In [3]:

%matplotlib inline

import pandas as pd

import matplotlib.pyplot as plt

from textblob import TextBlob

import sklearn

import numpy as np

from sklearn.feature\_extraction.text import CountVectorizer

from sklearn.naive\_bayes import MultinomialNB

from sklearn.metrics import classification\_report, f1\_score, accuracy\_score, confusion\_matrix

np.random.seed(47)

In [4]:

reviews = pd.read\_csv('DataScienceJobsResumes.csv', encoding = 'unicode\_escape')

In [6]:

print(reviews.head())

print(reviews.tail())

jobOrResumeDescription role \

0 : Artificial Intelligence / Machine Learning D... Developer

1 : Data Scientist/Architect\n: 6+ months + Hig... Data Scientist

2 : Data Analyst\n: Davidson, NC\n: 04+ Months\... Data Analyst

3 : Big Data Architect or Data Scientist\n: New... Data Scientist

4 : Data Engineer\n: Woonsocket, RI\n: 6+ Months... Data Engineer

sourceType

0 job recruiter

1 job recruiter

2 job recruiter

3 job recruiter

4 job recruiter

jobOrResumeDescription role \

196 \n\n and Experience:\n\* Ten years of experienc... Data Scientist

197 \n \n~ and ~\n \t \n- Python \t- \tMolecula... Data Scientist

198 M.S. Statistics graduate looking for a full-ti... Data Scientist

199 \nIm a data scientist with a background in ne... Data Scientist

200 \n:\n\* Machine learning in R, and other softwa... Data Scientist

sourceType

196 resume

197 resume

198 resume

199 resume

200 resume

In [8]:

print(reviews.shape)

(201, 3)

In [9]:

import regex

def preprocessor(text):

text = regex.sub('<[^>]\*>', '', text)

emoticons = regex.findall('(?::|;|=)(?:-)?(?:\)|\(|D|P)', text)

text = regex.sub('[\W]+', ' ', text.lower()) +\

' '.join(emoticons).replace('-', '')

return text

In [10]:

reviews.tail()

Out[10]:

|  | **jobOrResumeDescription** | **role** | **sourceType** |
| --- | --- | --- | --- |
| 196 | \n\n and Experience:\n\* Ten years of experienc... | Data Scientist | resume |
| 197 | \n \n~ and ~\n \t \n- Python \t- \tMolecula... | Data Scientist | resume |
| 198 | M.S. Statistics graduate looking for a full-ti... | Data Scientist | resume |
| 199 | \nIm a data scientist with a background in ne... | Data Scientist | resume |
| 200 | \n:\n\* Machine learning in R, and other softwa... | Data Scientist | resume |

In [11]:

import numpy as np

reviews = reviews.reindex(np.random.permutation(reviews.index))

print(reviews.head())

print(reviews.tail())

jobOrResumeDescription role \

5 PhD or MS with 3 years post MS experience Comp... Data Scientist

156 \* A data scientist with 7 years of progressiv... Data Scientist

0 : Artificial Intelligence / Machine Learning D... Developer

99 Senior Data Scientist\nKareo Inc2 re - Irvine,... Data Scientist

187 My life experience has been in assistant mana... Data Scientist

sourceType

5 job recruiter

156 resume

0 job recruiter

99 Job post

187 resume

jobOrResumeDescription role \

72 Senior Healthcare Data Analyst\nAdvantmed98 re... Data Analyst

8 Research Computing SME - 19-03563\n \n\n : ... Data Scientist

71 Sr Financial Analyst\nApria Healthcare1,853 re... Data Analyst

134 ? Highly productive and effective Financial An... Data Scientist

135 Business Data Analyst who maximizes productivi... Data Scientist

sourceType

72 Job post

8 job recruiter

71 Job post

134 resume

135 resume

In [12]:

reviews.groupby('sourceType').describe()

Out[12]:

|  | **jobOrResumeDescription** | | | | **role** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **unique** | **top** | **freq** | **count** | **unique** | **top** | **freq** |
| **sourceType** |  |  |  |  |  |  |  |  |
| Job post | 59 | 58 | Senior Data Analyst\nKelly Services14,506 re -... | 2 | 59 | 3 | Data Analyst | 41 |
| job recruiter | 42 | 42 | : Big Data Lead\n\n: Raritan- NJ\n\n: -12 mont... | 1 | 42 | 5 | Data Scientist | 17 |
| resume | 100 | 100 | M.S. Statistics graduate looking for a full-ti... | 1 | 100 | 1 | Data Scientist | 100 |

