Assignment 4 Enhong Ma course: 3030

```
# Homework: Explain what this does and how it does it
# Again make sure that you have the correct file location

from operator import add
```

.from operator package, import an accumulator operator – add. This will be used to calculate the parallel result from cluster node. In our case, only one local node.

```
f = sc.textFile("/var/tmp/temp1/spark/bin/shakespeare.txt")
```

.use textFile method of spark content object – sc to open text file on disk

```
wc = f.flatMap(lambda x: x.split(' ')).map(lambda x: (x, 1)).reduceByKey(add)
```

.first split the text file into separate word, using flatMap to create a list, one word on each line. Then map the text file and append number after each word through map function. Finally, use the accumulator operator – add to aggregate same word together. It generates two text files contains word counting .

```
wc.saveAsTextFile("wc_out.txt")
```

.**create directory wc_out.txt and write output to files under this directory.**