In [ ]:

```
'''

(10 points) Part 3a: MapRedce
In this task, design a MapReduce program in python that reads all the original twee
ts (no retweets) in the sample tweets (congress-sample-10k.json.gz) and if a tweet
 is a reply to another tweet then output a record of the form <src_id, src_user, ds
t_id, dst_user>.

Create a small cluster (2 or 3 nodes) as per the AWS Guide and then ssh to your clu
ster and use Hadoop streaming to execute your mapreduce program.

Note: the Hadoop streaming jar file can be found at /usr/lib/hadoop-mapreduce/hadoo
p-streaming.jar

'''
```

In [21]:

```
congress_sample = spark.read.csv("s3://us-congress-tweets/congress-sample-10k.json.
gz", header=True)
```

In [31]:

```
# your mapper function
# congress_sample_reply = congress_sample.select("in_reply_to_status_id_str")
# congress_sample_reply.show()


ak47 = congress_sample.limit(1).toPandas
```

In [35]:

```
ak47.head
```

```
'function' object has no attribute 'head'
Traceback (most recent call last):
AttributeError: 'function' object has no attribute 'head'
```

In [33]:

```
congress_sample.printSchema()
```

```
root
 |-- {"extended_tweet":{"entities":{"urls":[]: string (nullable = true)
 |-- "hashtags":[]1: string (nullable = true)
 |-- "user_mentions":[{"indices":[02: string (nullable = true)
 |-- 8]3: string (nullable = true)
 |-- "screen_name":"jillwow"4: string (nullable = true)
 |-- "id_str":"464222379"5: string (nullable = true)
 |-- "name":"jillie"6: string (nullable = true)
 |-- "id":464222379}7: string (nullable = true)
 |-- {"indices":[98: string (nullable = true)
 |-- 23]9: string (nullable = true)
 |-- "screen_name":"RepAdamSchiff"10: string (nullable = true)
 |-- "id_str":"29501253"11: string (nullable = true)
 |-- "name":"Adam Schiff"12: string (nullable = true)
 |-- "id":29501253}]13: string (nullable = true)
 |-- "symbols":[]}14: string (nullable = true)
 |-- "full_text":"@jillwow @RepAdamSchiff That is what is so disturbing
&amp; ominous 😳  Also: string (nullable = true)
 |--  we must remember our world history - didn\u2019t Hiler only have
30% or so support from the German ppl before he began his persecution o
f the Jewish ppl? 😕 ": string (nullable = true)
 |-- "display_text_range":[2417: string (nullable = true)
 |-- 226]}: string (nullable = true)
 |-- "in_reply_to_status_id_str":"1047522510839975936": string (nullabl
e = true)
 |-- "in_reply_to_status_id":1047522510839975936: string (nullable = tr
ue)
 |-- "created_at":"Wed Oct 03 17:30:03 +0000 2018": string (nullable =
true)
 |-- "in_reply_to_user_id_str":"464222379": string (nullable = true)
 |-- "source":"<a href=\"http://twitter.com/download/iphone\" rel=\"nof
ollow\">Twitter for iPhone<\/a>": string (nullable = true)
 |-- "retweet_count":0: string (nullable = true)
 |-- "retweeted":false: string (nullable = true)
 |-- "geo":null: string (nullable = true)
 |-- "filter_level":"low": string (nullable = true)
 |-- "in_reply_to_screen_name":"jillwow": string (nullable = true)
 |-- "is_quote_status":false: string (nullable = true)
 |-- "id_str":"1047539288705880064": string (nullable = true)
 |-- "in_reply_to_user_id":464222379: string (nullable = true)
 |-- "favorite_count":0: string (nullable = true)
 |-- "id":1047539288705880064: string (nullable = true)
 |-- "text":"@jillwow @RepAdamSchiff That is what is so disturbing &am
p; ominous 😳  Also: string (nullable = true)
 |--  we must remember our world history - didn\u2019t\u2026 https://t.
