实验报告-HW11

文件结构

```
BIGDATASYSTEM_HW11
├── report.md
├── report.pdf
├── generator
└── simple_app_python
    ├── SimpleApp.py
    ├── log4j.properties
    ├── submit.sh
    ├── output_files
    └── stream
```

- `report.md`：实验报告
- `report.pdf`：实验报告PDF版本
- `generator`：生成数据的脚本文件夹
- `simple_app_python`：Spark Streaming应用程序文件夹
  - `SimpleApp.py`：Spark Streaming应用程序
  - `log4j.properties`：log4j配置文件
  - `submit.sh`：提交Spark应用程序的脚本
  - `output_files`：输出文件夹（存放应用程序5分钟内依次输出的文件）
  - `stream`：流数据文件夹（存放5分钟内产生的原始Streaming数据，从 `hdfs` 中拷贝得到）

运行方式

生成数据流：

```
cd generator
./generator.sh
```

获取结果：

```
cd simple_app_python
./submit.sh
```

代码逻辑

在原有程序基础上，首先通过 `getpass.getuser()` 获取当前用户名，得到待监听及存储检查点的 `hdfs` 文件路径；然后使用 `DStream` 的 `updateStateByKey` 方法，对每个 `word` 进行累积计数；最后使用 `foreachRDD` 方法，借助自定义的 `get_top_words` 函数，输出每个 `batch` 的 `top 100` 单词并将其存入文件。

实验结果

实验结果存放在 `simple_app_python/output_files` 文件夹中，每个文件对应一个 `batch` 的 `top 100` 单词；存放于 `simple_app_python/stream` 文件夹中的 `stream` 文件为拷贝自 `hdfs` 的原始数据流，可用于检查程序是否正确运行。具体控制台输出结果截图如下：

数据生成器：

```
2024210897@intro00:~/BigDataSystem_HW11/generator$ ./generator.sh
mkdir: `/user/2024210897/stream': File exists
Deleted /user/2024210897/stream/words.1733160338.txt
Deleted /user/2024210897/stream/words.1733160396.txt
Deleted /user/2024210897/stream/words.1733160455.txt
Deleted /user/2024210897/stream/words.1733160513.txt
Deleted /user/2024210897/stream/words.1733160571.txt
2024-12-03 02:31:42 generating words.1733164300.txt succeed
2024-12-03 02:32:40 generating words.1733164358.txt succeed
2024-12-03 02:33:38 generating words.1733164416.txt succeed
2024-12-03 02:34:36 generating words.1733164474.txt succeed
2024-12-03 02:35:35 generating words.1733164532.txt succeed
^C
2024210897@intro00:~/BigDataSystem_HW11/generator$
```

Spark Streaming应用程序：

```
k14 13
n25 13
o30 13
m23 13
m37 13
m33 13
o5 13
j22 13
h15 13
j2 13
o10 13
m1 13
m20 13
n20 13
n26 13
k31 13
n3 13
n32 13
k13 13
l1 12
p9 12
l10 12
i9 12
r3 12
p36 12
r1 12
p11 12
m13 12
Top 100 words for batch 4 saved to ./output_files/top_100_words_batch_4.txt
^CERROR:root:Exception while sending command.
Traceback (most recent call last):
  File "/spark/python/lib/py4j-0.10.9.7-src.zip/py4j/clientserver.py", line 511, in send_command
    answer = smart_decode(self.stream.readline()[:-1])
RuntimeError: reentrant call inside < io.BufferedReader name=3>
```