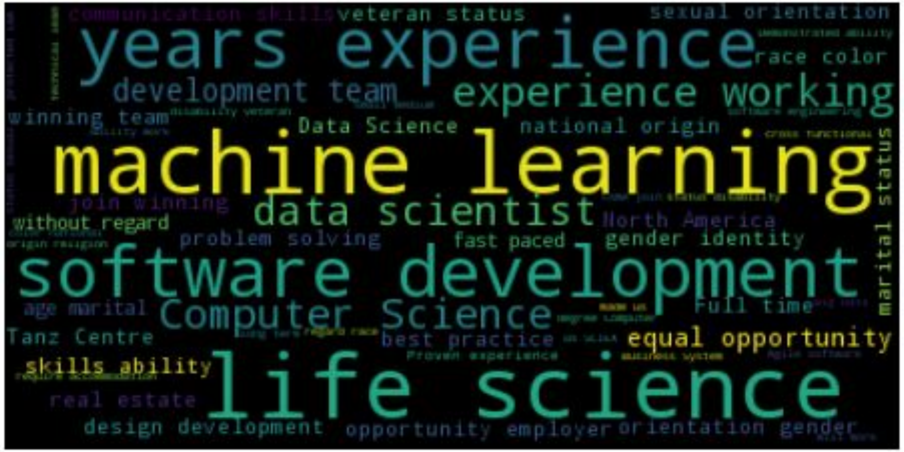
Personalized Learning Analytics Assignment

By Skanda Vaidyanath

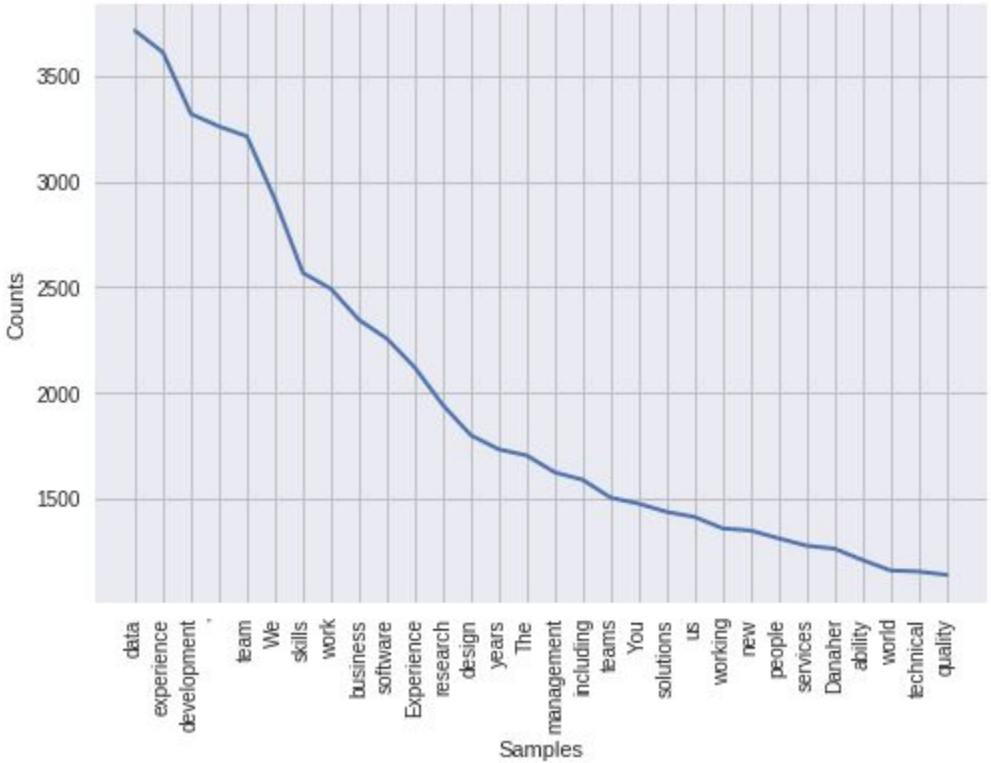
Roadmap of techniques used

1) Exploring the data
2) Basic text searches
3) High level technical and non-technical features
4) Clustering high level features
5) Topic Modelling and LDA
6) Vector Space Model
7) Self Organising Maps
8) Text Classifier
9) Curriculum for the course ‘Data Analytics and AI’

