

project queries - Databricks Com x Untitled document - Google Doc x Course Homepage - CS 4371.50 x Project.pdf x YouTube x + -

community.cloud.databricks.com/?o=8578117171331102#notebook/1548904151009498/command/1548904151009499

Apps reddit GitHub Looking to land a s... Learn to Code - for... Ruby on Rails | A w... HackerRank Advice for new dev... How to get up to 3... Other bookmarks

project queries (Python) The University of Tex...

My Cluster File Edit View: Code Permissions Run All Clear Publish Comments Runs Revision history

```
11 df = sqlContext.read.json("/FileStore/tables/tweets.json")
12 df.show()
13
14
```

(2) Spark Jobs

df: pyspark.sql.dataframe.DataFrame

_corrupt_record: string
country: string
latitude: string
longitude: string
sentiment: string
state: string
tweet: string

_corrupt_record	country	latitude	longitude	sentiment	state	tweet
null	null	null	null	neutral	null	Inthiswks:Asweliv...
null	UnitedStatesofAme...	44.0581728	-121.3153896	neutral	Oregon	SomeoneIlovevery...
null	UnitedKingdom	54.4253879	-6.4448525	neutral	NorthernIreland	IpromiseIwillneve...
null	UnitedKingdom	51.5073219	-0.1276474	positive	England	PrivateEyemakesaf...
null	UnitedStatesofAme...	31.2437872	-90.4531536	neutral	Mississippi	Bumpfortheafterno...
null	null	null	null	neutral	null	Ourcountryisfacin...
null	null	null	null	neutral	null	SoTHISwouldn'thap...
null	UnitedKingdom	52.7954791	-0.5402402866174321	positive	England	PrivateEyemakesaf...
null	India	10.3528744	76.5128396	neutral	Kerala	EvenaftertheSpani...
null	UnitedStatesofAme...	40.7127281	-74.0060152	neutral	NewYork	April'sEmployment...
null	UnitedStatesofAme...	38.6529545	-90.2411656024635	neutral	Missouri	ThatsgreatAmanda...
null	UnitedStatesofAme...	40.7101463	-73.9195255	neutral	NewYork	NewIllaIprevail...
null	null	null	null	neutral	null	ChrisOyakhilomeh...
null	null	null	null	neutral	null	Aninsidelookat'st...
null	null	null	null	neutral	null	Texashasbeenpho...
null	UnitedStatesofAme...	41.0751048	-111.9741925	positive	Utah	pleaseLookatourin...
null	Indonesia	0.5262455	101.4515727	neutral	Riau	BREAKING:confirmS...
null	France	48.584614	7.7507127	positive	GrandEst	Intheirresponseto...
null	UnitedStatesofAme...	35.925193	-86.8689365	neutral	Tennessee	Whenyouwearacloth...
null	UnitedStatesofAme...	25.7742658	-80.1936589	positive	Florida	Parentswithjointc...

only showing top 20 rows

Type here to search 9:03 PM 5/7/2020

project queries - Databricks Com x Untitled document - Google Doc x Course Homepage - CS 4371.50 x Project.pdf x YouTube x + -

community.cloud.databricks.com/?o=8578117171331102#notebook/1548904151009498/command/3411472298759296

Apps reddit GitHub Looking to land a s... Learn to Code - for... Ruby on Rails | A w... HackerRank Advice for new dev... How to get up to 3... Other bookmarks

project queries (Python) The University of Tex...

My Cluster File Edit View: Code Permissions Run All Clear Publish Comments Runs Revision history

_corrupt_record	country	latitude	longitude	sentiment	state	tweet
null	UnitedStatesofAme...	41.0751048	-111.9741925	positive	Utah	pleaseLookatourin...
null	Indonesia	0.5262455	101.4515727	neutral	Riau	BREAKING:confirmS...
null	France	48.584614	7.7507127	positive	GrandEst	Intheirresponseto...
null	UnitedStatesofAme...	35.925193	-86.8689365	neutral	Tennessee	Whenyouwearacloth...
null	UnitedStatesofAme...	25.7742658	-80.1936589	positive	Florida	Parentswithjointc...