co/h5SURki12V": string (nullable = true)
 |-- "place":null: string (nullable = true)
 |-- "lang":"en"37: string (nullable = true)
 |-- "quote_count":0: string (nullable = true)
 |-- "favorited":false: string (nullable = true)
 |-- "coordinates":null: string (nullable = true)
 |-- "truncated":true: string (nullable = true)
 |-- "timestamp_ms":"1538587803423": string (nullable = true)
 |-- "reply_count":0: string (nullable = true)
 |-- "entities":{"urls":[{"display_url":"twitter.com/i/web/status/1\u20
26": string (nullable = true)
 |-- "indices":[121: string (nullable = true)
```

```
 |-- 144]: string (nullable = true)
 |-- "expanded_url":"https://twitter.com/i/web/status/10475392887058800
64": string (nullable = true)
 |-- "url":"https://t.co/h5SURki12V"}]: string (nullable = true)
 |-- "hashtags":[]49: string (nullable = true)
 |-- "user_mentions":[{"indices":[050: string (nullable = true)
 |-- 8]51: string (nullable = true)
 |-- "screen_name":"jillwow"52: string (nullable = true)
 |-- "id_str":"464222379"53: string (nullable = true)
 |-- "name":"jillie"54: string (nullable = true)
 |-- "id":464222379}55: string (nullable = true)
 |-- {"indices":[956: string (nullable = true)
 |-- 23]57: string (nullable = true)
 |-- "screen_name":"RepAdamSchiff"58: string (nullable = true)
 |-- "id_str":"29501253"59: string (nullable = true)
 |-- "name":"Adam Schiff"60: string (nullable = true)
 |-- "id":29501253}]61: string (nullable = true)
 |-- "symbols":[]}62: string (nullable = true)
 |-- "display_text_range":[2463: string (nullable = true)
 |-- 140]: string (nullable = true)
 |-- "contributors":null: string (nullable = true)
 |-- "user":{"utc_offset":null: string (nullable = true)
 |-- "friends_count":1034: string (nullable = true)
 |-- "profile_image_url_https":"https://pbs.twimg.com/profile_images/82
9148920223707137/L0JxcBmq_normal.jpg": string (nullable = true)
 |-- "listed_count":3: string (nullable = true)
 |-- "profile_background_image_url":"http://abs.twimg.com/images/theme
s/theme1/bg.png": string (nullable = true)
 |-- "default_profile_image":false: string (nullable = true)
 |-- "favourites_count":47592: string (nullable = true)
 |-- "description":"Proud citizen; 💕 Obama: string (nullable = true)
 |--  Michelle & HRC; #Resistance: string (nullable = true)
 |--  #BoycottNRA: string (nullable = true)
 |--  #BanAssaultWeapons #ActiveMeasures": string (nullable = true)
 |-- "created_at":"Wed Jun 17 17:50:07 +0000 2015": string (nullable =
true)
 |-- "is_translator":false: string (nullable = true)
 |-- "profile_background_image_url_https":"https://abs.twimg.com/image
s/themes/theme1/bg.png": string (nullable = true)
 |-- "protected":false: string (nullable = true)
 |-- "screen_name":"rposey42": string (nullable = true)
 |-- "id_str":"3248076530": string (nullable = true)
 |-- "profile_link_color":"1DA1F2": string (nullable = true)
 |-- "translator_type":"none": string (nullable = true)
 |-- "id":3248076530: string (nullable = true)
 |-- "geo_enabled":false: string (nullable = true)
 |-- "profile_background_color":"C0DEED": string (nullable = true)
 |-- "lang":"en"88: string (nullable = true)
 |-- "profile_sidebar_border_color":"C0DEED": string (nullable = true)
 |-- "profile_text_color":"333333": string (nullable = true)
 |-- "verified":false: string (nullable = true)
 |-- "profile_image_url":"http://pbs.twimg.com/profile_images/829148920
223707137/L0JxcBmq_normal.jpg": string (nullable = true)
 |-- "time_zone":null: string (nullable = true)
 |-- "url":null: string (nullable = true)
 |-- "contributors_enabled":false: string (nullable = true)
 |-- "profile_background_tile":false: string (nullable = true)
```

```
 |-- "profile_banner_url":"https://pbs.twimg.com/profile_banners/324807
6530/1520646319": string (nullable = true)
 |-- "statuses_count":70007: string (nullable = true)
 |-- "follow_request_sent":null: string (nullable = true)
 |-- "followers_count":660: string (nullable = true)
 |-- "profile_use_background_image":true: string (nullable = true)
 |-- "default_profile":true: string (nullable = true)
 |-- "following":null: string (nullable = true)
 |-- "name":"Rhoda🦋 ": string (nullable = true)
 |-- "location":"California: string (nullable = true)
 |--  USA": string (nullable = true)
 |-- "profile_sidebar_fill_color":"DDEEF6": string (nullable = true)
 |-- "notifications":null}}: string (nullable = true)
```

In [37]:

```
congress_sample.rdd.map(lambda r: r.notifications).collect()
```

```
An error occurred while calling z:org.apache.spark.api.python.PythonRD
D.collectAndServe.
: org.apache.spark.SparkException: Job aborted due to stage failure: Ta
sk 0 in stage 5.0 failed 4 times, most recent failure: Lost task 0.3 in
stage 5.0 (TID 11, ip-172-31-12-174.ec2.internal, executor 19): org.apa
che.spark.api.python.PythonException: Traceback (most recent call las
t):
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/sql/types.
py", line 1527, in __getattr__
    idx = self.__fields__.index(item)
ValueError: 'notifications' is not in list

During handling of the above exception, another exception occurred:

Traceback (most recent call last):
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/worker.p
y", line 377, in main
    process()
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/worker.p
y", line 372, in process
    serializer.dump_stream(func(split_index, iterator), outfile)
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/serializer
s.py", line 393, in dump_stream
    vs = list(itertools.islice(iterator, batch))
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/util.py",
line 113, in wrapper
    return f(*args, **kwargs)
  File "<stdin>", line 1, in <lambda>
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/sql/types.
py", line 1532, in __getattr__
    raise AttributeError(item)
AttributeError: notifications

        at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.
