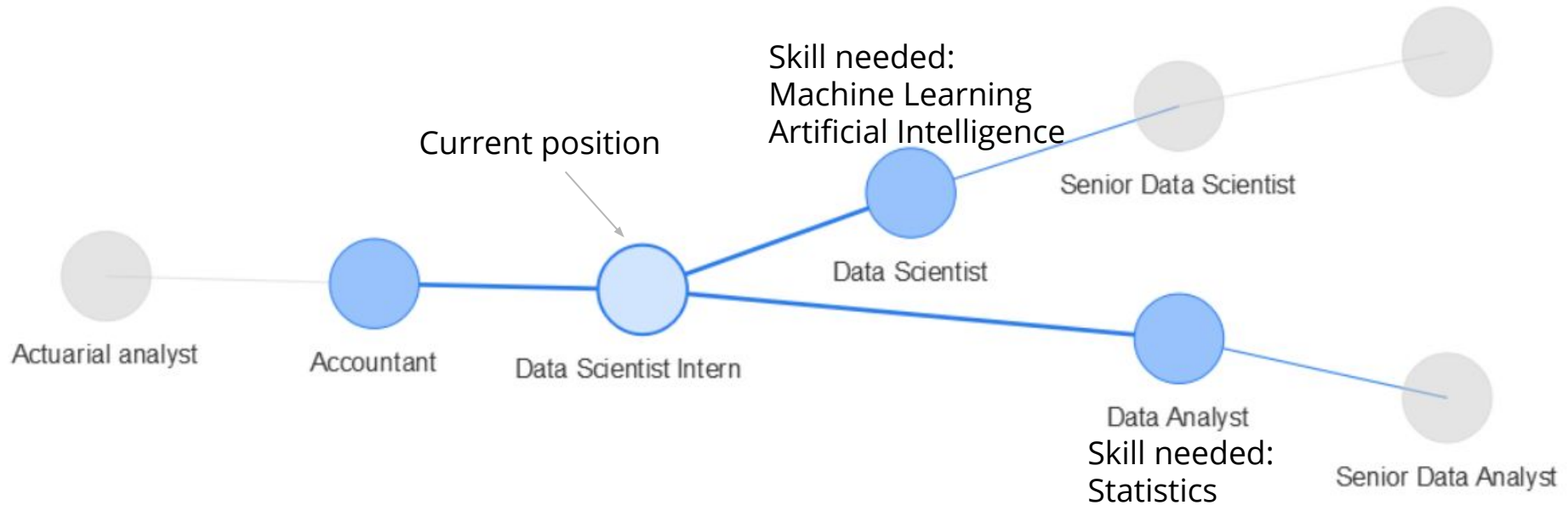
'analysts': 'analyst',
'administrator': 'manager',
'coordinator': 'administrator',
'analsyt': 'analyst',
'manger': 'manager',
'supervisor': 'manager',
'mgr': 'manager',
'director': 'manager',
'sr': 'senior',
'managementanalyst': 'management',
'managementit': 'management',
'planning': 'management',
'oversight': 'management',
'mgmt': 'management',
'control': 'management',
'assurancemanager': 'assurance',
'budget': 'financial',
'tax': 'financial',

'deduction': 'financial',
'program': 'project',
'technician': 'specialist',
'powerbuilder': 'consultant',
'aide': 'assistant',
'assist': 'assistant',
'asst': 'assistant',
'targeting': 'business',
'market': 'business',
'citibusiness': 'business',
'resource': 'resources',
'directors': 'director',
'system': 'systems',
'leader': 'lead',
'application': 'software',
'engineering': 'engineer',
'helpdesk': 'support',
'provided': 'support',

Career Guide (Gephi)



Career Guide (Interactive Function)

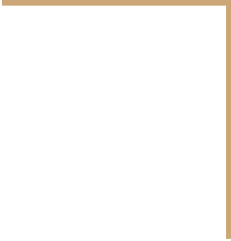


Reference

Den, V. (2017, October 14). Super Fast String Matching in Python. Retrieved from <https://bergvca.github.io/2017/10/14/super-fast-string-matching.html>

WB Advanced Analytics, & WB Advanced Analytics. (2017, July 26). Boosting selection of the most similar entities in large scale datasets. Retrieved from <https://medium.com/wbaa/https-medium-com-ingwbaa-boosting-selection-of-the-most-similar-entities-in-large-scale-datasets-450b3242e618>

Perone, C. S. (n.d.). Machine Learning :: Cosine Similarity for Vector Space Models (Part III). Retrieved from <http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/>



Q&A

Thank you

