question 327 views

3d not sure how to debug it.

i got following error message for 3d, i just applied create_one_hot_dict to parsed_train_df. I passed all test before. I am not quite understanding the error message. any help would be appreciated.

```
AnalysisException: u'resolved attribute(s) features#39 missing from label#4365,feature#4366 in operator !Generate explode(features#39), false,
 false, None, [col#4696];'
---------------------------------------------------------------------------
AnalysisException                         Traceback (most recent call last)
<ipython-input-141-868ee0035974> in <module>()
      1 # TODO: Replace <FILL IN> with appropriate code
      2
----> 3 ctr_ohe_dict = create_one_hot_dict(parsed_train_df)
      4 num_ctr_ohe_feats = len(ctr_ohe_dict)
      5 print num_ctr_ohe_feats

<ipython-input-59-1382c4453110> in create_one_hot_dict(input_df)
     11             unique integers.
     12     """
---> 13     sample_distinct_feats_df = (input_df.select(explode(sample_data_df.features))).distinct()
     14     sample_dict =(sample_distinct_feats_df
     15                   .rdd

/databricks/spark/python/pyspark/sql/dataframe.py in select(self, *cols)
    860         [Row(name=u'Alice', age=12), Row(name=u'Bob', age=15)]
    861         """
--> 862         jdf = self._jdf.select(self._jcols(*cols))
    863         return DataFrame(jdf, self.sql_ctx)
    864

/databricks/spark/python/lib/py4j-0.9-src.zip/py4j/java_gateway.py in __call__(self, *args)
    811         answer = self.gateway_client.send_command(command)
    812         return_value = get_return_value(
--> 813             answer, self.gateway_client, self.target_id, self.name)
    814
    815         for temp_arg in temp_args:

/databricks/spark/python/pyspark/sql/utils.py in deco(*a, **kw)
     49                                             e.java_exception.getStackTrace()))
     50             if s.startswith('org.apache.spark.sql.AnalysisException: '):
---> 51                 raise AnalysisException(s.split(': ', 1)[1], stackTrace)
     52             if s.startswith('java.lang.IllegalArgumentException: '):
```

```
    53                 raise IllegalArgumentException(s.split(': ', 1)[1], stackTrace)

AnalysisException: u'resolved attribute(s) features#39 missing from label#4365,feature#4366 in operator !Generate explode(features#39), false,
 false, None, [col#4696];'
```

lab3

Updated 6 days ago by Anonymous

the students' answer, *where students collectively construct a single answer*

From the error message we can read "*resolved attribute(s) features#39 missing from label#4365,feature*".
That indicates that you are asking for a column or attribute named 'features', but there exists no such column. 'label' and 'feature' exists though.

Updated 6 days ago by Bo Rosenquist

followup discussions *for lingering questions and comments*

Resolved Unresolved

Anonymous 6 days ago
change feature to features gives same error message.

```
AnalysisException: u'resolved attribute(s) features#7 missing from label#62,features#63 in operator !Generate explode(features#7), fals
e, false, None, [col#154];'
```

Bo Rosenquist 6 days ago
Now it seams you are referencing one column from one dataframe to another.

```
df = sqlContext.createDataFrame([(1,), (2,), (3,)], ('v',))
df.select((df.v + 1).alias('v')).select(df.v + 2).show()
```

This will fail, as the last df.v is referring to the first dataframe, although that column has been replaced by another one.

```
u'resolved attribute(s) v#1759L missing from v#1760L in operator !Project'
```

What you could do is to use the col function to reference the second column

```
from pyspark.sql.functions import col
df = sqlContext.createDataFrame([(1,), (2,), (3,)], ('v',))
df.select((df.v + 1).alias('v')).select((col('v') + 2).alias('v')).show()
```

Now we are not referensing the first v but the available v.

```
+---+
|  v|
+---+
|  4|
|  5|
|  6|
+---+
```

Cor van der Hulst 6 days ago I want one dataframe to be made up of two columns:

df = sqlContext.createDataFrame([(1,), (2,), (3,)], ('v',))
df2 = sqlContext.createDataFrame([(4,), (5,), (6,)], ('w',))
iwant = df2.withColumn("x", df.v)

and it gives the same cryptic message: u'resolved attribute(s) v#11907L missing from w#11908L in operator !Project [w#11908L,v#11907L AS x#11909L];'

how can i add two dataframes of the same length?

