```
%python
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
```

```
%python
# The path to our CSV file
file = "../Resources/StackOverflow_ML.csv"
```

```
%python
# Read our Kickstarter data into pandas
df = pd.read_csv(file)
df.head()
```

	guid/_isPermaLink	guid/__text	link	author/name/__text	category/0	category/1	category/2	category/3	category/4	title	
0	False	295343	https://stackoverflow.com/jobs/295343/machine-...	Innate, Inc.	machine-learning	algorithm	mongodb	sql-server	amazon-web-services	Machine Learning Engineer at Innate, Inc. (Was...	<p>We are look Learning (ML) ..
1	False	353959	https://stackoverflow.com/jobs/353959/software...	Facebook	java	c++	hadoop	NaN	NaN	Software Engineer, Machine Learning at Faceboo...	<p>Facebook's people the po...
2	False	206002	https://stackoverflow.com/jobs/206002/software...	Facebook	java	hadoop	hbase	mapreduce	NaN	Software Engineer, Machine Learning at Faceboo...	<p>Facebook's people the po...
3	False	206001	https://stackoverflow.com/jobs/206001/software...	Facebook	java	hadoop	mapreduce	NaN	NaN	Software Engineer, Machine Learning at Faceboo...	<p>Facebook's people the po...
4	False	334659	https://stackoverflow.com/jobs/334659/machine-...	Research Square	python	mysql	php	google-cloud-platform	git	Machine Learning Software Engineer (Remote, US...	<p><a href="https://ww r...

```
%python
# Get a list of all of our columns for easy reference
df.columns
```

```
Index(['guid/_isPermaLink', 'guid/__text', 'link', 'author/name/__text',
       'category/0', 'category/1', 'category/2', 'category/3', 'category/4',
       'title', 'description', 'pubDate', 'updated/__text', 'location/_xmlns',
       'location/__text', 'category/5'],
      dtype='object')
```

```
%python
# Extract 'author/name/__text','category/0', 'category/1', 'category/2', 'category/3', 'category/4',
#'title', 'description', 'pubDate','location/__text',

reduced_so_ml_df = df.loc[:, ["author/name/__text","category/0", "category/1", "category/2", "category/3", "category/4",
"title", "description", "pubDate","location/__text"]]
reduced_so_ml_df
```

	author/name/__text	category/0	category/1	category/2	category/3	category/4	title	description	pubDate	location/__text
0	Innate, Inc.	machine-learning	algorithm	mongodb	sql-server	amazon-web-services	Machine Learning Engineer at Innate, Inc. (Was...	<p>We are looking for a Machine Learning (ML) ...	Fri, 07 Feb 2020 18:52:39 Z	Washington, DC
1	Facebook	java	c++	hadoop	NaN	NaN	Software Engineer, Machine Learning at Faceboo...	<p>Facebook's mission is to give people the po...	Wed, 12 Feb 2020 17:05:10 Z	New York, NY
2	Facebook	java	hadoop	hbase	mapreduce	NaN	Software Engineer, Machine Learning at Faceboo...	<p>Facebook's mission is to give people the po...	Wed, 05 Feb 2020 11:35:27 Z	Boston, MA
3	Facebook	java	hadoop	mapreduce	NaN	NaN	Software Engineer, Machine	<p>Facebook's mission is to give	Wed, 05 Feb 2020	Bellevue, WA

3	Facebook	java	hadoop	mapreduce	NaN	NaN	Learning at Faceboo...	people the po...	2020 11:35:27 Z	Bellevue, WA
4	Research Square	python	mysql	php	google-cloud-platform	git	Machine Learning Software Engineer (Remote, US...	<p><a href="https://www.researchsquare.com/" r...	Tue, 04 Feb 2020 19:05:01 Z	Durham, NC
...	...	...	...	...	...	...	...	...	...	...
507	Facebook	web-services	c++	python	NaN	NaN	Production Engineer at Facebook (Boston, MA)	<p>Facebook's mission is to give people the po...	Wed, 22 Jan 2020 20:50:24 Z	Boston, MA
508	Facebook	r	sql	hadoop	NaN	NaN	Data Science Manager, Analytics Ads & Business...	<p>Facebook's mission is to give people the po...	Wed, 22 Jan 2020 20:50:57 Z	Menlo Park, CA
509	MasterPeace Solutions	user-experience	css	photoshop	adobe-illustrator	html	User Experience (UX) Designer Fully Cleared at...	<p>MasterPeace Solutions is seeking a ...	Thu, 16 Jan 2020 14:43:47 Z	Annapolis Junction, MD
510	Facebook	security	NaN	NaN	NaN	NaN	Privacy Program Manager, Product at Facebook (...	<p>Facebook's mission is to give people the po...	Wed, 05 Feb 2020 11:35:24 Z	Menlo Park, CA
511	Facebook	web-services	java	spring	swing	NaN	Manager, Production Engineering at Facebook (B...	<p>Facebook's mission is to give people the po...	Thu, 06 Feb 2020 01:25:01 Z	Boston, MA

