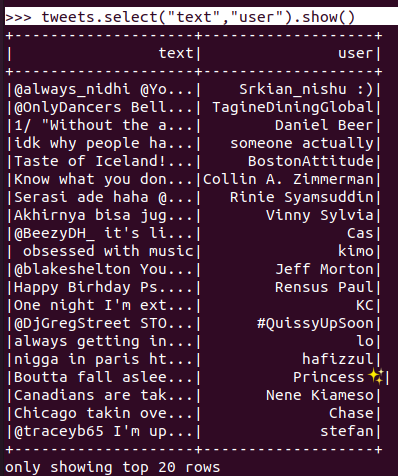
1. Find all the tweets by user

ans.>>> tweets.select("text","user").show()





1. Find how many tweets each user has

ans. >>> tweets.groupby("user").count().show()



def persons\_udf(x):

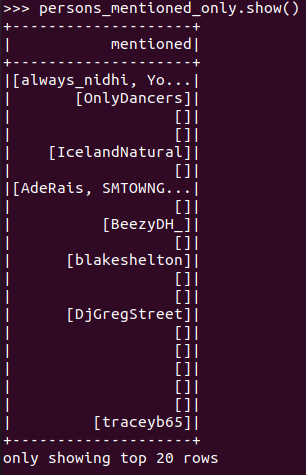
return re.findall(persons\_reg, x)

persons = udf(lambda x: persons\_udf(x), ArrayType(StringType()))

1. Find all the persons mentioned on tweets

>>> persons\_mentioned\_only=persons\_mentioned.select("mentioned")

>>> persons\_mentioned\_only.show()



1. Count how many times each person is mentioned

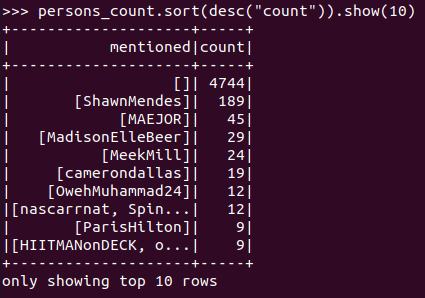
>>> persons\_count=persons\_mentioned.groupBy("mentioned").count()

>>> persons\_count.show()



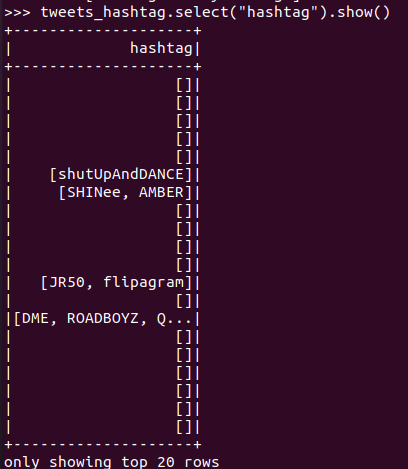
1. Find the 10 most mentioned persons

>>> persons\_count.sort(desc("count")).show(10)



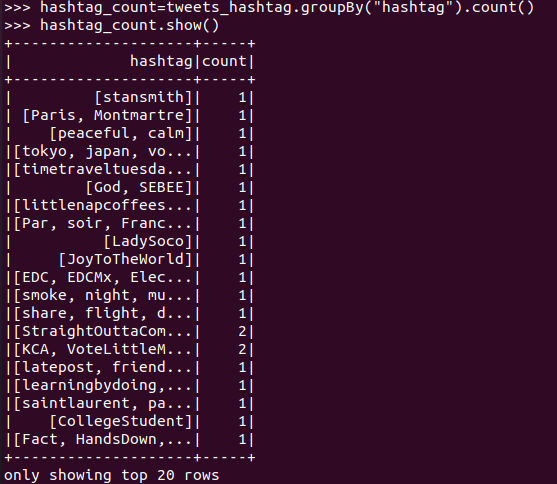
1. Find all the hashtags mentioned on a tweet

>>> tweets\_hashtag.select("hashtag").show()



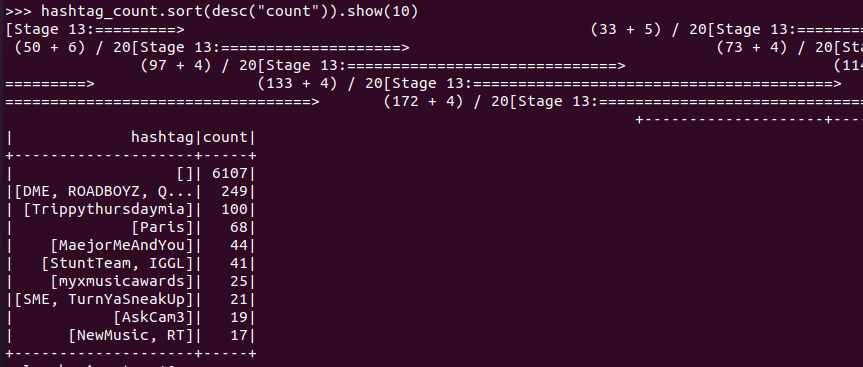
1. Count how many times each hashtag is mentioned

>>> hashtag\_count=tweets\_hashtag.groupBy("hashtag").count().show()



1. Find the 10 most popular Hashtags

>>> hashtag\_count.sort(desc("count")).show(10)



1. Find the top 5 countries which tweet the most.

>>> tweet=sqlCtx.read.json('/home/Downloads/tweets.json')

>>> a1=tweet.groupBy("country").count()

>>> a1.show()

>>> from pyspark.sql.functions import desc

>>> a1.sort(desc("count")).show(5)