Bo Rosenquist 6 days ago
If you want to merge two DataFrames you should use a join.
NB, you should have a key to connect correct rows between the two DataFrames. If you don't supply a key you will get all possible combinations of rows.

```
df = sqlContext.createDataFrame([(1,), (2,), (3,)], ('v',))
df2 = sqlContext.createDataFrame([(4,), (5,), (6,)], ('w',))
nokey = df2.join(df)
nokey.show()
```

```
+---+---+
|  w|  v|
+---+---+
|  4|  1|
|  4|  2|
|  4|  3|
|  5|  1|
|  5|  2|
|  5|  3|
|  6|  1|
|  6|  2|
|  6|  3|
+---+---+
```

If no is key available, you can create one. But you can't be sure which row from df will be joined with which from df2

```
df = df.rdd.zipWithIndex().map(lambda (r,i): (r[0], i)).toDF(['v','key'])
df2 = df2.rdd.zipWithIndex().map(lambda (r,i): (r[0], i)).toDF(['w', 'key'])
yougot = df2.join(df, 'key')
yougot.show()


+---+---+---+
|key|  w|  v|
+---+---+---+
|  0|  4|  1|
|  1|  5|  2|
|  2|  6|  3|
+---+---+---+
```

Looking on the above code, the key issue is a major one in my view.

Bo Rosenquist 6 days ago You can also zip two rdds, with the same result as making you own key, that is, you can't be sure about what item gets zipped with what

```
df.rdd.zip(df2.rdd).map(lambda (v, w): (v[0], w[0])).toDF(['v','w']).show()

+---+---+
|  v|  w|
+---+---+
|  1|  4|
|  2|  5|
```

```
|  3|  6|
+---+---+
```

Bo Rosenquist 6 days ago
@Cor, I saw your question on @473. The suggestions I show above are certainly not suited for that task.

Cor van der Hulst 6 days ago
Bo,

thanks for your answer, i now understand the error message, and how to solve it. Thanks again

Anonymous 4 days ago I don't understand how Bo's answer solves the error of why withColumnRenamed('feature', 'features') doesn't work to resolve the error.

Aditya Athalye 2 days ago withColumnRenamed worked for me. Not sure why that approach does not work for others. Or maybe I am missing something.

Anonymous 2 days ago Hi - The withColumnRenamed does not work for me either, though when I display the DF I see it has features
p=parsed_train_df.select().withColumnRenamed('feature','features').show()

label| features|
+-----+--------------------+ |
0.0|[[0,], [1,-1], [2...|
| 0.0|[[0,], [1,-1], [2...|

But ctr_ohe_dict = create_one_hot_dict(p)
gives this issue
AnalysisException: u'resolved attribute(s) features#1692 missing from label#1909,features#2662 in operator !Generate explode(features#1692), false, false, None, [col#2663];'

AnalysisException Traceback (most recent call last)
<ipython-input-241-6968c55cd8ce> in <module>()
2 p=parsed_train_df.select('label','feature').withColumnRenamed('feature','features')
3
----> 4 ctr_ohe_dict = create_one_hot_dict(p)
5 #num_ctr_ohe_feats = len(ctr_ohe_dict)
6 #print num_ctr_ohe_feats
<ipython-input-186-441be41e70f7> in create_one_hot_dict(input_df)
12 """
13 input_distinct_feats_df = (input_df

```
---> 14 .select(explode(sample_data_df.features)).distinct())
15 input_ohe_dict = (input_distinct_feats_df
16 .rdd
```