512 rows × 10 columns

```
%python
reduced_so_ml_df=reduced_so_ml_df.rename(columns={"author/name/__text": "company", "location/__text": "location"})
```

```
%python
print("\nSplitting 'location' column into two different columns :\n",
                                    reduced_so_ml_df["location"].str.split(",",expand=True))
```

```
Splitting 'location' column into two different columns :
                       0    1
0             Washington   DC
1               New York   NY
2                 Boston   MA
3               Bellevue   WA
4                 Durham   NC
..                   ...  ...
507               Boston   MA
508           Menlo Park   CA
509  Annapolis Junction   MD
510           Menlo Park   CA
511               Boston   MA

[512 rows x 2 columns]
```

```
%python
reduced_so_ml_df[['city','state']]=reduced_so_ml_df["location"].str.split(",",expand=True)
```

```
%python
reduced_so_ml_df
```

	company	category/0	category/1	category/2	category/3	category/4	title	description	pubDate	location	city	state
0	Innate, Inc.	machine-learning	algorithm	mongodb	sql-server	amazon-web-services	Machine Learning Engineer at Innate, Inc. (Was...	<p>We are looking for a Machine Learning (ML) ...	Fri, 07 Feb 2020 18:52:39 Z	Washington, DC	Washington	DC
1	Facebook	java	c++	hadoop	NaN	NaN	Software Engineer, Machine Learning at Faceboo...	<p>Facebook's mission is to give people the po...	Wed, 12 Feb 2020 17:05:10 Z	New York, NY	New York	NY
2	Facebook	java	hadoop	hbase	mapreduce	NaN	Software Engineer, Machine Learning at Faceboo...	<p>Facebook's mission is to give people the po...	Wed, 05 Feb 2020 11:35:27 Z	Boston, MA	Boston	MA
3	Facebook	java	hadoop	mapreduce	NaN	NaN	Software Engineer, Machine Learning at Faceboo...	<p>Facebook's mission is to give people the po...	Wed, 05 Feb 2020 11:35:27 Z	Bellevue, WA	Bellevue	WA
4	Research Square	python	mysql	php	google-cloud-platform	git	Machine Learning Software Engineer (Remote, US...	<p><a href="https://www.researchsquare.com/" r...	Tue, 04 Feb 2020 19:05:01 Z	Durham, NC	Durham	NC
...	...	...	...	...	...	...	...	...	...	...	...	...
507	Facebook	web-services	c++	python	NaN	NaN	Production Engineer at Facebook (Boston, MA)	<p>Facebook's mission is to give people the po...	Wed, 22 Jan 2020 20:50:24 Z	Boston, MA	Boston	MA
							Data Science Manager,		Wed, 22			

508	Facebook	r	sql	hadoop	NaN	NaN	Data Science Manager, Analytics Ads & Business...	<p>Facebook's mission is to give people the po...	Wed, 22 Jan 2020 20:50:57 Z	Menlo Park, CA	Menlo Park	CA
509	MasterPeace Solutions	user-experience	css	photoshop	adobe-illustrator	html	User Experience (UX) Designer Fully Cleared at...	<p>MasterPeace Solutions is seeking a ...	Thu, 16 Jan 2020 14:43:47 Z	Annapolis Junction, MD	Annapolis Junction	MD
510	Facebook	security	NaN	NaN	NaN	NaN	Privacy Program Manager, Product at Facebook (...	<p>Facebook's mission is to give people the po...	Wed, 05 Feb 2020 11:35:24 Z	Menlo Park, CA	Menlo Park	CA
511	Facebook	web-services	java	spring	swing	NaN	Manager, Production Engineering at Facebook (B...	<p>Facebook's mission is to give people the po...	Thu, 06 Feb 2020 01:25:01 Z	Boston, MA	Boston	MA