Stage 1 : Exploring the data



Stage 1

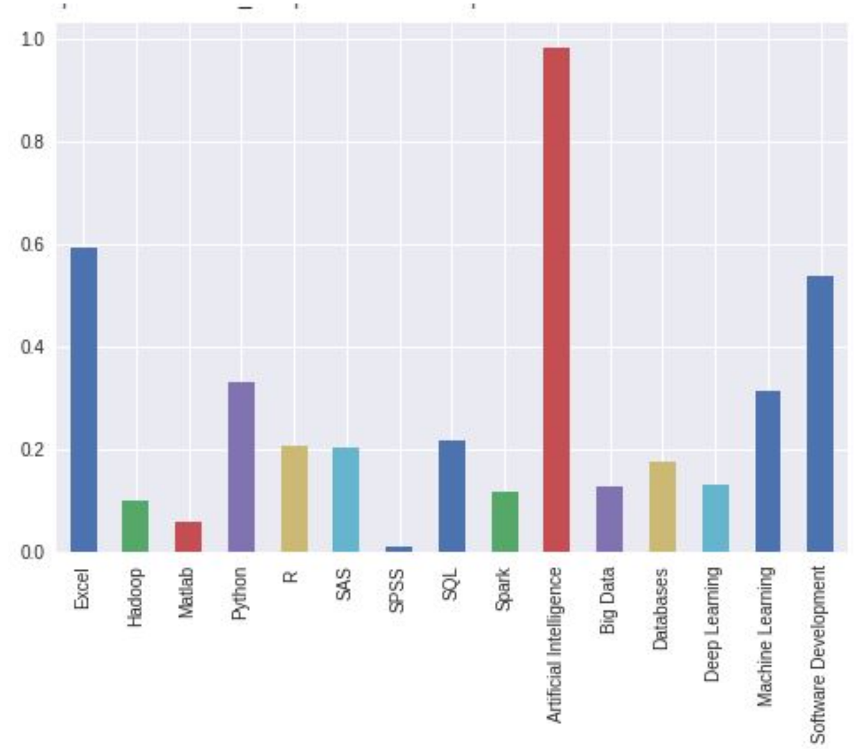


3500
3000
2500
2000
1500
Counts
data
experience
development
,
team
We
skills
work
business
software
Experience
research
design
years
The
management
including
teams
You
solutions
us
working
new
people
services
Danaher
ability
world
technical
quality
Samples

Stage 1

This stage just tells us about the data we are using and the different words that are a part of the Job Description column of the data. We observe some keywords such as ‘Machine Learning’, ‘Software development’, etc.

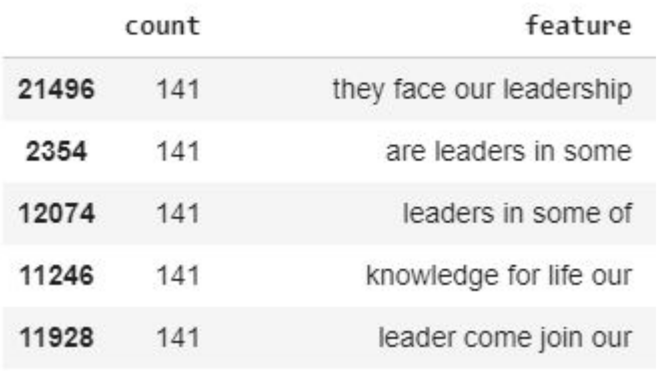
Stage 2 : Basic text searches



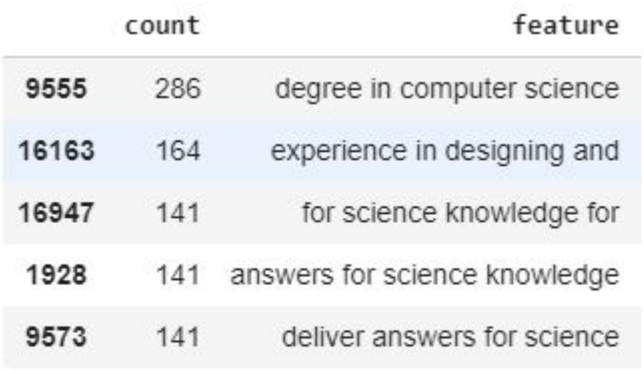
Stage 2

Gives us an idea of what the important skills are from the job postings. Shows that AI is a key skill which is a requirement of nearly all job postings while SPSS is not a very sought after skill. Comparisons between Python and R, SPSS and SAS etc. can also be made.