Resolved Unresolved

Carolyn Ownby 2 days ago
I can't get this to work either. I did a select with alias before calling create_one_hot_dict, and the resulting dataframe LOOKS ok, I think:

```
+--------------------+
| features|
+--------------------+
|[[0,], [1,-1], [2...|
|[[0,], [1,-1], [2...|
+--------------------+
only showing top 2 rows
...
AnalysisException: u'resolved attribute(s) features#876 missing from features#1010 in operator !Generate explode(features#876), false,
 false, None, [col#1021];'
```

So, is this feature/features mess a mistake? Can I change 3c to reference 'features' instead of 'feature'? Nevermind on 3c modification - same error. What am I missing?

Carolyn Ownby 2 days ago Found my blunder! It was a cut/paste error in create_one_hot_dict: I was using sample_data_df instead of input_df in my explode.

Juan Tapia Osorio 2 hours ago thanks I had this problem too, and I solved it when I did read this answer

Resolved Unresolved

Akshay Jain 18 hours ago
@Bo
I think you can help me.
I am getting weird error which I havent seen on piazza till now and I stuck on this part from a long time now.

I am using the following code :
ctr_ohe_dict = create_one_hot_dict(create_one_hot_dict(parsed_train_df.withColumnRenamed('feature','features')))

print ctr_ohe_dict
num_ctr_ohe_feats = len(ctr_ohe_dict.keys())
print num_ctr_ohe_feats
print ctr_ohe_dict[(0, '')]

and getting this weird error.

KeyError: (0, '')
{(2, u'mouse'): 5, (0, u'cat'): 6, (0, u'bear'): 0, (2, u'salmon'): 4, (1, u'tabby'): 3, (1, u'black'): 2, (0, u'mouse'): 1} 7 ---------- KeyError Traceback (most recent call last) <ipython-input-36-d22fb25add7e> in <module>() 5 num_ctr_ohe_feats = len(ctr_ohe_dict.keys()) 6 print num_ctr_ohe_feats ----> 7 print ctr_ohe_dict[(0, '')] KeyError: (0, '')
Please help me and i am really stuck
Thanks a lot

Bo Rosenquist 18 hours ago The `KeyError: (0, '')` means as you probably know that `ctr_ohe_dict` does not contain that key.
So something is going wrong when you build the `cre_ohe_dict` in the beginning of the above code.
On that line there are two calls to `create_one_hot_dict`. Is that a typo now? I think it is since the code would not run otherwise, so I overlook that.
The `print ctr_ohe_dict` prints the ohe from the sample data! It tells me you are using the wrong df in create_one_hot_dict.
I guess that you are using `sample_data_df` instead of `input_df` in create_one_hot_dict?

Akshay Jain 18 hours ago Thanks for your reply Bo.

Yes that was a typo.

This is my function:
def create_one_hot_dict(input_df):

return sample_distinct_feats_df.rdd.map(lambda r: tuple(r[0])).zipWithIndex().collectAsMap()

sample_ohe_dict_auto = create_one_hot_dict(sample_data_df)

which I think is correct.
I am seriously not able to figure out , where I am going wrong.

Bo Rosenquist 17 hours ago
In the return statement you are using `sample_distinct_feats_df` instead of `input_df`.
The result is that no matter what df you pass off to `create_one_hot_dict`, you will always get the dictionary for `sample_distinct_feats_df`.
You need to replace it with `input_df`.

Bo Rosenquist 17 hours ago When you do the replace, you will hopefully note that the test in (2c) will fail. You need get the distinct values in input_df.

Akshay Jain 16 hours ago Thank you so so much.
I would not have been able to do it without your help.

I really appreciate your help.
Thanks

Bo Rosenquist 16 hours ago
Happy to support.
Glad it worked out.