only showing top 20 rows

Command took 1.13 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 7:47:08 PM on My Cluster

Cmd 2

```
1 df.printSchema()
```

root
|-- _corrupt_record: string (nullable = true)
|-- country: string (nullable = true)
|-- latitude: string (nullable = true)
|-- longitude: string (nullable = true)
|-- sentiment: string (nullable = true)
|-- state: string (nullable = true)
|-- tweet: string (nullable = true)

Command took 0.03 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:03:57 PM on My Cluster

Cmd 3

```
1 df.select("country").show()
```

(1) Spark Jobs

country
null
UnitedStatesofAme...
UnitedKingdom
UnitedKingdom
UnitedStatesofAme...
null

Type here to search 9:04 PM 5/7/2020

project queries (Python) | The University of Tex...

My Cluster | File | Edit | View: Code | Permissions | Run All | Clear | Publish | Comments | Runs | Revision history

```
-- tweet: string (nullable = true)
```

Command took 0.03 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:03:57 PM on My Cluster

Cnd 3

```
1 df.select("country").show()
```

▶ (1) Spark Jobs

country
null
UnitedStatesofAme...
UnitedKingdom
UnitedKingdom
UnitedStatesofAme...
null
null
UnitedKingdom
India
UnitedStatesofAme...
UnitedStatesofAme...
UnitedStatesofAme...
null
null
null
UnitedStatesofAme...
Indonesia
France
UnitedStatesofAme...
UnitedStatesofAme...

only showing top 20 rows

Command took 0.30 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:04:15 PM on My Cluster

Cnd 4

```
1 df.select(df['country']).where(df.sentiment == "negative").show()
```

project queries (Python) | The University of Tex...

My Cluster | File | Edit | View: Code | Permissions | Run All | Clear | Publish | Comments | Runs | Revision history

only showing top 20 rows

Command took 0.30 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:04:15 PM on My Cluster

Cnd 4

```
1 df.select(df['country']).where(df.sentiment == "negative").show()
2 df.select(df['country']).where(df.sentiment == "neutral").show()
3 df.select(df['country']).where(df.sentiment == "positive").show()
```

▶ (3) Spark Jobs

country
null

▶ (3) Spark Jobs

country
null
UnitedStatesofAme...
UnitedKingdom
UnitedStatesofAme...
null
null
India
UnitedStatesofAme...
UnitedStatesofAme...
UnitedStatesofAme...
null
null

only showing top 20 rows

Command took 0.91 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:04:31 PM on My Cluster

Cnd 5

```
1 df.groupBy("sentiment").count().show()
```

community.cloud.databricks.com/?o=8578117171331102#notebook/1548904151009498/command/169941611426259

project queries (Python)

My Cluster

only showing top 20 rows

Command took 0.30 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:04:15 PM on My Cluster

Cmd 4

```
1 df.select(df['country']).where(df.sentiment == "negative").show()
2 df.select(df['country']).where(df.sentiment == "neutral").show()
3 df.select(df['country']).where(df.sentiment == "positive").show()
```

(3) Spark Jobs

country
null
null
Indonesia
UnitedStatesofAme...
UnitedStatesofAme...
UnitedStatesofAme...
null
null
UnitedKingdom

only showing top 20 rows

country
UnitedKingdom
UnitedKingdom
UnitedStatesofAme...
France
UnitedStatesofAme...
null

Command took 0.91 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:04:31 PM on My Cluster

Cmd 5

```
1 df.groupBy("sentiment").count().show()
```

community.cloud.databricks.com/?o=8578117171331102#notebook/1548904151009498/command/169941611426261

project queries (Python)

My Cluster

sentiment
France
UnitedStatesofAme...
null

Command took 0.91 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:04:31 PM on My Cluster

Cmd 5

```
1 df.groupBy("sentiment").count().show()
```

(5) Spark Jobs

sentiment	count
positive	6
null	2
neutral	28
negative	1

Command took 1.96 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:05:10 PM on My Cluster

Cmd 6

```
1 df.groupBy("country").count().show()
2 df.groupBy("state").count().show()
```

(10) Spark Jobs

country	count
Utah	1
Kerala	1
Oregon	1
GrandEst	1
null	16
DistrictofColumbia	1
Missouri	1
Riau	1
Michigan	1

project queries - Databricks Com x Untitled document - Google Doc x Course Homepage - CS 4371.50 x Project.pdf x YouTube x + -

community.cloud.databricks.com/?o=8578117171331102#notebook/1548904151009498/command/169941611426260

Apps reddit GitHub Looking to land a s... Learn to Code - for... Ruby on Rails | A w... HackerRank Advice for new dev... How to get up to 3... Other bookmarks

project queries (Python) The University of Tex...