512 rows × 12 columns

```
%python
state_groups =reduced_so_ml_df.groupby("state")
state_groups["city"].count()
```

```
state
 AR      1
 AZ      1
 CA    221
 CO      8
 DC     15
 FL      1
 GA      3
 IA      1
 ID      3
 IL     20
 IN      1
 KY      1
 MA     28
 MD     12
 MI      3
 MO      1
 NC      7
 NJ      8
 NM      1
 NV      1
 NY     52
 OH      9
 OR      3
 PA     11
 RI      2
 SC      1
 TN      1
 TX     18
 UK     14
```

```
%python
#plt.bar(state_groups.index, job_cnt, color='r', alpha=0.5, align="center")
job_count=state_groups.count()
job_cnt=job_count['city']
```

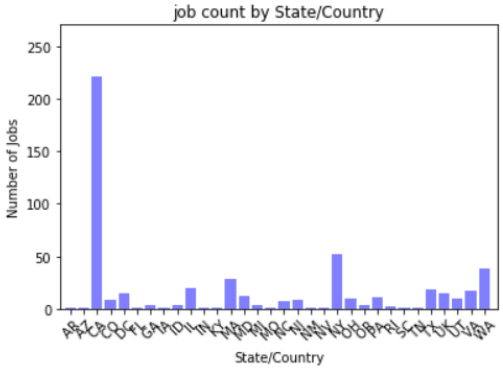
```
%python
plt.bar(job_count.index, job_count['city'], color='b', alpha=0.5, align="center")

plt.xlim(-0.75, len(job_count.index))
plt.ylim(0, max(job_count['city'])+50)
plt.title("job count by State/Country")
plt.xlabel("State/Country")
plt.ylabel("Number of Jobs")

plt.xticks(rotation=45)
```

```
([0,
  1,
  2,
  3,
  4,
  5,
  6,
  7,
  8,
  9,
  10,
  11,
  12,
  13,
  14,
  15,
  16,
  17,
  18,
  19,
  20,
  21,
  22,
  23,
  24,
```

```
24,
25,
26,
27,
28,
29,
```




```
%python
df=reduced_so_ml_df.loc[:, ["company","title", "description", "city","state"]]
df.to_csv("../Resources/clean_data.csv",index=False)
```

```
%pyspark
# Read in data from S3 Buckets
from pyspark import SparkFiles
url ="https://ucb-bhumikasharma.s3-us-west-1.amazonaws.com/clean_data.csv"
spark.sparkContext.addFile(url)
start_data = spark.read.csv(SparkFiles.get("clean_data.csv"), sep=",", header=True)

# Show DataFrame
start_data.show()
```

```
+--------+--------------------+--------------------+--------------------+--------------------+--------------------+
|   class|             company|               title|         description|                city|               state|
+--------+--------------------+--------------------+--------------------+--------------------+--------------------+
|positive|        Innate, Inc.|Machine Learning ...|<p>We are looking...|          Washington|                  DC|
|positive|            Facebook|Software Engineer...|<p>Facebook's mis...|            New York|                  NY|
|positive|            Facebook|Software Engineer...|<p>Facebook's mis...|              Boston|                  MA|
|positive|            Facebook|Software Engineer...|<p>Facebook's mis...|             Bellevue|                  WA|
|positive|     Research Square|Machine Learning ...|"<p><a href=""htt...| which provides s...|         hardworking|
|positive|   Rebellion Defense|Senior Software E...|As a Senior Softw...|          Washington|                  DC|
|positive|        7 Chord inc.|Data Engineer  Al...|This is an except...|            Brooklyn|                  NY|
|positive|Applied Research ...|Software Engineer...|"<p><u>Title</u>:...|          and design|         development|
|positive|        Carbon Relay|Senior Machine Le...|"<p>Carbon Relay ...| tackles hard pro...| and has fun!</p>...|
|positive|            Facebook|Data Engineer, Ma...|<p>Facebook's mis...|             Fremont|                  CA|
|positive|            Facebook|Software Engineer...|<p>Facebook's mis...|          Menlo Park|                  CA|
|positive|        OmniSci, Inc|Software Engineer...|<p>Our mission at...|       San Francisco|                  CA|
|positive|            Facebook|Instagram - Softw...|<p>Facebook's mis...|            New York|                  NY|
|positive|         PlayStation|Senior Software E...|<p><strong>Senior...|           San Mateo|                  CA|
|positive|Sailpoint Technol...|Software Engineer...|<p>At SailPoint, ...|              Austin|                  TX|
|positive|Eastman Chemical ...|Machine Learning ...|<p><strong>Summar...|           Kingsport|                  TN|
|positive|           StyleSeat|Senior Machine Le...|<p><strong>Senior...|       San Francisco|                  CA|
|positive|               Apple|SW Engineer, Mach...|Want to own the p...|             Seattle|                  WA|
|positive|               Apple|Machine Learning ...|Have you ever sch...|           Cupertino|                  CA|