In [13]:

reviews.groupby('role').describe()

Out[13]:

|  | **jobOrResumeDescription** | | | | **sourceType** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **unique** | **top** | **freq** | **count** | **unique** | **top** | **freq** |
| **role** |  |  |  |  |  |  |  |  |
| Data Analyst | 49 | 48 | Senior Data Analyst\nKelly Services14,506 re -... | 2 | 49 | 2 | Job post | 41 |
| Data Engineer | 13 | 13 | \n and Key : Data Engineer, Python, SQL and U... | 1 | 13 | 2 | job recruiter | 8 |
| Data Scientist | 130 | 130 | \n\* Dedicated IT with around7 years of exper... | 1 | 130 | 3 | resume | 100 |
| Developer | 8 | 8 | : Hadoop Developer\n\n: SFO, CA / Bentonville,... | 1 | 8 | 1 | job recruiter | 8 |
| Software Engineer | 1 | 1 | Ref. : 19-62193\n ... | 1 | 1 | 1 | job recruiter | 1 |

In [14]:

reviews['length'] = reviews['jobOrResumeDescription'].map(lambda text: len(text))

print(reviews.head())

jobOrResumeDescription role \

5 PhD or MS with 3 years post MS experience Comp... Data Scientist

156 \* A data scientist with 7 years of progressiv... Data Scientist

0 : Artificial Intelligence / Machine Learning D... Developer

99 Senior Data Scientist\nKareo Inc2 re - Irvine,... Data Scientist

187 My life experience has been in assistant mana... Data Scientist

sourceType length

5 job recruiter 801

156 resume 22827

0 job recruiter 909

99 Job post 4129

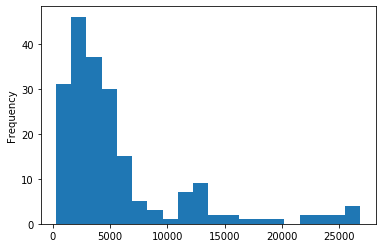
187 resume 4127

In [15]:

reviews.length.plot(bins=20, kind='hist')

Out[15]:

<matplotlib.axes.\_subplots.AxesSubplot at 0x22adc045208>



In [16]:

reviews.length.describe()

Out[16]:

count 201.000000

mean 5694.482587

std 5759.567865

min 303.000000

25% 2202.000000

50% 3873.000000

75% 6120.000000

max 26846.000000

Name: length, dtype: float64

In [18]:

print(list(reviews.jobOrResumeDescription[reviews.length > 40].index))

print(list(reviews.role[reviews.length > 40]))

[5, 156, 0, 99, 187, 74, 128, 36, 160, 141, 132, 20, 63, 108, 16, 133, 25, 158, 96, 9, 1, 148, 62, 172, 15, 32, 112, 164, 45, 10, 154, 162, 168, 3, 13, 31, 183, 92, 82, 121, 192, 182, 26, 178, 193, 130, 190, 131, 83, 78, 40, 115, 69, 102, 37, 139, 149, 165, 47, 85, 189, 35, 106, 42, 81, 46, 18, 107, 6, 155, 39, 64, 76, 77, 61, 188, 151, 70, 24, 111, 38, 91, 195, 67, 94, 54, 44, 11, 127, 119, 33, 170, 150, 200, 143, 145, 136, 117, 57, 29, 19, 116, 22, 126, 104, 58, 177, 171, 122, 197, 60, 88, 129, 191, 174, 110, 98, 105, 43, 49, 14, 194, 118, 103, 97, 167, 93, 30, 95, 181, 41, 152, 142, 17, 109, 120, 157, 138, 28, 100, 56, 12, 140, 185, 186, 146, 199, 124, 89, 176, 153, 175, 166, 144, 2, 75, 52, 84, 51, 90, 159, 86, 68, 125, 147, 21, 27, 161, 79, 137, 101, 163, 4, 66, 34, 73, 48, 53, 184, 169, 196, 180, 123, 87, 65, 114, 50, 55, 113, 173, 198, 7, 59, 23, 80, 179, 72, 8, 71, 134, 135]

