Spark Commands

Run pyspark and then execute the following commands in the spark environment.

```
rdd=sc.parallelize(range(1,1000))
rdd

#collect
x = sc.parallelize([1,2,3,4,5])
y = x.collect()
print(x)  # distributed
print(y)  # not distributed

# take
y = x.take(num = 3)
print(y)

# first
y = x.first()
print(y)

# filter
y = x.filter(lambda x: x%2 == 1)  # filters out even
elements
print(y.collect())

# map
```

```
y = x.map(lambda x: (x,x**2))
y.collect()
```

reduce

```
y = x.reduce(lambda obj, accumulated: obj + accumulated)  # computes a cumulative sum
print(y)
```

reduceByKey

```
x = sc.parallelize([('B',1),('B',2),('A',3),('A',4),('A',5)])
y = x.reduceByKey(lambda v1, v2: v1 + v2)
print(y.collect())
```

MapReduce

```
x.map(lambda gender:(data[1],1).reduceByKey(lambda x,y:(x+y)).collect())
```

flatMap

```
x = sc.parallelize([1,2,3,4,5])
y1 = x.map(lambda x: (x, 100*x, x**2))
y2 = x.flatMap(lambda x: (x, 100*x, x**2))
print(x.collect())
print(y1.collect())
print(y2.collect())
```

union

```
x = sc.parallelize(['A','A','B'])
y = sc.parallelize(['D','C','A'])
z = x.union(y)
print(z.collect())
```

###Reading from Files

```
inputTxt=sc.textFile("input.txt")
inputTxt.take(10)
```