In [3]:

```
'''
(30 points) Part 5: Community Detection

User-hashtag relations have been extracted and saved in the file s3://us-congress-t
weets/user_hashtags.csv. If a user uses a hashtag there will be a record with the u
serid and the hashtag.
Use the Trawling algorithm discussed in class to find potential user communities in
the dataset. (Hint: use FPGrowth in the Spark ML package). Explore different values
for the support parameter.
'''
```

```
An error was encountered:
Invalid status code '404' from https://172.31.6.195:18888/sessions/1 wi
th error payload: {"msg":"Session '1' not found."}
```

In [2]:

```
user_hashtags = spark.read.csv("s3://us-congress-tweets/user_hashtags.csv", header=
True)

reply_network = spark.read.csv("s3://us-congress-tweets/reply_network.csv", header=
True)
```

```
'path s3://aws-logs-358879944178-us-east-1/tweets_all already exists.;'
Traceback (most recent call last):
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/sql/readwriter.p
y", line 931, in csv
    self._jwrite.csv(path)
  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 69, in deco
    raise AnalysisException(s.split(': ', 1)[1], stackTrace)
pyspark.sql.utils.AnalysisException: 'path s3://aws-logs-358879944178-u
s-east-1/tweets_all already exists.;'
```

In [ ]:

```

```