My Cluster File Edit View: Code Permissions Run All Clear Publish Comments Runs Revision history

Command took 1.96 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:05:10 PM on My Cluster

Cmd 6

```
1 df.groupBy("country").count().show()
2 df.groupBy("state").count().show()
```

(10) Spark Jobs

country	count
France	1
null	16
India	1
UnitedKingdom	4
Indonesia	1
UnitedStatesofAme...	14

state	count
Utah	1
Kerala	1
Oregon	1
Grandest	1
null	16
DistrictofColumbia	1
Missouri	1

Command took 3.09 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:05:10 PM on My Cluster

Cmd 7

```
1 df.createOrReplaceTempView("tweet")
2
3 sqlDF = spark.sql("SELECT country as Country, sentiment FROM tweet")
4 spark.sql("SELECT country as Country FROM tweet GROUP BY country").show()
5
6 sqlDF.show()
```

Type here to search 9:05 PM 5/7/2020

project queries - Databricks Com x Untitled document - Google Doc x Course Homepage - CS 4371.50 x Project.pdf x YouTube x + -

community.cloud.databricks.com/?o=8578117171331102#notebook/1548904151009498/command/169941611426260

Apps reddit GitHub Looking to land a s... Learn to Code - for... Ruby on Rails | A w... HackerRank Advice for new dev... How to get up to 3... Other bookmarks

project queries (Python) The University of Tex...

My Cluster File Edit View: Code Permissions Run All Clear Publish Comments Runs Revision history

Command took 1.96 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:05:10 PM on My Cluster

Cmd 6

```
1 df.groupBy("country").count().show()
2 df.groupBy("state").count().show()
```

(10) Spark Jobs

country	count
France	1
null	16
India	1
UnitedKingdom	4
Indonesia	1
UnitedStatesofAme...	14

state	count
Utah	1
Kerala	1
Oregon	1
GrandEst	1
null	16
DistrictofColumbia	1
Missouri	1
Riau	1
Michigan	1
Arizona	1
NorthernIreland	1
Mississippi	1
Tennessee	1
Florida	2
Indiana	1
NewYork	3
England	2
Scotland	1

Command took 3.09 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:05:10 PM on My Cluster

Cmd 7

```
1 df.createOrReplaceTempView("tweet")
2
3 sqlDF = spark.sql("SELECT country as Country, sentiment FROM tweet")
4 spark.sql("SELECT country as Country FROM tweet GROUP BY country").show()
5
6 sqlDF.show()
```

Type here to search 9:05 PM 5/7/2020

project queries - Databricks Com x | Untitled document - Google Doc x | Course Homepage - CS 4371.50 x | Project.pdf x | YouTube x | +

community.cloud.databricks.com/?o=8578117171331102#notebook/1548904151009498/command/169941611426262

Apps reddit GitHub Looking to land a s... Learn to Code - for... Ruby on Rails | A w... HackerRank Advice for new dev... How to get up to 3... Other bookmarks

project queries (Python) ? The University of Tex...

My Cluster File Edit View: Code Permissions Run All Clear Publish Comments Runs Revision history

Command took 3.09 seconds -- by rxa16063@outdallas.edu at 5/7/2020, 9:05:30 PM on My Cluster

Cmd 7

```
1 df.createOrReplaceTempView("tweet")
2
3 sqlDF = spark.sql("SELECT country as Country, sentiment FROM tweet")
4 spark.sql("SELECT country as Country FROM tweet GROUP BY country").show()
5
6 sqlDF.show()
```

▶ (6) Spark Jobs

sqDF: pyspark.sql.dataframe.DataFrame = [Country: string, sentiment: string]

Country
France
null
India
UnitedKingdom
Indonesia
UnitedStatesofAme...

