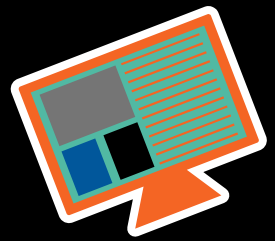
Hacking the Data Science Resume

Garreth Cline



01

02

03

TABLE OF CONTENTS

01 Background

02 Data

03 Models

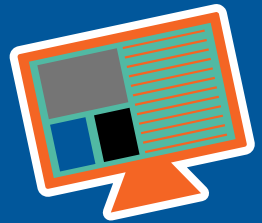
01

02

03

01

BACKGROUND



A LITTLE BACKGROUND

AI GENERATES JOBS

Gartner expects to see 2.3 million jobs starting in 2021

LOTS OF DATA

90% of the world's data was generated within the past two years alone

LOTS OF JOBS

Fifty Thousand Jobs

01

02

03

GOAL

Recommends what are important skills, based on job descriptions.

- Helps the user keep up with new trends
- Helps the user understand differences in job titles

01

02

03

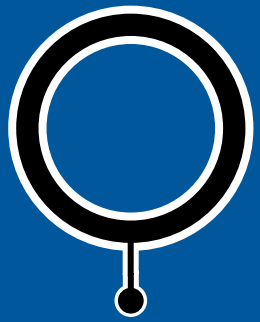
The data I used was scraped from Indeed.com

01

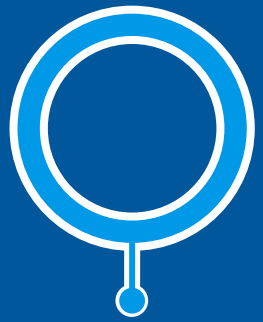
02

03

METHODS USED



SUPERVISED LEARNING



UNSUPERVISED LEARNING

NATURAL LANGUAGE PROCESSING

DATA CLEANING

01

02

03

Features

Features

SQL, Python Ect.

