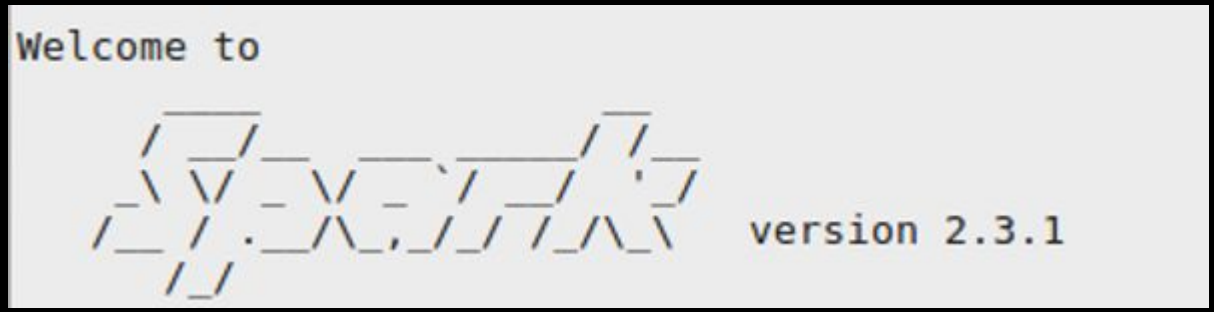
Welcome to
version 2.3.1

```
>>> thing = spark.read.csv("/usr/spark-2.3.0/data/coverage.txt")

>>> thing.groupBy("_c0").count().show()


|            _c0 |  count
 -------------------------
|  building-or |          3
|  lowenstam |    7948
|  aa-images | 128285
 -------------------------
```

```
>>> thing.createOrReplaceTempView("thingview")
>>> query = spark.sql("""SELECT _c0, COUNT(*) AS num FROM thingview
GROUP BY _c0""")
>>> query.show()
+-------------+------+
|          _c0|   num|
+-------------+------+
| building-or |     3|
| lowenstam   |  7948|
| aa-images   |128285|
+-------------+------+
```

```
f = open('/usr/spark-2.3.0/data/oimb2_data.txt', 'r')
for line in f:
  arr = line.split('\t')
  for i in range(2,8):
    if bool(arr[i].strip()):
      query = spark.sql("SELECT %s FROM thingview
        WHERE title = '%s'" %(fields[i], arr[1]))
      j = query.toJSON().first()
      print(arr[0], j)
```

https://spark.apache.org

https://github.com/spark4lib/code4lib2018

https://gist.github.com/lsat12357

Linda Sato, University of Oregon