handlePythonException(PythonRunner.scala:456)
        at org.apache.spark.api.python.PythonRunner$$anon$1.read(Python
Runner.scala:592)
        at org.apache.spark.api.python.PythonRunner$$anon$1.read(Python
Runner.scala:575)
        at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.
hasNext(PythonRunner.scala:410)
        at org.apache.spark.InterruptibleIterator.hasNext(Interruptible
Iterator.scala:37)
        at scala.collection.Iterator$class.foreach(Iterator.scala:891)
        at org.apache.spark.InterruptibleIterator.foreach(Interruptible
Iterator.scala:28)
        at scala.collection.generic.Growable$class.$plus$plus$eq(Growab
le.scala:59)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuff
er.scala:104)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuff
```

```
er.scala:48)
        at scala.collection.TraversableOnce$class.to(TraversableOnce.sc
ala:310)
        at org.apache.spark.InterruptibleIterator.to(InterruptibleItera
tor.scala:28)
        at scala.collection.TraversableOnce$class.toBuffer(TraversableO
nce.scala:302)
        at org.apache.spark.InterruptibleIterator.toBuffer(Interruptibl
eIterator.scala:28)
        at scala.collection.TraversableOnce$class.toArray(TraversableOn
ce.scala:289)
        at org.apache.spark.InterruptibleIterator.toArray(Interruptible
Iterator.scala:28)
        at org.apache.spark.rdd.RDD$$anonfun$collect$1$$anonfun$13.appl
y(RDD.scala:945)
        at org.apache.spark.rdd.RDD$$anonfun$collect$1$$anonfun$13.appl
y(RDD.scala:945)
        at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkC
ontext.scala:2101)
        at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkC
ontext.scala:2101)
        at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.sca
la:90)
        at org.apache.spark.scheduler.Task.run(Task.scala:123)
        at org.apache.spark.executor.Executor$TaskRunner$$anonfun$10.ap
ply(Executor.scala:408)
        at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:
1360)
        at org.apache.spark.executor.Executor$TaskRunner.run(Executor.s
cala:414)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPool
Executor.java:1149)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoo
lExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)

Driver stacktrace:
        at org.apache.spark.scheduler.DAGScheduler.org$apache$spark$sch
eduler$DAGScheduler$$failJobAndIndependentStages(DAGScheduler.scala:204
1)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage
$1.apply(DAGScheduler.scala:2029)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage
$1.apply(DAGScheduler.scala:2028)
        at scala.collection.mutable.ResizableArray$class.foreach(Resiza
bleArray.scala:59)
        at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.sca
la:48)
        at org.apache.spark.scheduler.DAGScheduler.abortStage(DAGSchedu
ler.scala:2028)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskS
etFailed$1.apply(DAGScheduler.scala:966)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskS
etFailed$1.apply(DAGScheduler.scala:966)
        at scala.Option.foreach(Option.scala:257)
        at org.apache.spark.scheduler.DAGScheduler.handleTaskSetFailed
(DAGScheduler.scala:966)
```

```
        at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.doOn
Receive(DAGScheduler.scala:2262)
        at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onRe
ceive(DAGScheduler.scala:2211)
        at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onRe
ceive(DAGScheduler.scala:2200)
        at org.apache.spark.util.EventLoop$$anon$1.run(EventLoop.scala:
49)
        at org.apache.spark.scheduler.DAGScheduler.runJob(DAGScheduler.
scala:777)
        at org.apache.spark.SparkContext.runJob(SparkContext.scala:206
1)
        at org.apache.spark.SparkContext.runJob(SparkContext.scala:208
2)
        at org.apache.spark.SparkContext.runJob(SparkContext.scala:210
1)
        at org.apache.spark.SparkContext.runJob(SparkContext.scala:212
6)
        at org.apache.spark.rdd.RDD$$anonfun$collect$1.apply(RDD.scala:
945)
        at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperati
onScope.scala:151)
        at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperati
onScope.scala:112)
        at org.apache.spark.rdd.RDD.withScope(RDD.scala:363)
        at org.apache.spark.rdd.RDD.collect(RDD.scala:944)
        at org.apache.spark.api.python.PythonRDD$.collectAndServe(Pytho
nRDD.scala:166)
        at org.apache.spark.api.python.PythonRDD.collectAndServe(Python
RDD.scala)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAcce
ssorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMe
thodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:244)
        at py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.jav
a:357)
        at py4j.Gateway.invoke(Gateway.java:282)
        at py4j.commands.AbstractCommand.invokeMethod(AbstractCommand.j
ava:132)
        at py4j.commands.CallCommand.execute(CallCommand.java:79)
        at py4j.GatewayConnection.run(GatewayConnection.java:238)
        at java.lang.Thread.run(Thread.java:748)
Caused by: org.apache.spark.api.python.PythonException: Traceback (most
recent call last):
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/sql/types.
py", line 1527, in __getattr__
    idx = self.__fields__.index(item)
ValueError: 'notifications' is not in list

During handling of the above exception, another exception occurred:

Traceback (most recent call last):
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
```

```
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/worker.p
y", line 377, in main
    process()
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/worker.p
y", line 372, in process
    serializer.dump_stream(func(split_index, iterator), outfile)
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/serializer
s.py", line 393, in dump_stream
    vs = list(itertools.islice(iterator, batch))
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/util.py",
line 113, in wrapper
    return f(*args, **kwargs)
  File "<stdin>", line 1, in <lambda>
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/sql/types.