['Data Scientist', 'Data Scientist', 'Developer', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Engineer', 'Data Scientist', 'Data Engineer', 'Data Scientist', 'Data Scientist', 'Software Engineer', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Analyst', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Developer', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Engineer', 'Developer', 'Data Scientist', 'Data Engineer', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Developer', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Analyst', 'Data Engineer', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Engineer', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Developer', 'Data Analyst', 'Data Analyst', 'Developer', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Analyst', 'Data Analyst', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Engineer', 'Data Analyst', 'Data Analyst', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Developer', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Engineer', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Engineer', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Engineer', 'Data Scientist', 'Data Analyst', 'Data Engineer', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Analyst', 'Data Analyst', 'Data Analyst', 'Data Analyst', 'Data Engineer', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Developer', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Engineer', 'Data Analyst', 'Data Scientist', 'Data Analyst', 'Data Analyst', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Analyst', 'Data Analyst', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Scientist', 'Data Analyst', 'Data Analyst', 'Data Analyst', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Analyst', 'Data Scientist', 'Data Scientist']

In [19]:

%%time

reviews.hist(column='length', by='role', bins=10)

Wall time: 344 ms

Out[19]:

array([[<matplotlib.axes.\_subplots.AxesSubplot object at 0x0000022ADC11EBE0>,

<matplotlib.axes.\_subplots.AxesSubplot object at 0x0000022ADC16F2E8>],

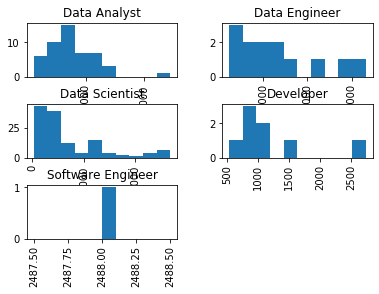
[<matplotlib.axes.\_subplots.AxesSubplot object at 0x0000022ADC19E8D0>,

<matplotlib.axes.\_subplots.AxesSubplot object at 0x0000022ADC1D0E80>],

[<matplotlib.axes.\_subplots.AxesSubplot object at 0x0000022ADC20C470>,

<matplotlib.axes.\_subplots.AxesSubplot object at 0x0000022ADC240A20>]],

dtype=object)



In [26]:

def split\_into\_tokens(review):

#review = unicode(review, 'iso-8859-1')# in python 3 the default of str() previously python2 as unicode() is utf-8

return TextBlob(review).words

In [27]:

reviews.jobOrResumeDescription.head().apply(split\_into\_tokens)

Out[27]:

5 [PhD, or, MS, with, 3, years, post, MS, experi...

156 [A, data, scientist, with, 7, years, of, progr...

0 [Artificial, Intelligence, Machine, Learning, ...

99 [Senior, Data, Scientist, Kareo, Inc2, re, Irv...

187 [My, life, experience, has, been, in, assistan...

Name: jobOrResumeDescription, dtype: object

In [28]:

TextBlob("hello world, how is it going?").tags # list of (word, POS) pairs

Out[28]:

[('hello', 'JJ'),

('world', 'NN'),

('how', 'WRB'),

('is', 'VBZ'),

('it', 'PRP'),

('going', 'VBG')]

In [29]:

import nltk

nltk.download('stopwords')

[nltk\_data] Downloading package stopwords to

[nltk\_data] C:\Users\m\AppData\Roaming\nltk\_data...

[nltk\_data] Package stopwords is already up-to-date!

Out[29]:

True

In [30]:

from nltk.corpus import stopwords

stop = stopwords.words('english')

stop = stop + [u'a',u'b',u'c',u'd',u'e',u'f',u'g',u'h',u'i',u'j',u'k',u'l',u'm',u'n',u'o',u'p',u'q',u'r',u's',u't',u'v',u'w',u'x',u'y',u'z']

In [32]:

def split\_into\_lemmas(review):

#review = unicode(review, 'iso-8859-1')

review = review.lower()

#review = unicode(review, 'utf8').lower()

#review = str(review).lower()

words = TextBlob(review).words

# for each word, take its "base form" = lemma

return [word.lemma for word in words if word not in stop]

reviews.jobOrResumeDescription.head().apply(split\_into\_lemmas)

Out[32]:

5 [phd, m, 3, year, post, m, experience, compute...