Country	sentiment
null	neutral
UnitedStatesofAme...	neutral
UnitedKingdom	neutral
UnitedKingdom	positive
UnitedStatesofAme...	neutral
null	neutral
null	neutral

Command took 1.83 seconds -- by rxa16063@outdallas.edu at 5/7/2020, 9:06:01 PM on My Cluster

Cmd 8

```
1 #show_sentiment_for_countries
```

Type here to search

9:06 PM 5/7/2020

project queries - Databricks Com x | Untitled document - Google Doc x | Course Homepage - CS 4371.50 x | Project.pdf x | YouTube x | +

community.cloud.databricks.com/?o=8578117171331102#notebook/1548904151009498/command/169941611426262

Apps reddit GitHub Looking to land a s... Learn to Code - for... Ruby on Rails | A w... HackerRank Advice for new dev... How to get up to 3... Other bookmarks

project queries (Python) ? The University of Tex...

My Cluster File Edit View: Code Permissions Run All Clear Publish Comments Runs Revision history

Command took 3.09 seconds -- by rxa16063@outdallas.edu at 5/7/2020, 9:05:30 PM on My Cluster

Cmd 7

```
1 df.createOrReplaceTempView("tweet")
2
3 sqlDF = spark.sql("SELECT country as Country, sentiment FROM tweet")
4 spark.sql("SELECT country as Country FROM tweet GROUP BY country").show()
5
6 sqlDF.show()
```

▶ (6) Spark Jobs

sqDF: pyspark.sql.dataframe.DataFrame = [Country: string, sentiment: string]

null	neutral
UnitedStatesofAme...	neutral
UnitedKingdom	neutral
UnitedKingdom	positive
UnitedStatesofAme...	neutral
null	neutral
null	neutral
UnitedKingdom	positive
India	neutral
UnitedStatesofAme...	neutral
UnitedStatesofAme...	neutral
UnitedStatesofAme...	neutral
null	neutral
null	neutral
null	neutral
UnitedStatesofAme...	positive
Indonesia	neutral
France	positive
UnitedStatesofAme...	neutral
UnitedStatesofAme...	positive

Command took 1.83 seconds -- by rxa16063@outdallas.edu at 5/7/2020, 9:06:01 PM on My Cluster

Cmd 8

```
1 #show_sentiment_for_countries
```

Type here to search

9:06 PM 5/7/2020

community.cloud.databricks.com/?o=8578117171331102#notebook/1548904151009498/command/169941611426263

project queries (Python)

My Cluster

```
1 #show sentiment for countries
2 df.groupby("country").agg(functions.count('sentiment')).show()
3
4 #show sentiment for state
5 df.groupby("state").agg(functions.count('sentiment')).show()
6
7
8
9 df.groupby("country", "state").agg(functions.count('sentiment')).show()
```

(15) Spark Jobs

country	count(sentiment)
France	1
null	14
India	1
UnitedKingdom	4
Indonesia	1
UnitedStatesofAme...	14

state	count(sentiment)
Utah	1
Kerala	1
Oregon	1
GrandEst	1
null	14
DistrictofColumbia	1
Missouri	1

Command took 5.95 seconds -- by rxa16063@utdallas.edu at 5/7/2020, 9:06:39 PM on My Cluster

```
Cmd 9
1 #Trump
2 df.groupby("tweet").agg(functions.count('tweet')).filter(df.tweet.contains("Trump")).show()
```

community.cloud.databricks.com/?o=8578117171331102#notebook/1548904151009498/command/169941611426263

project queries (Python)

My Cluster

```
1 #show sentiment for countries
2 df.groupby("country").agg(functions.count('sentiment')).show()
3
4 #show sentiment for state
5 df.groupby("state").agg(functions.count('sentiment')).show()
6
7
8
9 df.groupby("country", "state").agg(functions.count('sentiment')).show()
```

