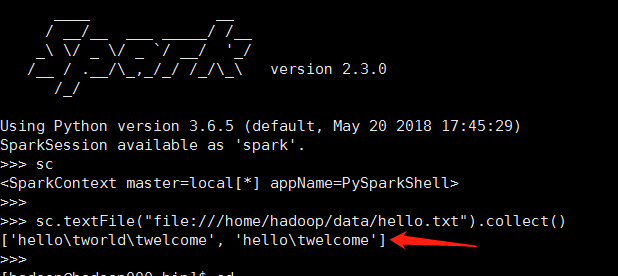
把外部文件创建为rdd：

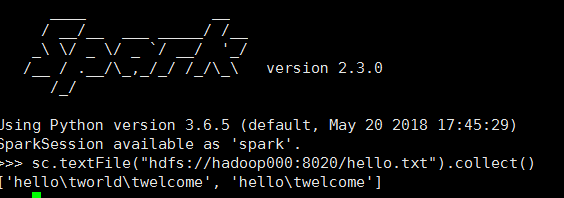
data=sc.textFile("file:///home/hadoop/data/hello.txt")



把hdfs 上的文件创建为rdd的方式:服务端口加上文件路径

sc.textFile("hdfs://hadoop000:8020/hello.txt")

在成产环境中需要保证数在每个datanode上的路径都一样



Rdd 保存到文件系统：

Data=[1,2,3]

Data.saveAsTextFile("file:///home/hadoop/data/output/")

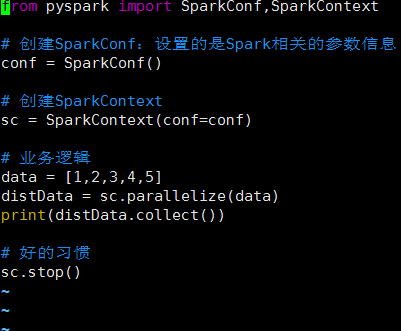


把脚本在linux 上运行：

在pycharm 上编辑好的脚本，上传或者复制到liunux 上就可以运行了

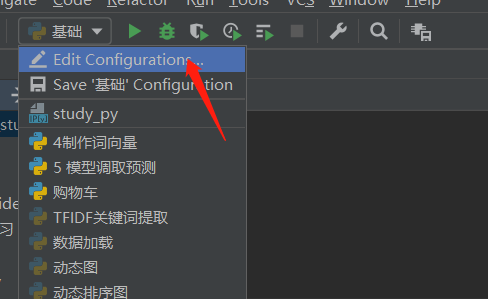
但不需要conf=pyspark.SparkConf().setMaster('local[2]').setAppName('sparktest')

改成：原因是不需要硬编码；msater 和appname 不需要设置了

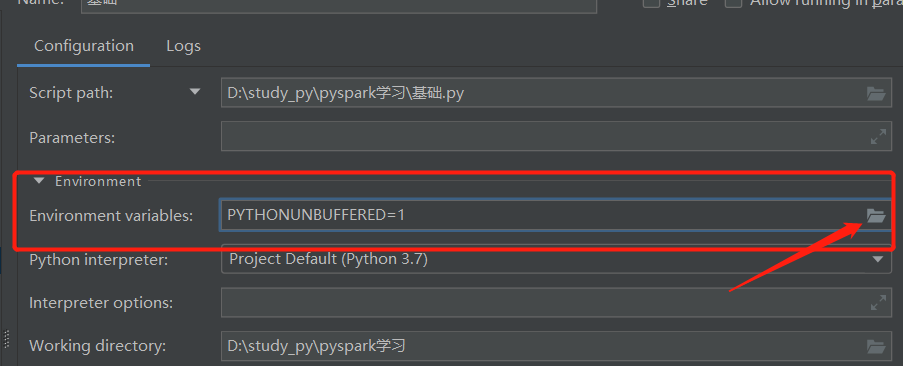


然后在spark/bin目录下提交脚本了： 输入命令： ./spark-submit –master local[2] –name spark0301 /home/Hadoop/script/spark0301

Pycharm 写spark 程序：

点击edit Configuration  


点击文件夹



三个模式：

Local 模式:

Standalone 模式：需要在集群上每个节点都部署spark，然后需要启动spark集群（需要启动master）

Yarn模式： 只需要一个节点，然后提交作业，不需要spark集群（不需要启动master）

此模式：./spark-submit –master yarn –name 随便取 /home./Hadoop/scricp/脚本地址