1 mrJob

mrJob (map-reduce job) is a python module enable python to communicate with Hadoop framework. Here is a simple word count application.

```
"""The classic MapReduce job: count the frequency of words.
"""
from mrjob.job import MRJob
import re

WORD_RE = re.compile(r"[\w']+")


class MRWordFreqCount(MRJob):

    def mapper(self, _, line):
        for word in WORD_RE.findall(line):
            yield (word.lower(), 1)

    def combiner(self, word, counts):
        yield (word, sum(counts))

    def reducer(self, word, counts):
        yield (word, sum(counts))


if __name__ == '__main__':
     MRWordFreqCount.run()
```

1.1 Another example is to calculate the histogram of students in our department. We can pipe the input text file into the code by using command

```
python status.pt < data.csv
```

```
#status.py
from mrjob.job import MRJob

class statusCount(MRJob):
 def getStatus(self, key, record):
       data = record.split(';')[3]
       yield data,1

 def frequency(self, state , value):
       yield state, sum(value)
 def sumMapper(self, state, freq):
       yield 1,freq
 def sumReducer(self, key, value):
       yield 'Number of students',sum(value)
 def steps(self):
       return [self.mr(mapper=self.getStatus,reducer=self.frequency),
             self.mr(mapper=self.sumMapper,reducer=self.sumReducer)]

if __name__ == '__main__' :
 statusCount.run()
```