Stage 3 : High level technical and non-technical features



	count	feature
21496	141	they face our leadership
2354	141	are leaders in some
12074	141	leaders in some of
11246	141	knowledge for life our
11928	141	leader come join our

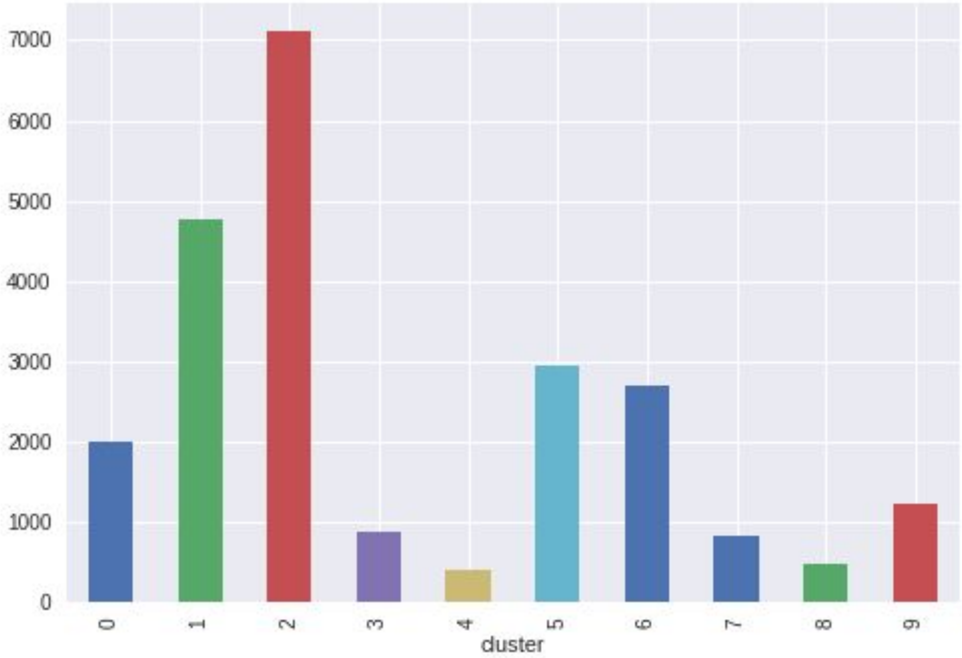


	count	feature
9555	286	degree in computer science
16163	164	experience in designing and
16947	141	for science knowledge for
1928	141	answers for science knowledge
9573	141	deliver answers for science

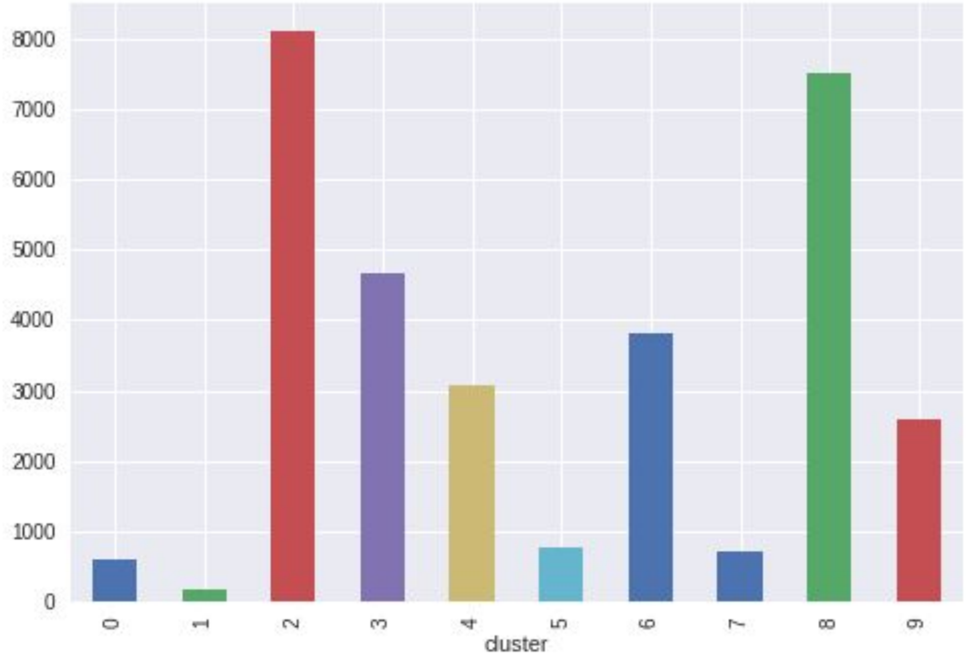
Stage 3

The image on the left in the previous slide shows some high level non-technical features while the image on the right shows some high level technical features.

Stage 4 : Clustering high level features



Stage 4


8000
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
8
9
cluster

Stage 4

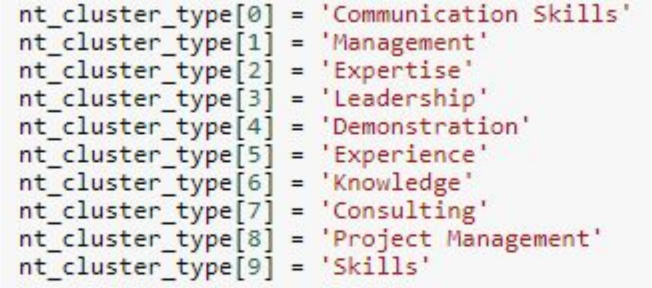
The figure on slide 10 gives cluster densities of non-technical features.

The figure on slide 11 gives cluster densities of technical features.

There are ten clusters of each. The description of each cluster is given in the following slides.

Stage 4

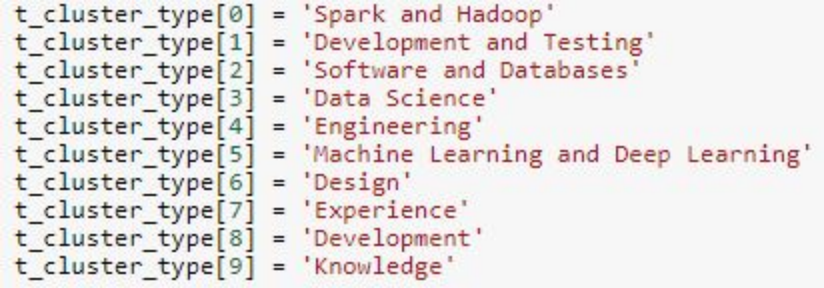
Non-technical features:



```
nt_cluster_type[0] = 'Communication Skills'
nt_cluster_type[1] = 'Management'
nt_cluster_type[2] = 'Expertise'
nt_cluster_type[3] = 'Leadership'
nt_cluster_type[4] = 'Demonstration'
nt_cluster_type[5] = 'Experience'
nt_cluster_type[6] = 'Knowledge'
nt_cluster_type[7] = 'Consulting'
nt_cluster_type[8] = 'Project Management'
nt_cluster_type[9] = 'Skills'
```

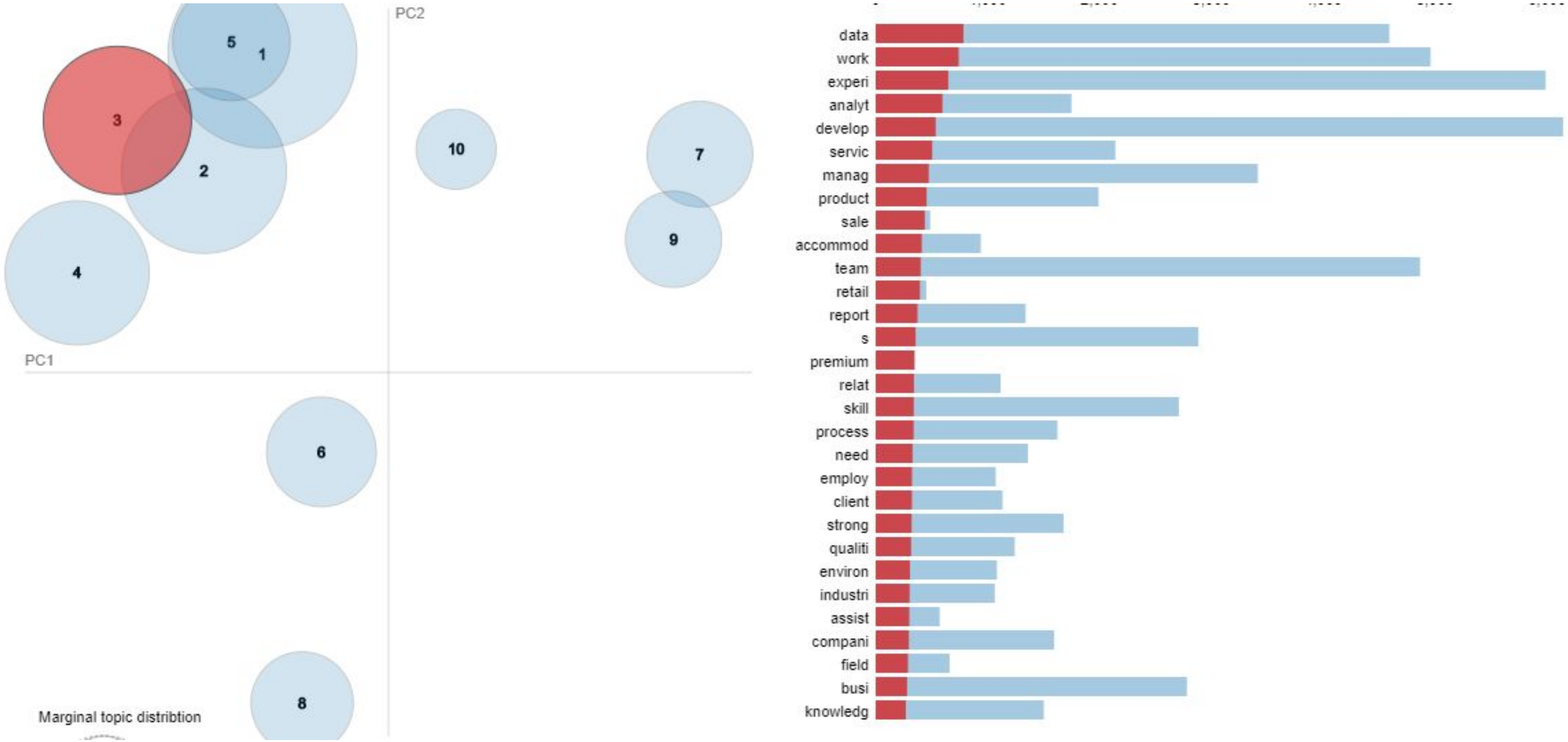
Stage 4

Technical features:



```
t_cluster_type[0] = 'Spark and Hadoop'
t_cluster_type[1] = 'Development and Testing'
t_cluster_type[2] = 'Software and Databases'
t_cluster_type[3] = 'Data Science'
t_cluster_type[4] = 'Engineering'
t_cluster_type[5] = 'Machine Learning and Deep Learning'
t_cluster_type[6] = 'Design'
t_cluster_type[7] = 'Experience'
t_cluster_type[8] = 'Development'
t_cluster_type[9] = 'Knowledge'
```

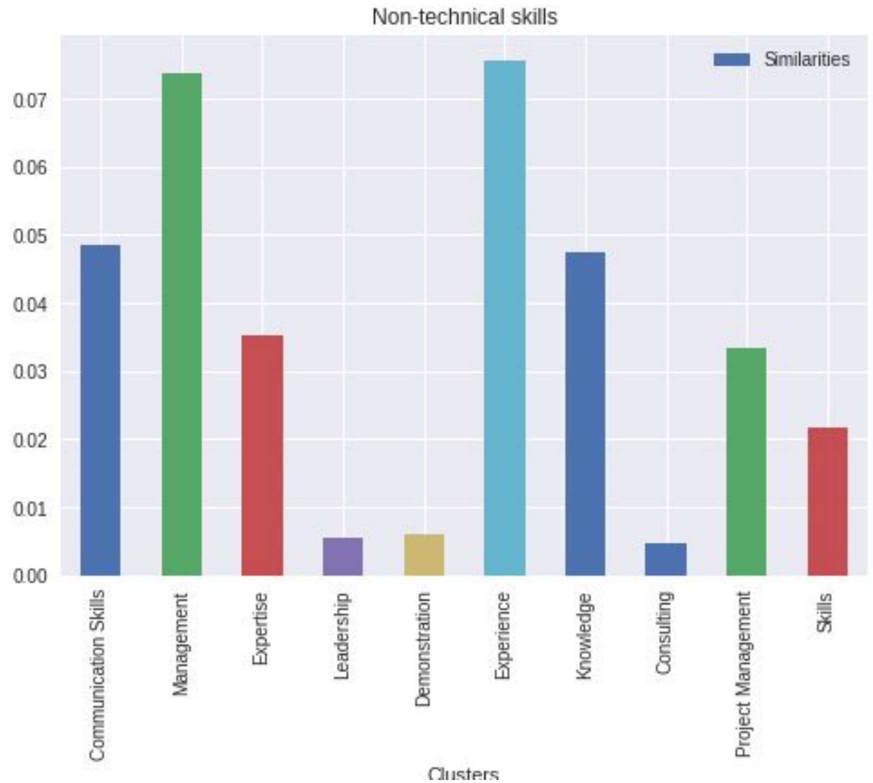
Stage 5 : Topic Modelling and LDA



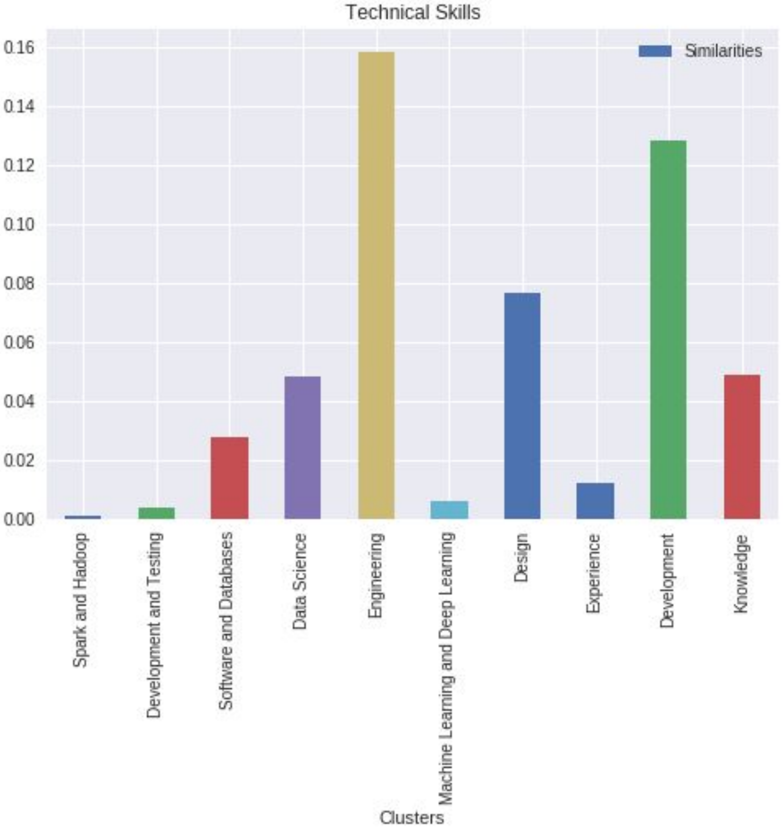
Stage 5

The picture depicts the term distribution for topic number 3 on the right side. On the left side we can see the overall distribution of topics. The animation in the notebook is a great aid for understanding. I have used LDA only for visualisation in this particular assignment. This is mainly because our topic modelling is not very efficient in the sense that it does not spread across all four quadrants (left side of image) and there are many overlaps between topics.

Stage 6 : Vector Space Model



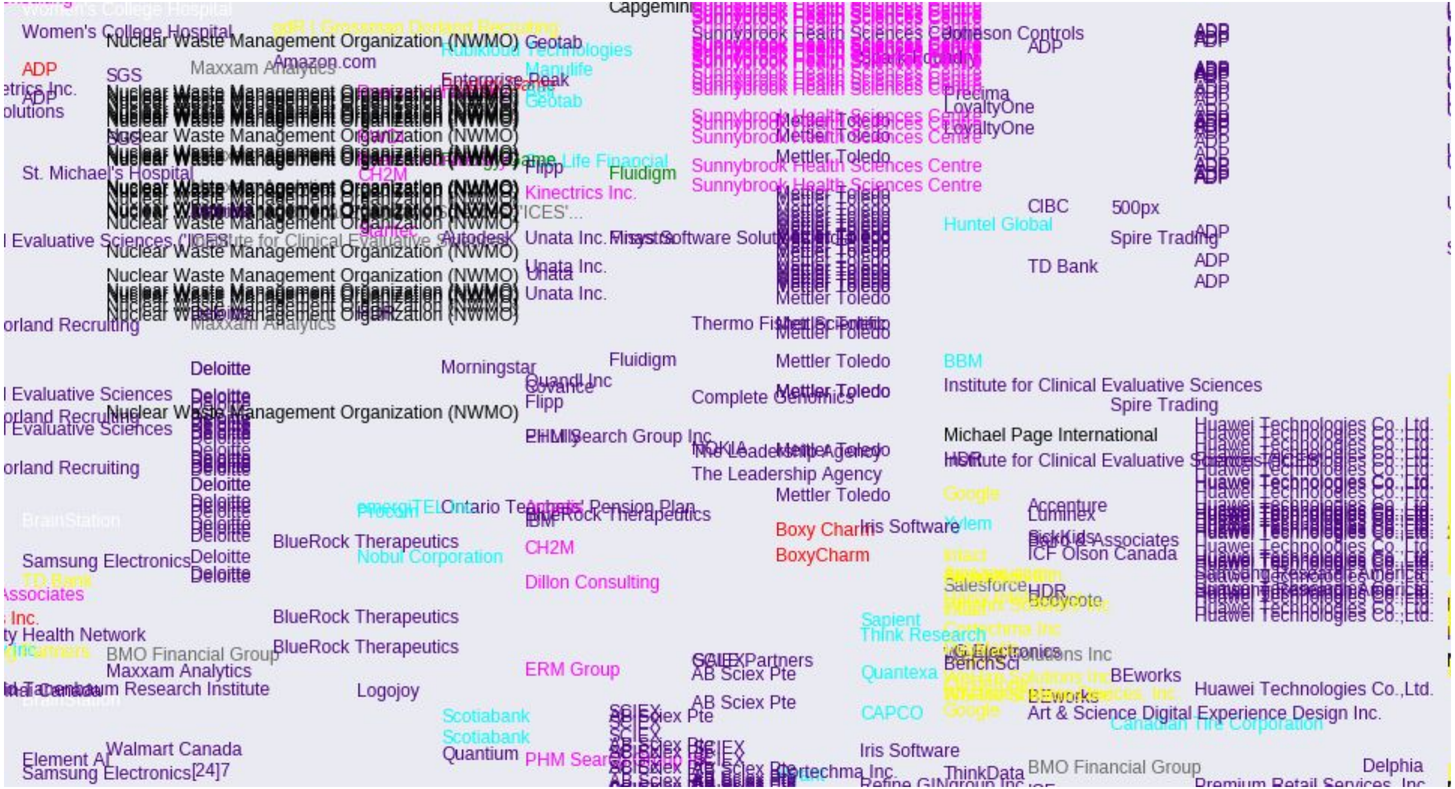
Stage 6


Technical Skills
Similarities
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Spark and Hadoop
Development and Testing
Software and Databases
Data Science
Engineering
Machine Learning and Deep Learning
Design
Experience
Development
Knowledge
Clusters

Stage 6

The figure in slide 17 shows the similarity with each cluster for document number 500. The figure in slide 18 shows the similarity with each cluster for document number 500. For this particular document, the most important non-technical skills are Management and Experience. Leadership and Consulting are not sought after. Similarly for technical skills, Spark and Hadoop and Machine Learning and Deep Learning are not sought after. Development and Design are important technical factors.

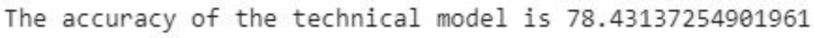
Stage 7 : Self Organising Maps



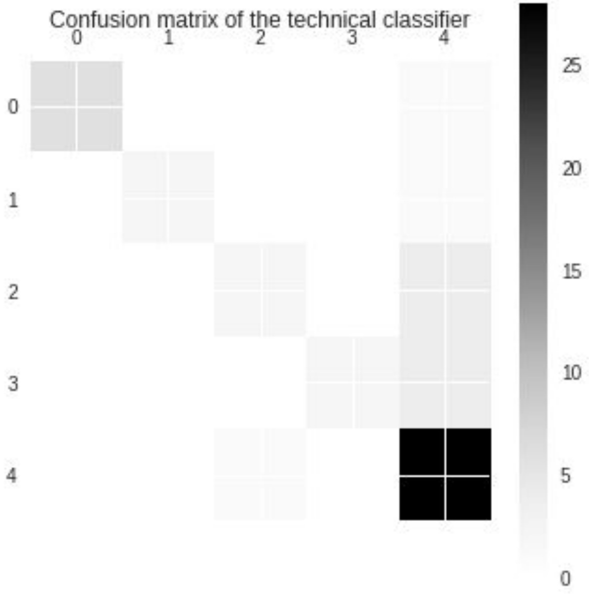
Stage 7

A section of the self organising map output is shown in the previous slide. From the map, we can make inferences about how similar or different the jobs of two companies are. For example, we can say Boxy Charm and Iris Software are offering similar jobs while Deloitte and TDB Bank are not.

Stage 8 : Text Classifier

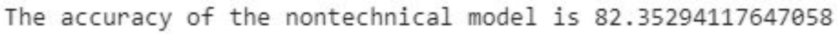


```
The accuracy of the technical model is 78.43137254901961
```

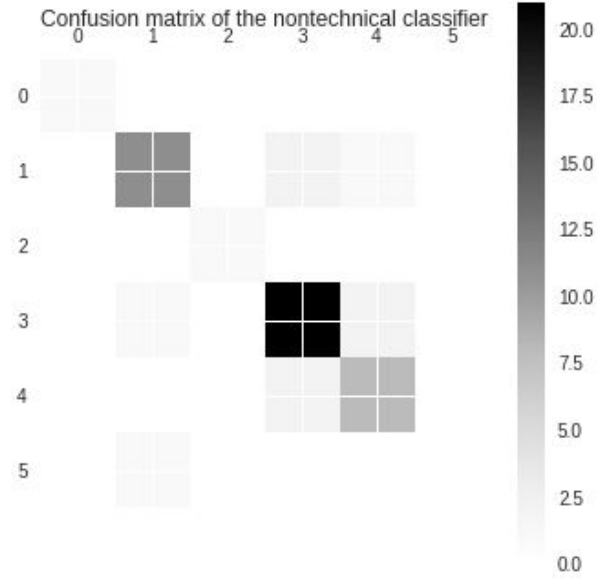


Confusion matrix of the technical classifier

Stage 8



```
The accuracy of the nontechnical model is 82.35294117647058
```



Stage 8

In slide 22, the confusion matrix of the technical text classifier is shown. The accuracy is 78%

In slide 23, the confusion matrix of the non-technical text classifier is shown. The accuracy is 82%

This classifier can be used to classify a new job posting into a technical bucket and a non-technical bucket.

Stage 9 : Design a curriculum for Data Analytics and AI

The details of this section are given in the Python notebook. Please refer the notebook for extensive details.