py", line 1532, in __getattr__
    raise AttributeError(item)
AttributeError: notifications

        at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.
handlePythonException(PythonRunner.scala:456)
        at org.apache.spark.api.python.PythonRunner$$anon$1.read(Python
Runner.scala:592)
        at org.apache.spark.api.python.PythonRunner$$anon$1.read(Python
Runner.scala:575)
        at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.
hasNext(PythonRunner.scala:410)
        at org.apache.spark.InterruptibleIterator.hasNext(Interruptible
Iterator.scala:37)
        at scala.collection.Iterator$class.foreach(Iterator.scala:891)
        at org.apache.spark.InterruptibleIterator.foreach(Interruptible
Iterator.scala:28)
        at scala.collection.generic.Growable$class.$plus$plus$eq(Growab
le.scala:59)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuff
er.scala:104)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuff
er.scala:48)
        at scala.collection.TraversableOnce$class.to(TraversableOnce.sc
ala:310)
        at org.apache.spark.InterruptibleIterator.to(InterruptibleItera
tor.scala:28)
        at scala.collection.TraversableOnce$class.toBuffer(TraversableO
nce.scala:302)
        at org.apache.spark.InterruptibleIterator.toBuffer(Interruptibl
eIterator.scala:28)
        at scala.collection.TraversableOnce$class.toArray(TraversableOn
ce.scala:289)
        at org.apache.spark.InterruptibleIterator.toArray(Interruptible
Iterator.scala:28)
        at org.apache.spark.rdd.RDD$$anonfun$collect$1$$anonfun$13.appl
y(RDD.scala:945)
        at org.apache.spark.rdd.RDD$$anonfun$collect$1$$anonfun$13.appl
y(RDD.scala:945)
```

```
        at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkC
ontext.scala:2101)
        at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkC
ontext.scala:2101)
        at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.sca
la:90)
        at org.apache.spark.scheduler.Task.run(Task.scala:123)
        at org.apache.spark.executor.Executor$TaskRunner$$anonfun$10.ap
ply(Executor.scala:408)
        at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:
1360)
        at org.apache.spark.executor.Executor$TaskRunner.run(Executor.s
cala:414)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPool
Executor.java:1149)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoo
lExecutor.java:624)
        ... 1 more

Traceback (most recent call last):
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/rdd.py", line 81
6, in collect
    sock_info = self.ctx._jvm.PythonRDD.collectAndServe(self._jrdd.rdd
())
  File "/usr/lib/spark/python/lib/py4j-0.10.7-src.zip/py4j/java_gatewa
y.py", line 1257, in __call__
    answer, self.gateway_client, self.target_id, self.name)
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/sql/utils.py", li
ne 63, in deco
    return f(*a, **kw)
  File "/usr/lib/spark/python/lib/py4j-0.10.7-src.zip/py4j/protocol.p
y", line 328, in get_return_value
    format(target_id, ".", name), value)
py4j.protocol.Py4JJavaError: An error occurred while calling z:org.apac
he.spark.api.python.PythonRDD.collectAndServe.
: org.apache.spark.SparkException: Job aborted due to stage failure: Ta
sk 0 in stage 5.0 failed 4 times, most recent failure: Lost task 0.3 in
stage 5.0 (TID 11, ip-172-31-12-174.ec2.internal, executor 19): org.apa
che.spark.api.python.PythonException: Traceback (most recent call las
t):
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/sql/types.
py", line 1527, in __getattr__
    idx = self.__fields__.index(item)
ValueError: 'notifications' is not in list

During handling of the above exception, another exception occurred:

Traceback (most recent call last):
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/worker.p
y", line 377, in main
    process()
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/worker.p
y", line 372, in process
    serializer.dump_stream(func(split_index, iterator), outfile)
```

```
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/serializer
s.py", line 393, in dump_stream
    vs = list(itertools.islice(iterator, batch))
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/util.py",
line 113, in wrapper
    return f(*args, **kwargs)
  File "<stdin>", line 1, in <lambda>
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/sql/types.
py", line 1532, in __getattr__
    raise AttributeError(item)
AttributeError: notifications

        at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.
handlePythonException(PythonRunner.scala:456)
        at org.apache.spark.api.python.PythonRunner$$anon$1.read(Python
Runner.scala:592)
        at org.apache.spark.api.python.PythonRunner$$anon$1.read(Python
Runner.scala:575)
        at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.