156 [data, scientist, 7, year, progressive, experi...

0 [artificial, intelligence, machine, learning, ...

99 [senior, data, scientist, kareo, inc2, irvine,...

187 [life, experience, assistant, manager, managem...

Name: jobOrResumeDescription, dtype: object

In [33]:

%%time

bow\_transformer = CountVectorizer(analyzer=split\_into\_lemmas).fit(reviews['jobOrResumeDescription'])

print(len(bow\_transformer.vocabulary\_))

10625

Wall time: 7.36 s

In [34]:

review4 = reviews['jobOrResumeDescription'][42]

print(review4)

At TGS, data is a core part of our business and our data focused software developers are among our most valued resources. This is mission critical, so we?re for an uncommonly reliable professional who enjoys coding, ing with and analyzing data, and providing support for production systems. In addition to programming and data analysis, this is likely to involve interaction with external resources such as data providers, brokers, dealers, and software vendors.

Successful candis will have experience in a number of the following general areas:

Programming: strong experience with Python, Java, SQL, and/or similar languages

Large data sets: experience developing programs to parse, process, clean, organize, and analyze large data sets

Applications: experience designing, developing, and maintaining software applications

Vendor interaction: ing with external resources to solve problems, acquire data, and improve relationships.

System : experience with scripting languages (Perl, Python, shell scripts, etc.), and Unix-based operating systems (especially Linux)

Indeed Hire and TGS Management Company, LLC are ing together to find the best candi for this .

By ing, you agree to be contacted by our agent, Indeed Hire, and receive ups via text and phone about your application.

Type: Full-time

Experience:

Python, Java, SQL, and/or similar languages: 1 year ()

designing, developing, and maintaining software applications: 1 year ()

Large data sets: 2 years ()

distance:

Irvine, CA: Between 31 and 40 miles ()

In [35]:

bow4 = bow\_transformer.transform([review4])

print(bow4)

(0, 38) 2

(0, 152) 1

(0, 311) 1

(0, 345) 1

(0, 651) 1

(0, 698) 1

(0, 791) 1

(0, 809) 1

(0, 934) 1

(0, 945) 1

(0, 970) 1

(0, 977) 1

(0, 982) 2

(0, 1136) 4

(0, 1195) 1

(0, 1548) 1

(0, 1742) 1

(0, 1785) 1

(0, 1827) 1

(0, 1873) 1

(0, 1875) 1

(0, 2137) 1

(0, 2224) 1

(0, 2302) 1

(0, 2478) 1

: :

(0, 8047) 1

(0, 8149) 3

(0, 8515) 1

(0, 8518) 1

(0, 8668) 3

(0, 8718) 1

(0, 8767) 2

(0, 8878) 4

(0, 8900) 1

(0, 9017) 2

(0, 9207) 1

(0, 9265) 1

(0, 9323) 1

(0, 9389) 3

(0, 9563) 1

(0, 9576) 2

(0, 9734) 1

(0, 9940) 1

(0, 9983) 1

(0, 10035) 1

(0, 10084) 1

(0, 10180) 1

(0, 10219) 2

(0, 10251) 1

(0, 10561) 3

In [36]:

%%time

reviews\_bow = bow\_transformer.transform(reviews['jobOrResumeDescription'])

print('sparse matrix shape:', reviews\_bow.shape)

print('number of non-zeros:', reviews\_bow.nnz)

print('sparsity: %.2f%%' % (100.0 \* reviews\_bow.nnz / (reviews\_bow.shape[0] \* reviews\_bow.shape[1])))

sparse matrix shape: (201, 10625)

number of non-zeros: 61674

sparsity: 2.89%

Wall time: 6.97 s

In [37]:

# Split/splice into training ~ 80% and testing ~ 20%

reviews\_bow\_train = reviews\_bow[:155]

reviews\_bow\_test = reviews\_bow[155:]

reviews\_sentiment\_train = reviews['role'][:155]

reviews\_sentiment\_test = reviews['role'][155:]

print(reviews\_bow\_train.shape)

print(reviews\_bow\_test.shape)