|positive|               Apple|Siri - Machine Le...|The Siri Client G...|             Seattle|                  WA|
+--------+--------------------+--------------------+--------------------+--------------------+--------------------+
only showing top 20 rows
```

```
%pyspark
from pyspark.sql.functions import length
# Create a length column to be used as a future feature
data_df = start_data.withColumn('length', length(start_data['description']))
data_df.show()
```

```
+--------+--------------------+--------------------+--------------------+--------------------+--------------------+------+
|   class|             company|               title|         description|                city|               state|length|
+--------+--------------------+--------------------+--------------------+--------------------+--------------------+------+
|positive|        Innate, Inc.|Machine Learning ...|<p>We are looking...|          Washington|                  DC|  1839|
|positive|            Facebook|Software Engineer...|<p>Facebook's mis...|            New York|                  NY|  3353|
|positive|            Facebook|Software Engineer...|<p>Facebook's mis...|              Boston|                  MA|  3297|
|positive|            Facebook|Software Engineer...|<p>Facebook's mis...|             Bellevue|                  WA|  3353|
|positive|     Research Square|Machine Learning ...|"<p><a href=""htt...| which provides s...|         hardworking|   282|
|positive|   Rebellion Defense|Senior Software E...|As a Senior Softw...|          Washington|                  DC|  3146|
|positive|        7 Chord inc.|Data Engineer  Al...|This is an except...|            Brooklyn|                  NY|  2678|
|positive|Applied Research ...|Software Engineer...|"<p><u>Title</u>:...|          and design|         development|  1846|
|positive|        Carbon Relay|Senior Machine Le...|"<p>Carbon Relay ...| tackles hard pro...| and has fun!</p>...|   530|
|positive|            Facebook|Data Engineer, Ma...|<p>Facebook's mis...|             Fremont|                  CA|  4815|
|positive|            Facebook|Software Engineer...|<p>Facebook's mis...|          Menlo Park|                  CA|  3357|
|positive|        OmniSci, Inc|Software Engineer...|<p>Our mission at...|       San Francisco|                  CA|  4301|
|positive|            Facebook|Instagram - Softw...|<p>Facebook's mis...|            New York|                  NY|  3458|
|positive|         PlayStation|Senior Software E...|<p><strong>Senior...|           San Mateo|                  CA|  2196|
|positive|Sailpoint Technol...|Software Engineer...|<p>At SailPoint, ...|              Austin|                  TX|  3960|
|positive|Eastman Chemical ...|Machine Learning ...|<p><strong>Summar...|           Kingsport|                  TN|  4479|
|positive|           StyleSeat|Senior Machine Le...|<p><strong>Senior...|       San Francisco|                  CA|  3457|
|positive|               Apple|SW Engineer, Mach...|Want to own the p...|             Seattle|                  WA|  1310|
|positive|               Apple|Machine Learning ...|Have you ever sch...|           Cupertino|                  CA|  1549|
|positive|               Apple|Siri - Machine Le...|The Siri Client G...|             Seattle|                  WA|  2050|
+--------+--------------------+--------------------+--------------------+--------------------+--------------------+------+
only showing top 20 rows
```

```
%pyspark
from pyspark.ml.feature import Tokenizer, StopWordsRemover, HashingTF, IDF, StringIndexer
# Create all the features to the data set
pos_neg_to_num = StringIndexer(inputCol='class',outputCol='label')
tokenizer = Tokenizer(inputCol="description", outputCol="token_text")
stopremove = StopWordsRemover(inputCol='token_text',outputCol='stop_tokens')
hashingTF = HashingTF(inputCol="token_text", outputCol='hash_token')
idf = IDF(inputCol='hash_token', outputCol='idf_token')
```

```
%pyspark
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.linalg import Vector

# Create feature vectors
clean_up = VectorAssembler(inputCols=['idf_token', 'length'], outputCol='features')
```

```
%pyspark
# Create a and run a data processing Pipeline
from pyspark.ml import Pipeline
data_prep_pipeline = Pipeline(stages=[pos_neg_to_num, tokenizer, stopremove, hashingTF, idf, clean_up])
```

```
%pyspark
# Fit and transform the pipeline
cleaner = data_prep_pipeline.fit(data_df)
cleaned = cleaner.transform(data_df)
```

```
%pyspark
# Show label and resulting features
cleaned.select(['label', 'features']).show()
```

```
+-----+--------------------+
|label|            features|
+-----+--------------------+
|  0.0|(262145,[882,1836...|
|  0.0|(262145,[619,1156...|
|  0.0|(262145,[619,1156...|
|  0.0|(262145,[619,1156...|
|  0.0|(262145,[9639,178...|
|  0.0|(262145,[784,1836...|
|  0.0|(262145,[966,1342...|
|  0.0|(262145,[1079,183...|
|  0.0|(262145,[619,1836...|
|  0.0|(262145,[842,966,...|
|  0.0|(262145,[619,1156...|
|  0.0|(262145,[553,619,...|
|  0.0|(262145,[1156,123...|
|  0.0|(262145,[1836,242...|
|  0.0|(262145,[1079,183...|
|  0.0|(262145,[1115,135...|
|  0.0|(262145,[1836,211...|
|  0.0|(262145,[1836,353...|
|  0.0|(262145,[2111,271...|
|  0.0|(262145,[1222,183...|
+-----+--------------------+
only showing top 20 rows
```

```
%pyspark
from pyspark.ml.classification import NaiveBayes
# Break data down into a training set and a testing set
training, testing = cleaned.randomSplit([0.7, 0.3])

# Create a Naive Bayes model and fit training data
nb = NaiveBayes()
predictor = nb.fit(training)
```

```
%pyspark
# Tranform the model with the testing data
test_results = predictor.transform(testing)
test_results.show(5)
```

```
+--------+-------+--------------------+--------------------+---------+-----+------+-----+--------------------+--------------------+--------------------+--------------------+
+------------------+--------------------+--------------------+----------+
|   class|company|               title|         description|     city|state|length|label|          token_text|         stop_tokens|          hash_token|           idf_token|
features|       rawPrediction|         probability|prediction|
+--------+-------+--------------------+--------------------+---------+-----+------+-----+--------------------+--------------------+--------------------+--------------------+
+------------------+--------------------+--------------------+----------+
|negative|  Apple|Apple Podcasts Da...|At Apple, new ide...|Cupertino|   CA|  1324|  1.0|[at, apple,, new,...|[apple,, new, ide...|(262144,[3114,388...|(262144,[3114,388...|
(262145,[3114,388...|[-4971.7128041385...|[1.0,2.7247667923...|       0.0|
|negative|  Apple|Data Engineer - S...|The SWE Data Anal...|Cupertino|   CA|  1642|  1.0|[the, swe, data, ...|[swe, data, analy...|(262144,[1222,337...|(262144,[1222,337...|
(262145,[1222,337...|[-4926.0297107875...|[1.0,2.4278527177...|       0.0|
|negative|  Apple|Enterprise BI Pla...|Do you like to be...|Cupertino|   CA|  2743|  1.0|[do, you, like, t...|[like, around, pe...|(262144,[630,2642...|(262144,[630,2642...|
(262145,[630,2642...|[-10646.706232059...|[1.0,2.2750919159...|       0.0|
|negative|  Apple|Senior iOS Engine...|At Apple, we work...|Cupertino|   CA|   952|  1.0|[at, apple,, we, ...|[apple,, work, ev...|(262144,[6369,880...|(262144,[6369,880...|
(262145,[6369,880...|[-3717.2087861501...|[1.0,2.1801168123...|       0.0|
|negative|  Apple|Software Engineer...|Are you ready to ...|Cupertino|   CA|  1274|  1.0|[are, you, ready,...|[ready, explore, ...|(262144,[2111,264...|(262144,[2111,264...|
(262145,[2111,264...|[-4324.9234463106...|[1.0,1.7259995595...|       0.0|
+--------+-------+--------------------+--------------------+---------+-----+------+-----+--------------------+--------------------+--------------------+--------------------+
+------------------+--------------------+--------------------+----------+
only showing top 5 rows
```

```
%pyspark
# Use the Class Evaluator for a cleaner description
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

acc_eval = MulticlassClassificationEvaluator()
acc = acc_eval.evaluate(test_results)
print("Accuracy of model at predicting job profile was: %f" % acc)
```

```
Accuracy of model at predicting job profile was: 0.760296
```

Interpreter: python.