hasNext(PythonRunner.scala:410)
        at org.apache.spark.InterruptibleIterator.hasNext(Interruptible
Iterator.scala:37)
        at scala.collection.Iterator$class.foreach(Iterator.scala:891)
        at org.apache.spark.InterruptibleIterator.foreach(Interruptible
Iterator.scala:28)
        at scala.collection.generic.Growable$class.$plus$plus$eq(Growab
le.scala:59)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuff
er.scala:104)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuff
er.scala:48)
        at scala.collection.TraversableOnce$class.to(TraversableOnce.sc
ala:310)
        at org.apache.spark.InterruptibleIterator.to(InterruptibleItera
tor.scala:28)
        at scala.collection.TraversableOnce$class.toBuffer(TraversableO
nce.scala:302)
        at org.apache.spark.InterruptibleIterator.toBuffer(Interruptibl
eIterator.scala:28)
        at scala.collection.TraversableOnce$class.toArray(TraversableOn
ce.scala:289)
        at org.apache.spark.InterruptibleIterator.toArray(Interruptible
Iterator.scala:28)
        at org.apache.spark.rdd.RDD$$anonfun$collect$1$$anonfun$13.appl
y(RDD.scala:945)
        at org.apache.spark.rdd.RDD$$anonfun$collect$1$$anonfun$13.appl
y(RDD.scala:945)
        at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkC
ontext.scala:2101)
        at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkC
ontext.scala:2101)
        at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.sca
la:90)
        at org.apache.spark.scheduler.Task.run(Task.scala:123)
```

```
        at org.apache.spark.executor.Executor$TaskRunner$$anonfun$10.ap
ply(Executor.scala:408)
        at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:
1360)
        at org.apache.spark.executor.Executor$TaskRunner.run(Executor.s
cala:414)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPool
Executor.java:1149)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoo
lExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)

Driver stacktrace:
        at org.apache.spark.scheduler.DAGScheduler.org$apache$spark$sch
eduler$DAGScheduler$$failJobAndIndependentStages(DAGScheduler.scala:204
1)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage
$1.apply(DAGScheduler.scala:2029)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage
$1.apply(DAGScheduler.scala:2028)
        at scala.collection.mutable.ResizableArray$class.foreach(Resiza
bleArray.scala:59)
        at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.sca
la:48)
        at org.apache.spark.scheduler.DAGScheduler.abortStage(DAGSchedu
ler.scala:2028)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskS
etFailed$1.apply(DAGScheduler.scala:966)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskS
etFailed$1.apply(DAGScheduler.scala:966)
        at scala.Option.foreach(Option.scala:257)
        at org.apache.spark.scheduler.DAGScheduler.handleTaskSetFailed
(DAGScheduler.scala:966)
        at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.doOn
Receive(DAGScheduler.scala:2262)
        at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onRe
ceive(DAGScheduler.scala:2211)
        at org.apache.spark.scheduler.DAGSchedulerEventProcessLoop.onRe
ceive(DAGScheduler.scala:2200)
        at org.apache.spark.util.EventLoop$$anon$1.run(EventLoop.scala:
49)
        at org.apache.spark.scheduler.DAGScheduler.runJob(DAGScheduler.
scala:777)
        at org.apache.spark.SparkContext.runJob(SparkContext.scala:206
1)
        at org.apache.spark.SparkContext.runJob(SparkContext.scala:208
2)
        at org.apache.spark.SparkContext.runJob(SparkContext.scala:210
1)
        at org.apache.spark.SparkContext.runJob(SparkContext.scala:212
6)
        at org.apache.spark.rdd.RDD$$anonfun$collect$1.apply(RDD.scala:
945)
        at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperati
onScope.scala:151)
        at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperati
onScope.scala:112)
```

```
        at org.apache.spark.rdd.RDD.withScope(RDD.scala:363)
        at org.apache.spark.rdd.RDD.collect(RDD.scala:944)
        at org.apache.spark.api.python.PythonRDD$.collectAndServe(Pytho
nRDD.scala:166)
        at org.apache.spark.api.python.PythonRDD.collectAndServe(Python
RDD.scala)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAcce
ssorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMe
thodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:244)
        at py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.jav
a:357)
        at py4j.Gateway.invoke(Gateway.java:282)
        at py4j.commands.AbstractCommand.invokeMethod(AbstractCommand.j
ava:132)
        at py4j.commands.CallCommand.execute(CallCommand.java:79)
        at py4j.GatewayConnection.run(GatewayConnection.java:238)
        at java.lang.Thread.run(Thread.java:748)
Caused by: org.apache.spark.api.python.PythonException: Traceback (most
recent call last):
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/sql/types.
py", line 1527, in __getattr__
    idx = self.__fields__.index(item)
ValueError: 'notifications' is not in list

During handling of the above exception, another exception occurred:

Traceback (most recent call last):
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/worker.p
y", line 377, in main
    process()
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/worker.p
y", line 372, in process
    serializer.dump_stream(func(split_index, iterator), outfile)
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/serializer
s.py", line 393, in dump_stream
    vs = list(itertools.islice(iterator, batch))
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/util.py",
line 113, in wrapper
    return f(*args, **kwargs)
  File "<stdin>", line 1, in <lambda>
  File "/mnt/yarn/usercache/livy/appcache/application_1572637375434_000
1/container_1572637375434_0001_01_000087/pyspark.zip/pyspark/sql/types.
py", line 1532, in __getattr__
    raise AttributeError(item)
AttributeError: notifications

        at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.
handlePythonException(PythonRunner.scala:456)
```

```
        at org.apache.spark.api.python.PythonRunner$$anon$1.read(Python
Runner.scala:592)
        at org.apache.spark.api.python.PythonRunner$$anon$1.read(Python
Runner.scala:575)
        at org.apache.spark.api.python.BasePythonRunner$ReaderIterator.
hasNext(PythonRunner.scala:410)
        at org.apache.spark.InterruptibleIterator.hasNext(Interruptible
Iterator.scala:37)
        at scala.collection.Iterator$class.foreach(Iterator.scala:891)
        at org.apache.spark.InterruptibleIterator.foreach(Interruptible
Iterator.scala:28)
        at scala.collection.generic.Growable$class.$plus$plus$eq(Growab
le.scala:59)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuff
er.scala:104)
        at scala.collection.mutable.ArrayBuffer.$plus$plus$eq(ArrayBuff
er.scala:48)
        at scala.collection.TraversableOnce$class.to(TraversableOnce.sc
ala:310)
        at org.apache.spark.InterruptibleIterator.to(InterruptibleItera
tor.scala:28)
        at scala.collection.TraversableOnce$class.toBuffer(TraversableO
nce.scala:302)
        at org.apache.spark.InterruptibleIterator.toBuffer(Interruptibl
eIterator.scala:28)
        at scala.collection.TraversableOnce$class.toArray(TraversableOn
ce.scala:289)
        at org.apache.spark.InterruptibleIterator.toArray(Interruptible
Iterator.scala:28)
        at org.apache.spark.rdd.RDD$$anonfun$collect$1$$anonfun$13.appl
y(RDD.scala:945)
        at org.apache.spark.rdd.RDD$$anonfun$collect$1$$anonfun$13.appl
y(RDD.scala:945)
        at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkC
ontext.scala:2101)
        at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkC
ontext.scala:2101)
        at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.sca
la:90)
        at org.apache.spark.scheduler.Task.run(Task.scala:123)
        at org.apache.spark.executor.Executor$TaskRunner$$anonfun$10.ap
ply(Executor.scala:408)
        at org.apache.spark.util.Utils$.tryWithSafeFinally(Utils.scala:
1360)
        at org.apache.spark.executor.Executor$TaskRunner.run(Executor.s
cala:414)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPool
Executor.java:1149)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoo
lExecutor.java:624)
        ... 1 more
```

In [19]:

```
```

6

In [ ]:

```
# map
rdd = congress_sample
sorted(rdd.map(lambda x: (x, 1)).collect())
```

In [ ]:

```
#Reduce
wordsRDD = sc.parallelize(congress_sample, 100)
wordCountsCollected = (wordsRDD
                       .map(lambda w: (w, 1))
                       .reduceByKey(lambda x,y: x+y)
                       .collect())
```

In [ ]:

```
```

In [39]:

```
congress_sample = spark.read.text("s3://us-congress-tweets/raw/files.txt")
congress_sample.collect()
```

[Row(value='s3://us-congress-tweets/raw/part-00000.snappy'), Row(value
='s3://us-congress-tweets/raw/part-00001.snappy'), Row(value='s3://us-c
ongress-tweets/raw/part-00002.snappy'), Row(value='s3://us-congress-twe
ets/raw/part-00003.snappy'), Row(value='s3://us-congress-tweets/raw/par
t-00004.snappy'), Row(value='s3://us-congress-tweets/raw/part-00005.sna
ppy'), Row(value='s3://us-congress-tweets/raw/part-00006.snappy'), Row
(value='s3://us-congress-tweets/raw/part-00007.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00008.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00009.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00010.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00011.snappy'), Row(value='s3://us-congress-tweets/raw/part-00012.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00013.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00014.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00015.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00016.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00017.snappy'), Row(value='s3://us-congress-tweets/raw/part-00018.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00019.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00020.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00021.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00022.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00023.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00024.snappy'), Row(value='s3://us-congress-tweets/raw/part-00025.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00026.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00027.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00028.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00029.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00030.snappy'), Row(value='s3://us-congress-tweets/raw/part-00031.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00032.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00033.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00034.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00035.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00036.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00037.snappy'), Row(value='s3://us-congress-tweets/raw/part-00038.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00039.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00040.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00041.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00042.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00043.snappy'), Row(value='s3://us-congress-tweets/raw/part-00044.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00045.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00046.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00047.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00048.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00049.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00050.snappy'), Row(value='s3://us-congress-tweets/raw/part-00051.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00052.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00053.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00054.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00055.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00056.snappy'), Row(value='s3://us-congress-tweets/raw/part-00057.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00058.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00059.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00060.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00061.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00062.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00063.snappy'), Row(value='s3://us-congress-tweets/raw/part-00064.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00065.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00066.snappy'), Row(value='s3://u

s-congress-tweets/raw/part-00067.snappy'), Row(value='s3://us-congress-tweets/raw/part-00068.snappy'), Row(value='s3://us-congress-tweets/raw/part-00069.snappy'), Row(value='s3://us-congress-tweets/raw/part-00070.snappy'), Row(value='s3://us-congress-tweets/raw/part-00071.snappy'), Row(value='s3://us-congress-tweets/raw/part-00072.snappy'), Row(value='s3://us-congress-tweets/raw/part-00073.snappy'), Row(value='s3://us-congress-tweets/raw/part-00074.snappy'), Row(value='s3://us-congress-tweets/raw/part-00075.snappy'), Row(value='s3://us-congress-tweets/raw/part-00076.snappy'), Row(value='s3://us-congress-tweets/raw/part-00077.snappy'), Row(value='s3://us-congress-tweets/raw/part-00078.snappy'), Row(value='s3://us-congress-tweets/raw/part-00079.snappy'), Row(value='s3://us-congress-tweets/raw/part-00080.snappy'), Row(value='s3://us-congress-tweets/raw/part-00081.snappy'), Row(value='s3://us-congress-tweets/raw/part-00082.snappy'), Row(value='s3://us-congress-tweets/raw/part-00083.snappy'), Row(value='s3://us-congress-tweets/raw/part-00084.snappy'), Row(value='s3://us-congress-tweets/raw/part-00085.snappy'), Row(value='s3://us-congress-tweets/raw/part-00086.snappy'), Row(value='s3://us-congress-