(155, 10625)

(46, 10625)

In [38]:

%time review\_sentiment = MultinomialNB().fit(reviews\_bow\_train, reviews\_sentiment\_train)

Wall time: 438 ms

In [41]:

print('predicted:', review\_sentiment.predict(bow4)[0])

print('expected:', reviews.role[42])

predicted: Data Analyst

expected: Developer

In [42]:

predictions = review\_sentiment.predict(reviews\_bow\_test)

print(predictions)

['Data Analyst' 'Data Analyst' 'Data Analyst' 'Data Analyst'

'Data Analyst' 'Data Scientist' 'Data Scientist' 'Data Analyst'

'Data Scientist' 'Data Scientist' 'Data Scientist' 'Data Scientist'

'Data Scientist' 'Data Analyst' 'Data Scientist' 'Data Scientist'

'Data Scientist' 'Data Scientist' 'Data Analyst' 'Data Scientist'

'Data Analyst' 'Data Analyst' 'Data Analyst' 'Data Scientist'

'Data Scientist' 'Data Scientist' 'Data Scientist' 'Data Scientist'

'Data Scientist' 'Data Analyst' 'Data Scientist' 'Data Analyst'

'Data Analyst' 'Data Scientist' 'Data Scientist' 'Data Scientist'

'Data Scientist' 'Data Analyst' 'Data Analyst' 'Data Analyst'

'Data Scientist' 'Data Analyst' 'Data Scientist' 'Data Analyst'

'Data Scientist' 'Data Scientist']

In [43]:

print('accuracy', accuracy\_score(reviews\_sentiment\_test, predictions))

print('confusion matrix\n', confusion\_matrix(reviews\_sentiment\_test, predictions))

print('(row=expected, col=predicted)')

accuracy 0.9347826086956522

confusion matrix

[[18 0 0 0]

[ 1 0 1 0]

[ 0 0 25 0]

[ 0 0 1 0]]

(row=expected, col=predicted)

In [44]:

print(classification\_report(reviews\_sentiment\_test, predictions))

#The F1 score can be interpreted as a weighted average of the precision and recall,

#where an F1 score reaches its best value at 1 and worst score at 0.

precision recall f1-score support

Data Analyst 0.95 1.00 0.97 18

Data Engineer 0.00 0.00 0.00 2

Data Scientist 0.93 1.00 0.96 25

Developer 0.00 0.00 0.00 1

accuracy 0.93 46

macro avg 0.47 0.50 0.48 46

weighted avg 0.87 0.93 0.90 46

c:\users\m\anaconda2\envs\python36\lib\site-packages\sklearn\metrics\classification.py:1437: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples.

'precision', 'predicted', average, warn\_for)

In [46]:

def predict\_review(new\_review):

new\_sample = bow\_transformer.transform([new\_review])

print(new\_review, '\nThe respective order new or pro probabilities:\n',np.around(review\_sentiment.predict\_proba(new\_sample), decimals=2),'\n')

In [51]:

reviews.role.unique()

#These will be alphabatized in the probabilities returned for the 5 categories.

#1-data analyst, 2-data engineer, 3-data scientist, 4-developer, 5-software engineer

Out[51]:

array(['Data Scientist', 'Developer', 'Data Analyst', 'Data Engineer',

'Software Engineer'], dtype=object)

In [48]:

predict\_review('R. programming. years. RDD. Python. Spark.')

predict\_review('AWS. Azure. php. excel')

predict\_review('access. microsoft. MS. office. sharepoint')

predict\_review('analyst. tutoring. teaching. Assistant.')

R. programming. years. RDD. Python. Spark.

The respective order new or pro probabilities:

[[0. 0. 1. 0. 0.]]

AWS. Azure. php. excel

The respective order new or pro probabilities:

[[0. 0. 1. 0. 0.]]

access. microsoft. MS. office. sharepoint

The respective order new or pro probabilities:

[[0.01 0. 0.99 0. 0. ]]

analyst. tutoring. teaching. Assistant.

The respective order new or pro probabilities:

[[0.02 0.01 0.96 0. 0. ]]

Most of the above are Data Scientists, but there are some percentage probabilties of being a job posting or resume for data analyst and data engineer.

In [ ]: