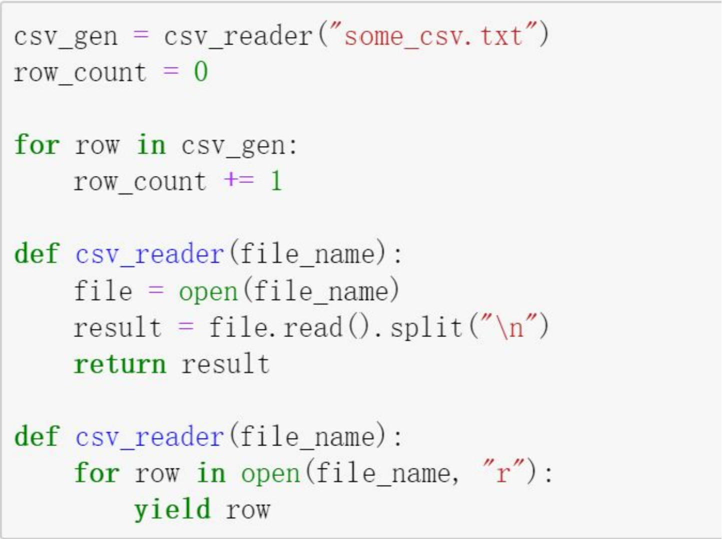
EECS E6893 Big Data Analytics
HW2: Friends Recommendation, GraphFrame (2)

Yunan Lu, yl4021@columbia.edu

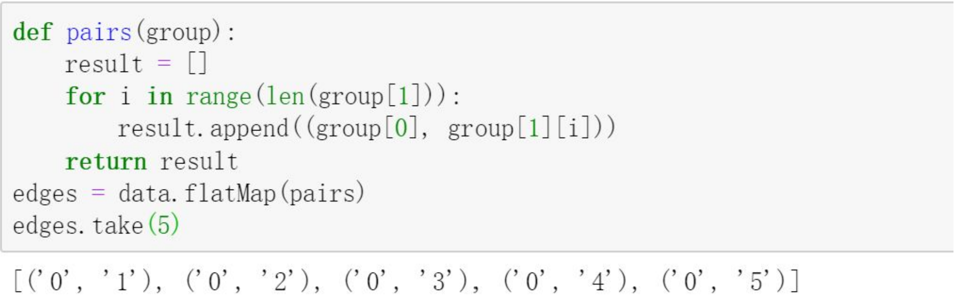
Yield

- When you use yield statement in any function, it turns it into a generator function.
- Generator functions are a special kind of function that return a lazy iterator. These are objects that you can loop over like a list. However, unlike lists, lazy iterators do not store their contents in memory.
- Reading Large File



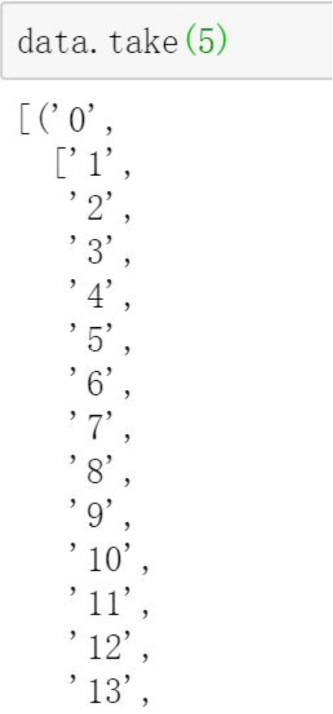
```
csv_gen = csv_reader(”some_csv.txt”)
row_count = 0

for row in csv_gen:
    row_count += 1

def csv_reader(file_name):
    file = open(file_name)
    result = file.read().split(”\n”)
    return result

def csv_reader(file_name):
    for row in open(file_name, ”r”):
        yield row
```

Yield vs Return



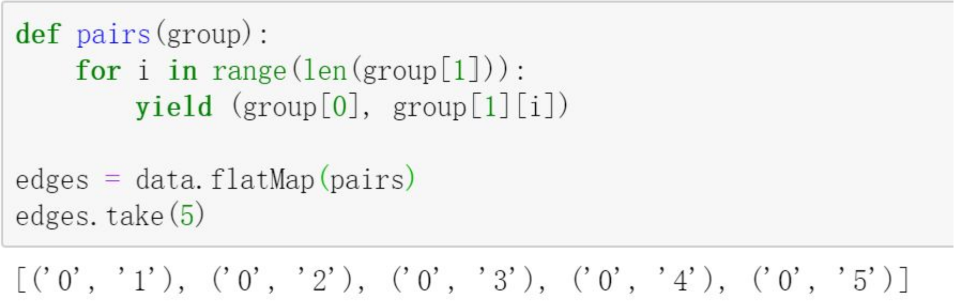
```
def pairs(group):
    result = []
    for i in range(len(group[1])):
        result.append((group[0], group[1][i]))
    return result
edges = data.flatMap(pairs)
edges.take(5)
```

```
[('0', '1'), ('0', '2'), ('0', '3'), ('0', '4'), ('0', '5')]
```



```
data.take(5)
```

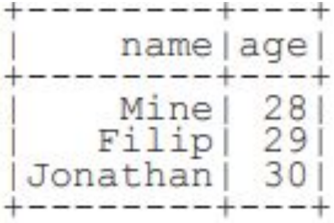
```
[('0',
  ['1',
   '2',
   '3',
   '4',
   '5',
   '6',
   '7',
   '8',
   '9',
   '10',
   '11',
   '12',
   '13',
```



```
def pairs(group):
    for i in range(len(group[1])):
        yield (group[0], group[1][i])

edges = data.flatMap(pairs)
edges.take(5)
```

```
[('0', '1'), ('0', '2'), ('0', '3'), ('0', '4'), ('0', '5')]
```

Spark DataFrames



```
+--------+---+
|    name|age|
+--------+---+
|    Mine| 28|
|   Filip| 29|
|Jonathan| 30|
+--------+---+
```

- df.groupBy("age").count().show()
- df.select("name").distinct().show()
- df.filter(df["age"]>24).show()
- df.sort("age", ascending=False).show()
- df.orderBy(["age","city"],ascending=[0,1]).show()   # orderby multiple columns
- df.dropDuplicates()
- df.na.fill(50).show()   # Replace null values
- df.na.drop().show()

Reference

- https://www.datacamp.com/community/blog/pyspark-sql-cheat-sheet