Each one is split and is on its own

Job Description

Used a TfidfVectorizer to make the words into useable data

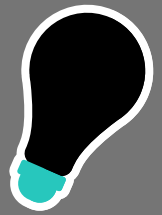
01

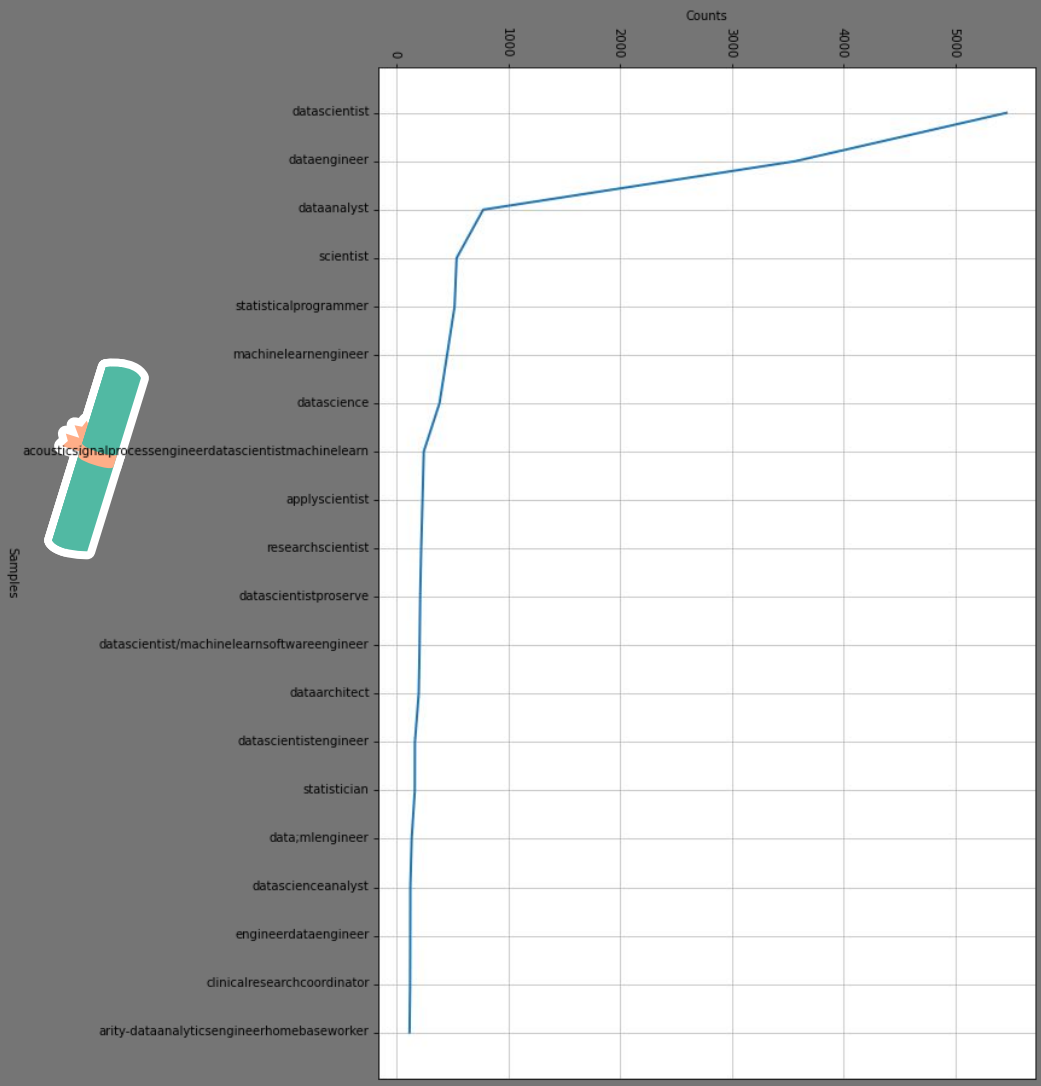
02

03

02

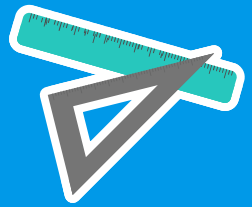
DATA




Counts
0
1000
2000
3000
4000
5000
Samples
datascientist
dataengineer
dataanalyst
scientist
statisticalprogrammer
machinelearnengineer
datascience
acousticsignalprocessengineerdatascientistmachinelearn
applyscientist
researchscientist
datascientistproserve
datascientist/machinelearnsoftwareengineer
dataarchitect
datascientistengineer
statistician
data;mlengineer
datascienceanalyst
engineerdataengineer
clinicalresearchcoordinator
arity-dataanalyticsengineerhomebaseworker

03

THE MODELS



Scores

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

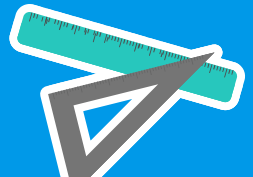
JACCARD SCORE

Jaccard similarity coefficient - size of the intersection divided by the size of the union of two label

$$\frac{1}{|N| \cdot |L|} \sum_{i=1}^{|N|} \sum_{j=1}^{|L|} \operatorname{xor}(y_{i,j}, z_{i,j})$$

HAMMING LOSS

the fraction of the wrong labels to the total number of labels



MODEL PLACEHOLDER

DESC

JACCARD SCORE

HAMMING LOSS

03

| Confusion Matrix |

[[2742 91]
 [78 1028]]

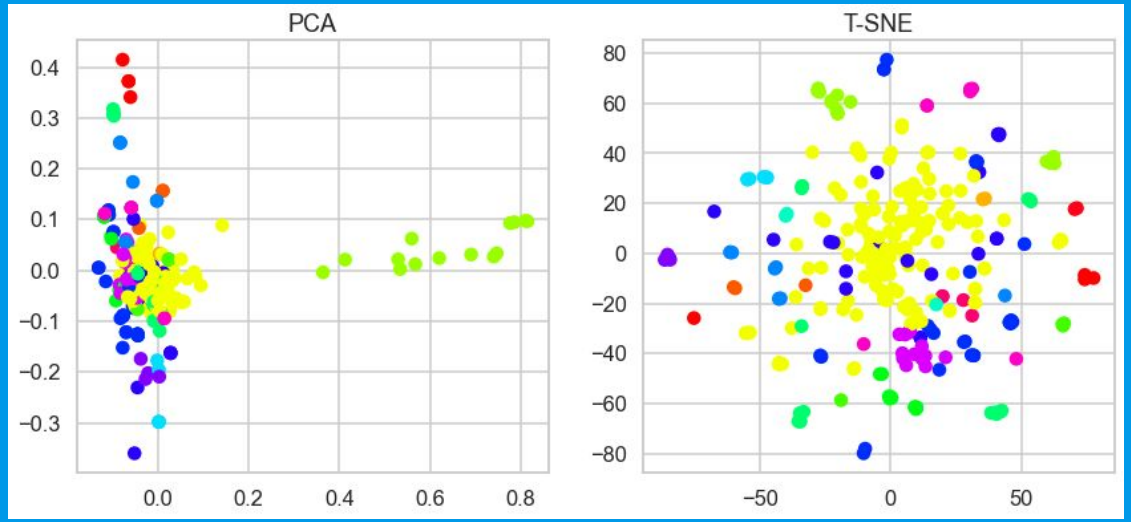
| Precision | Recall | F1 |

Precision Score: 0.95
F1 Score: 0.95
Recall Score: 0.95

| What title words are significant |

datascientist: algorithms vastly link div nine predictive python statistics ser unlimited

CLUSTERING


PCA
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0.0
0.2
0.4
0.6
0.8
T-SNE
80
60
40
20
0
−20
−40
−60
−80
−50
0
50