Spark 实践

通过 Spark 实现词频统计。

1. 下载与安装:

通过官网进行下载: https://spark.apache.org/downloads.html



We suggest the following location for your download:

https://dlcdn.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz

Alternate download locations are suggested below.

It is essential that you $\underbrace{\text{verify the integrity}}$ of the downloaded file using the PGP signature (.asc file) or a hash (.mds or .sha* file).

解压安装之后,安装成功,输入 pySpark:

```
Welcome to

----
/ __/ _ __ ____/ /__
__ / __/ _ _____/ /__
______/ /_____/ /___/ /__/ /__/
/__ / __/ /__/ /__/ /__/ version 3.5.0

Using Python version 3.11.5 (main, Sep 11 2023 08:31:25)
Spark context Web UI available at http://172.25.142.242:4040
Spark context available as 'sc' (master = local[*], app id = local-1701351226916).
SparkSession available as 'spark'.
```

2. 配置在 PyCharm 中调用 pySpark 的加载包

在 mac 上使用 pyspark,需要强行安装第三方包。将 spark 文件夹下的 pyspark 文件拷贝到 python/site-packages 文件夹目录下即可。

同时,也要在 Pycharm 中添加环境变量,将 spark 的安装路径复制过来。

3. 程序运行

运行下面这段词频统计代码:

from operator import add

from pyspark import SparkContext

文件 编辑 查看

good bad cool hadoop spark mlib good spark mlib cool spark bad

结果如下:

good: 2 hadoop: 1 bad: 2 cool: 2 spark: 3 mlib: 2