tweets/raw/part-00087.snappy'), Row(value='s3://us-congress-tweets/raw/part-00088.snappy'), Row(value='s3://us-congress-tweets/raw/part-00089.snappy'), Row(value='s3://us-congress-tweets/raw/part-00090.snappy'), Row(value='s3://us-congress-tweets/raw/part-00091.snappy'), Row(value='s3://us-congress-tweets/raw/part-00092.snappy'), Row(value='s3://us-congress-tweets/raw/part-00093.snappy'), Row(value='s3://us-congress-tweets/raw/part-00094.snappy'), Row(value='s3://us-congress-tweets/raw/part-00095.snappy'), Row(value='s3://us-congress-tweets/raw/part-00096.snappy'), Row(value='s3://us-congress-tweets/raw/part-00097.snappy'), Row(value='s3://us-congress-tweets/raw/part-00098.snappy'), Row(value='s3://us-congress-tweets/raw/part-00099.snappy'), Row(value='s3://us-congress-tweets/raw/part-00100.snappy'), Row(value='s3://us-congress-tweets/raw/part-00101.snappy'), Row(value='s3://us-congress-tweets/raw/part-00102.snappy'), Row(value='s3://us-congress-tweets/raw/part-00103.snappy'), Row(value='s3://us-congress-tweets/raw/part-00104.snappy'), Row(value='s3://us-congress-tweets/raw/part-00105.snappy'), Row(value='s3://us-congress-tweets/raw/part-00106.snappy'), Row(value='s3://us-congress-tweets/raw/part-00107.snappy'), Row(value='s3://us-congress-tweets/raw/part-00108.snappy'), Row(value='s3://us-congress-tweets/raw/part-00109.snappy'), Row(value='s3://us-congress-tweets/raw/part-00110.snappy'), Row(value='s3://us-congress-tweets/raw/part-00111.snappy'), Row(value='s3://us-congress-tweets/raw/part-00112.snappy'), Row(value='s3://us-congress-tweets/raw/part-00113.snappy'), Row(value='s3://us-congress-tweets/raw/part-00114.snappy'), Row(value='s3://us-congress-tweets/raw/part-00115.snappy'), Row(value='s3://us-congress-tweets/raw/part-00116.snappy'), Row(value='s3://us-congress-tweets/raw/part-00117.snappy'), Row(value='s3://us-congress-tweets/raw/part-00118.snappy'), Row(value='s3://us-congress-tweets/raw/part-00119.snappy'), Row(value='s3://us-congress-tweets/raw/part-00120.snappy'), Row(value='s3://us-congress-tweets/raw/part-00121.snappy'), Row(value='s3://us-congress-tweets/raw/part-00122.snappy'), Row(value='s3://us-congress-tweets/raw/part-00123.snappy'), Row(value='s3://us-congress-tweets/raw/part-00124.snappy'), Row(value='s3://us-congress-tweets/raw/part-00125.snappy'), Row(value='s3://us-congress-tweets/raw/part-00126.snappy'), Row(value='s3://us-congress-tweets/raw/part-00127.snappy'), Row(value='s3://us-congress-tweets/raw/part-00128.snappy'), Row(value='s3://us-congress-tweets/raw/part-00129.snappy'), Row(value='s3://us-congress-tweets/raw/part-00130.snappy'), Row(value='s3://us-congress-tweets/raw/part-00131.snappy'), Row(value='s3://us-congress-tweets/raw/part-00132.snappy'), Row(value='s3://us-congress-tweets/raw/part-00133.snappy'), Row(value='s3://us-congress-tweets/raw/

```
part-00134.snappy'), Row(value='s3://us-congress-tweets/raw/part-00135.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00136.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00137.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00138.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00139.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00140.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00141.snappy'), Row(value='s3://us-congress-tweets/raw/part-00142.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00143.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00144.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00145.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00146.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00147.snappy'), Row(value='s3://us-congress-tweets/raw/part-00148.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00149.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00150.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00151.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00152.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00153.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00154.snappy'), Row(value='s3://us-congress-tweets/raw/part-00155.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00156.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00157.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00158.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00159.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00160.snappy'), Row(value='s3://us-congress-tweets/raw/part-00161.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00162.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00163.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00164.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00165.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00166.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00167.snappy'), Row(value='s3://us-congress-tweets/raw/part-00168.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00169.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00170.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00171.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00172.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00173.snappy'), Row(value='s3://us-congress-tweets/raw/part-00174.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00175.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00176.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00177.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00178.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00179.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00180.snappy'), Row(value='s3://us-congress-tweets/raw/part-00181.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00182.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00183.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00184.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00185.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00186.snappy'), Row(value='s3://us-congress-tweets/raw/part-00187.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00188.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00189.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00190.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00191.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00192.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00193.snappy'), Row(value='s3://us-congress-tweets/raw/part-00194.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00195.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00196.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00197.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00198.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00199.snappy'), Row(value='s3://us-congress-tweets/raw/part-00200.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00201.snappy'), R
```

```