(15) Spark Jobs

state	count(sentiment)
Utah	1
Kerala	1
Oregon	1
GrandEst	1
null	14
DistrictofColumbia	1
Missouri	1
Riau	1
Michigan	1
Arizona	1
NorthernIreland	1
Mississippi	1
Tennessee	1
Florida	2
Indiana	1
NewYork	3
England	2
Scotland	1

Command took 5.95 seconds -- by rxa16063@utdallas.edu at 5/7/2020, 9:06:39 PM on My Cluster

```
Cmd 9
1 #Trump
2 df.groupby("tweet").agg(functions.count('tweet')).filter(df.tweet.contains("Trump")).show()
```

project queries - Databricks Com x Untitled document - Google Doc x Course Homepage - CS 4371.50 x Project.pdf x YouTube x + -

community.cloud.databricks.com/?o=8578117171331102#notebook/1548904151009498/command/169941611426263

Apps reddit GitHub Looking to land a s... Learn to Code - for... Ruby on Rails | A w... HackerRank Advice for new dev... How to get up to 3... Other bookmarks

project queries (Python) The University of Tex...

My Cluster File Edit View: Code Permissions Run All Clear Publish Comments Runs Revision history

```
5 df.groupBy("state").agg(functions.count('sentiment')).show()
6
7
8
9 df.groupBy("country", "state").agg(functions.count('sentiment')).show()
```

(15) Spark Jobs

country	state	count(sentiment)
France	GrandEst	1
UnitedKingdom	NorthernIreland	1
India	Kerala	1
null	null	14
UnitedStatesofAme...	Indiana	1
UnitedStatesofAme...	Missouri	1
UnitedStatesofAme...	DistrictofColumbia	1
UnitedStatesofAme...	Utah	1
UnitedStatesofAme...	Michigan	1
UnitedStatesofAme...	NewYork	3
UnitedStatesofAme...	Arizona	1
Indonesia	Riau	1
UnitedKingdom	England	2
UnitedKingdom	Scotland	1
UnitedStatesofAme...	Mississippi	1
UnitedStatesofAme...	Florida	2
UnitedStatesofAme...	Tennessee	1
UnitedStatesofAme...	Oregon	1

Command took 5.95 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:06:39 PM on My Cluster

Cmd 9

```
1 #Trump
2 df.groupBy("tweet").agg(functions.count('tweet')).filter(df.tweet.contains("Trump")).show()
3
4 df.groupBy("tweet").agg(functions.count('tweet')).filter(df.tweet.contains("fakenews")).show()
```

(10) Spark Jobs

Type here to search 9:07 PM 5/7/2020

project queries - Databricks Com x Untitled document - Google Doc x Course Homepage - CS 4371.50 x Project.pdf x YouTube x + -

community.cloud.databricks.com/?o=8578117171331102#notebook/1548904151009498/command/2883213139230025

Apps reddit GitHub Looking to land a s... Learn to Code - for... Ruby on Rails | A w... HackerRank Advice for new dev... How to get up to 3... Other bookmarks

project queries (Python) The University of Tex...

My Cluster File Edit View: Code Permissions Run All Clear Publish Comments Runs Revision history

```
1 Indonesia
2 Riau
3 UnitedKingdom
4 UnitedKingdom
5 UnitedStatesofAme...
6 UnitedStatesofAme...
7 UnitedStatesofAme...
8 UnitedStatesofAme...
9 UnitedStatesofAme...
```

tweet	count(tweet)
Texashasbeenpheno...	3
EvenaftertheSpan...	1

tweet	count(tweet)
Wasitsatan?Issata...	1
SomeoneIloveyerm...	1

Command took 3.38 seconds -- by rxa160630@utdallas.edu at 5/7/2020, 9:07:33 PM on My Cluster

Cmd 9

```
1 #Trump
2 df.groupBy("tweet").agg(functions.count('tweet')).filter(df.tweet.contains("Trump")).show()
3
4 df.groupBy("tweet").agg(functions.count('tweet')).filter(df.tweet.contains("fakenews")).show()
```

(10) Spark Jobs

Shift+Enter to run shortcuts

Type here to search 9:07 PM 5/7/2020