ow(value='s3://us-congress-tweets/raw/part-00202.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00203.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00204.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00205.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00206.snappy'), Row(value='s3://us-congress-tweets/raw/part-00207.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00208.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00209.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00210.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00211.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00212.snappy'), Row(value='s3://us-congress-tweets/raw/part-00213.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00214.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00215.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00216.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00217.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00218.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00219.snappy'), Row(value='s3://us-congress-tweets/raw/part-00220.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00221.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00222.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00223.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00224.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00225.snappy'), Row(value='s3://us-congress-tweets/raw/part-00226.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00227.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00228.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00229.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00230.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00231.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00232.snappy'), Row(value='s3://us-congress-tweets/raw/part-00233.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00234.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00235.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00236.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00237.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00238.snappy'), Row(value='s3://us-congress-tweets/raw/part-00239.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00240.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00241.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00242.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00243.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00244.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00245.snappy'), Row(value='s3://us-congress-tweets/raw/part-00246.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00247.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00248.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00249.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00250.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00251.snappy'), Row(value='s3://us-congress-tweets/raw/part-00252.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00253.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00254.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00255.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00256.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00257.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00258.snappy'), Row(value='s3://us-congress-tweets/raw/part-00259.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00260.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00261.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00262.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00263.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00264.snappy'), Row(value='s3://us-congress-tweets/raw/part-00265.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00266.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00267.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00268.snappy'), Row(value='s3://us-cong
```

```
ress-tweets/raw/part-00269.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00270.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00271.snappy'), Row(value='s3://us-congress-tweets/raw/part-00272.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00273.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00274.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00275.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00276.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00277.snappy'), Row(value='s3://us-congress-tweets/raw/part-00278.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00279.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00280.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00281.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00282.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00283.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00284.snappy'), Row(value='s3://us-congress-tweets/raw/part-00285.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00286.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00287.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00288.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00289.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00290.snappy'), Row(value='s3://us-congress-tweets/raw/part-00291.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00292.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00293.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00294.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00295.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00296.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00297.snappy'), Row(value='s3://us-congress-tweets/raw/part-00298.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00299.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00300.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00301.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00302.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00303.snappy'), Row(value='s3://us-congress-tweets/raw/part-00304.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00305.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00306.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00307.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00308.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00309.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00310.snappy'), Row(value='s3://us-congress-tweets/raw/part-00311.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00312.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00313.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00314.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00315.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00316.snappy'), Row(value='s3://us-congress-tweets/raw/part-00317.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00318.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00319.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00320.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00321.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00322.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00323.snappy'), Row(value='s3://us-congress-tweets/raw/part-00324.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00325.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00326.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00327.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00328.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00329.snappy'), Row(value='s3://us-congress-tweets/raw/part-00330.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00331.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00332.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00333.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00334.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00335.snappy'), Row(value='s3://us-congress-tweets/raw/part-
```

```
00336.snappy'), Row(value='s3://us-congress-tweets/raw/part-00337.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00338.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00339.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00340.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00341.snappy'), Row(value='s3://us-congress-tweets/raw/
part-00342.snappy'), Row(value='s3://us-congress-tweets/raw/part-00343.
snappy'), Row(value='s3://us-congress-tweets/raw/part-00344.snappy'), R
ow(value='s3://us-congress-tweets/raw/part-00345.snappy'), Row(value='s
3://us-congress-tweets/raw/part-00346.snappy'), Row(value='s3://us-cong
ress-tweets/raw/part-00347.snappy'), Row(value='s3://us-congress-tweet
s/raw/part-00348.snappy'), Row(value='s3://us-congress-tweets/raw/part-
00349.snappy'), Row(value='s3://us-congress-tweets/raw/part-00350.snapp
y'), Row(value='s3://us-congress-tweets/raw/part-00351.snappy'), Row(va
lue='s3://us-congress-tweets/raw/part-00352.snappy'), Row(value='s3://u
s-congress-tweets/raw/part-00353.snappy'), Row(value='s3://us-congress-
tweets/raw/part-00354.snappy')]
```

In [ ]: