Big Data Management
Griffith College Dublin

Assignment 1

Rohin Mehra
Enrolment number: 3082862
MSc.Big Data Analysis & Management
1 April 2023

Abstract

In this assignment, we are using Apache Hadoop to perform operations on a CSV file data which holds the following column names, which are as follows:

1. date: Download date
2. time: Download time (in UTC)
3. size: Package size (in bytes)
4. r_version: Version of R used to download package
5. r_arch: Processor architecture (i386 = 32 bit, x86_64 = 64 bit)
6. r_os: Operating System (darwin9.8.0 = mac, mingw32 = windows)
7. package: Name of the package downloaded
8. country: Two-letter ISO country code
9. ip_id: A unique daily id assigned to each IP address

By using this information we will perform the task of MapReduce.

We will also try to seek correct answers to the following questions during the process of MapReduce.

1. Show the number of downloads for package ggplot2.
2. List the highest number of downloads by country?
3. What were the top 10 most popular packages?
4. What is the most popular package in Ireland?
5. What OS is most popular among R programmers?

Commands Used

Q1.

$HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming- *.jar -input /user/bdm/assignment/input -output /user/bdm/assignment/output -file /home/bdm/ assignment/mapper.py -file /home/bdm/assignment/reducer.py -mapper 'python3 mapper.py' -reducer 'python3 reducer.py'

Output: Number of downloads for package ggplot2: 22360632

Q2.

$HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-

streaming-*.jar -input /user/bdm/assignment/input -output /user/bdm/

assignment/output2 -file /home/bdm/assignment/mapper.py -file /home/bdm/

assignment/reducer2.py -mapper 'python3 mapper.py' -reducer 'python3

reducer2.py'

Output: Highest number of downloads by a country: "NA" with 3225550 downloads

Q3.

$HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-

streaming-*.jar -input /user/bdm/assignment/input -output /user/bdm/

assignment/output3 -file /home/bdm/assignment/mapper.py -file /home/bdm/

assignment/reducer3.py -mapper 'python3 mapper.py' -reducer 'python3 reducer3.py'

Output: Top 10 most popular packages:

1. "NA": 3225550 downloads
2. ""mingw32"": 3194919 downloads
3. ""US"": 3061236 downloads
4. ""linux-gnu"": 778523 downloads
5. ""darwin17.0"": 648165 downloads
6. ""GB"": 569535 downloads
7. ""darwin20"": 328304 downloads
8. ""CN"": 282214 downloads
9. ""KR"": 254392 downloads
10. ""DE"": 236903 downloads

Q4.

$HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-

streaming-*.jar -input /user/bdm/assignment/input -output /user/bdm/

assignment/output4 -file /home/bdm/assignment/mapper.py -file /home/bdm/

assignment/reducer4.py -mapper 'python3 mapper.py' -reducer 'python3

reducer4.py’

Output: Most popular package in Ireland: ""mingw32"" with 3194919 downloads

Q5.

$HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-

streaming-*.jar -input /user/bdm/assignment/input -output /user/bdm/

assignment/output5 -file /home/bdm/assignment/mapper.py -file /home/bdm/

assignment/reducer5.py -mapper 'python3 mapper.py' -reducer 'python3

reducer5.py’

Output: Most popular OS among R programmers: ""mingw32"" with 3194919 downloads

Screenshots Of Hadoop Run

Question 1

```
untu:~/assignment$ $HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar -input /user/bdm/assignment/input -output /user/bdm/assignment/output -file /home/bdm/
apper.py -file /home/bdm/assignment/reducer.py -mapper 'python3 mapper.py' -reducer 'python3 reducer.py'
4-01 02:40:38,570 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
eJobJar: [/home/bdm/assignment/mapper.py, /home/bdm/assignment/reducer.py, /tmp/hadoop-unjar1329852730200824244/] [] /tmp/streamjob5501944513040243390.jar tmpDir=null
4-01 02:40:39,284 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
4-01 02:40:39,362 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
4-01 02:40:39,538 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/bdm/.staging/job_1680310999416_0015
4-01 02:40:40,089 INFO mapred.FileInputFormat: Total input files to process : 1
4-01 02:40:40,555 INFO mapreduce.JobSubmitter: number of splits:5
4-01 02:40:40,675 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1680310999416_0015
4-01 02:40:40,675 INFO mapreduce.JobSubmitter: Executing with tokens: []
4-01 02:40:40,816 INFO conf.Configuration: resource-types.xml not found
4-01 02:40:40,817 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
4-01 02:40:40,884 INFO impl.YarnClientImpl: Submitted application application_1680310999416_0015
4-01 02:40:40,916 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1680310999416_0015/
4-01 02:40:40,917 INFO mapreduce.Job: Running job: job_1680310999416_0015
4-01 02:40:46,053 INFO mapreduce.Job: Job job_1680310999416_0015 running in uber mode : false
4-01 02:40:46,057 INFO mapreduce.Job:  map 0% reduce 0%
4-01 02:41:07,720 INFO mapreduce.Job:  map 60% reduce 0%
4-01 02:41:13,985 INFO mapreduce.Job:  map 73% reduce 0%
4-01 02:41:19,157 INFO mapreduce.Job:  map 93% reduce 0%
4-01 02:41:20,172 INFO mapreduce.Job:  map 100% reduce 0%
4-01 02:41:29,241 INFO mapreduce.Job:  map 100% reduce 72%
4-01 02:41:35,284 INFO mapreduce.Job:  map 100% reduce 97%
4-01 02:41:36,307 INFO mapreduce.Job:  map 100% reduce 100%
4-01 02:41:37,339 INFO mapreduce.Job: Job job_1680310999416_0015 completed successfully
4-01 02:41:37,515 INFO mapreduce.Job: Counters: 55
  File System Counters
        FILE: Number of bytes read=575119970
        FILE: Number of bytes written=880190824
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=599519249
        HDFS: Number of bytes written=51
        HDFS: Number of read operations=20
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
  Job Counters
        Killed map tasks=1
        Launched map tasks=6
        Launched reduce tasks=1
        Data-local map tasks=6
        Total time spent by all maps in occupied slots (ms)=146278
        Total time spent by all reduces in occupied slots (ms)=26739
        Total time spent by all map tasks (ms)=146278
        Total time spent by all reduce tasks (ms)=26739
        Total vcore-milliseconds taken by all map tasks=146278
        Total vcore-milliseconds taken by all reduce tasks=26739
        Total megabyte-milliseconds taken by all map tasks=149788672
        Total megabyte-milliseconds taken by all reduce tasks=27380736
  Map-Reduce Framework
        Map input records=7453544
        Map output records=22360632
        Map output bytes=258672477
        Map output materialized bytes=303393771
        Input split bytes=570
        Combine input records=0
        Combine output records=0
        Reduce input groups=20035
        Reduce shuffle bytes=303393771
        Reduce input records=22360632
```

```
        FILE: Number of bytes read=575119970
        FILE: Number of bytes written=880190824
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=599519249
        HDFS: Number of bytes written=51
        HDFS: Number of read operations=20
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
  Job Counters
        Killed map tasks=1
        Launched map tasks=6
        Launched reduce tasks=1
        Data-local map tasks=6
        Total time spent by all maps in occupied slots (ms)=146278
        Total time spent by all reduces in occupied slots (ms)=26739
        Total time spent by all map tasks (ms)=146278
        Total time spent by all reduce tasks (ms)=26739
        Total vcore-milliseconds taken by all map tasks=146278
        Total vcore-milliseconds taken by all reduce tasks=26739
        Total megabyte-milliseconds taken by all map tasks=149788672
        Total megabyte-milliseconds taken by all reduce tasks=27380736
  Map-Reduce Framework
        Map input records=7453544
        Map output records=22360632
        Map output bytes=258672477
        Map output materialized bytes=303393771
        Input split bytes=570
        Combine input records=0
        Combine output records=0
        Reduce input groups=20035
        Reduce shuffle bytes=303393771
        Reduce input records=22360632
        Reduce output records=1
        Spilled Records=44744356
        Shuffled Maps =5
        Failed Shuffles=0
        Merged Map outputs=5
        GC time elapsed (ms)=2022
        CPU time spent (ms)=91770
        Physical memory (bytes) snapshot=2306871296
        Virtual memory (bytes) snapshot=14750511104
        Total committed heap usage (bytes)=2157969408
        Peak Map Physical memory (bytes)=482816000
        Peak Map Virtual memory (bytes)=2473078784
        Peak Reduce Physical memory (bytes)=494931968
        Peak Reduce Virtual memory (bytes)=2478227456
  Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
  File Input Format Counters
        Bytes Read=599518679
  File Output Format Counters
        Bytes Written=51
4-01 02:41:37,516 INFO streaming.StreamJob: Output directory: /user/bdm/assignment/output
untu:~/assignment$ hdfs dfs -cat /user/bdm/assignment/output/part-00000
 of downloads for package ggplot2: 22360632
```

Task 1

Question 2

```
untu:~/assignment$ $HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar -input /user/bdm/assignment/input -output /user/bdm/assignment/output2 -file /home/bdm/
mapper.py -file /home/bdm/assignment/reducer2.py -mapper 'python3 mapper.py' -reducer 'python3 reducer2.py'
4-01 02:47:47,592 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
eJobJar: [/home/bdm/assignment/mapper.py, /home/bdm/assignment/reducer2.py, /tmp/hadoop-unjar5414585757813547821/] [] /tmp/streamjob4273708677920036986.jar tmpDir=null
4-01 02:47:48,384 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
4-01 02:47:48,550 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
4-01 02:47:48,787 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/bdm/.staging/job_1680310999416_0016
4-01 02:47:49,966 INFO mapred.FileInputFormat: Total input files to process : 1
4-01 02:47:50,011 INFO mapreduce.JobSubmitter: number of splits:5
4-01 02:47:50,146 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1680310999416_0016
4-01 02:47:50,146 INFO mapreduce.JobSubmitter: Executing with tokens: []
4-01 02:47:50,290 INFO conf.Configuration: resource-types.xml not found
4-01 02:47:50,290 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
4-01 02:47:50,371 INFO impl.YarnClientImpl: Submitted application application_1680310999416_0016
4-01 02:47:50,412 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1680310999416_0016/
4-01 02:47:50,415 INFO mapreduce.Job: Running job: job_1680310999416_0016
4-01 02:47:56,584 INFO mapreduce.Job: Job job_1680310999416_0016 running in uber mode : false
4-01 02:47:56,588 INFO mapreduce.Job:  map 0% reduce 0%
4-01 02:48:18,100 INFO mapreduce.Job:  map 52% reduce 0%
4-01 02:48:19,278 INFO mapreduce.Job:  map 59% reduce 0%
4-01 02:48:24,516 INFO mapreduce.Job:  map 73% reduce 0%
4-01 02:48:29,958 INFO mapreduce.Job:  map 85% reduce 0%
4-01 02:48:34,598 INFO mapreduce.Job:  map 95% reduce 0%
4-01 02:48:35,814 INFO mapreduce.Job:  map 100% reduce 0%
4-01 02:48:42,033 INFO mapreduce.Job:  map 100% reduce 51%
4-01 02:48:48,297 INFO mapreduce.Job:  map 100% reduce 68%
4-01 02:48:54,634 INFO mapreduce.Job:  map 100% reduce 77%
4-01 02:49:00,780 INFO mapreduce.Job:  map 100% reduce 85%
4-01 02:49:06,840 INFO mapreduce.Job:  map 100% reduce 100%
4-01 02:49:06,858 INFO mapreduce.Job: Job job_1680310999416_0016 completed successfully
4-01 02:49:07,172 INFO mapreduce.Job: Counters: 55
  File System Counters
        FILE: Number of bytes read=575119970
        FILE: Number of bytes written=880190854
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=599519249
        HDFS: Number of bytes written=71
        HDFS: Number of read operations=20
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
  Job Counters
        Killed map tasks=1
        Launched map tasks=6
        Launched reduce tasks=1
        Data-local map tasks=6
        Total time spent by all maps in occupied slots (ms)=171835
        Total time spent by all reduces in occupied slots (ms)=44647
        Total time spent by all map tasks (ms)=171835
        Total time spent by all reduce tasks (ms)=44647
        Total vcore-milliseconds taken by all map tasks=171835
        Total vcore-milliseconds taken by all reduce tasks=44647
        Total megabyte-milliseconds taken by all map tasks=175959040
        Total megabyte-milliseconds taken by all reduce tasks=45718528
  Map-Reduce Framework
        Map input records=7453544
        Map output records=22360632
        Map output bytes=258672477
        Map output materialized bytes=303393771
        Input split bytes=570
        Combine input records=0
```

```
        FILE: Number of bytes read=575119970
        FILE: Number of bytes written=880190854
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=599519249
        HDFS: Number of bytes written=71
        HDFS: Number of read operations=20
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
  Job Counters
        Killed map tasks=1
        Launched map tasks=6
        Launched reduce tasks=1
        Data-local map tasks=6
        Total time spent by all maps in occupied slots (ms)=171835
        Total time spent by all reduces in occupied slots (ms)=44647
        Total time spent by all map tasks (ms)=171835
        Total time spent by all reduce tasks (ms)=44647
        Total vcore-milliseconds taken by all map tasks=171835
        Total vcore-milliseconds taken by all reduce tasks=44647
        Total megabyte-milliseconds taken by all map tasks=175959040
        Total megabyte-milliseconds taken by all reduce tasks=45718528
  Map-Reduce Framework
        Map input records=7453544
        Map output records=22360632
        Map output bytes=258672477
        Map output materialized bytes=303393771
        Input split bytes=570
        Combine input records=0
        Combine output records=0
        Reduce input groups=20035
        Reduce shuffle bytes=303393771
        Reduce input records=22360632
        Reduce output records=1
        Spilled Records=44744356
        Shuffled Maps =5
        Failed Shuffles=0
        Merged Map outputs=5
        GC time elapsed (ms)=6071
        CPU time spent (ms)=119550
        Physical memory (bytes) snapshot=2567557120
        Virtual memory (bytes) snapshot=14755532800
        Total committed heap usage (bytes)=2602565632
        Peak Map Physical memory (bytes)=482521088
        Peak Map Virtual memory (bytes)=2472869888
        Peak Reduce Physical memory (bytes)=467927040
        Peak Reduce Virtual memory (bytes)=2480345088
  Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
  File Input Format Counters
        Bytes Read=599518679
  File Output Format Counters
        Bytes Written=71
4-01 02:49:07,173 INFO streaming.StreamJob: Output directory: /user/bdm/assignment/output2
untu:~/assignment$ hdfs dfs -cat /user/bdm/assignment/output2/part-00000
t number of downloads by a country: "NA" with 3225550 downloads
```

Task 2

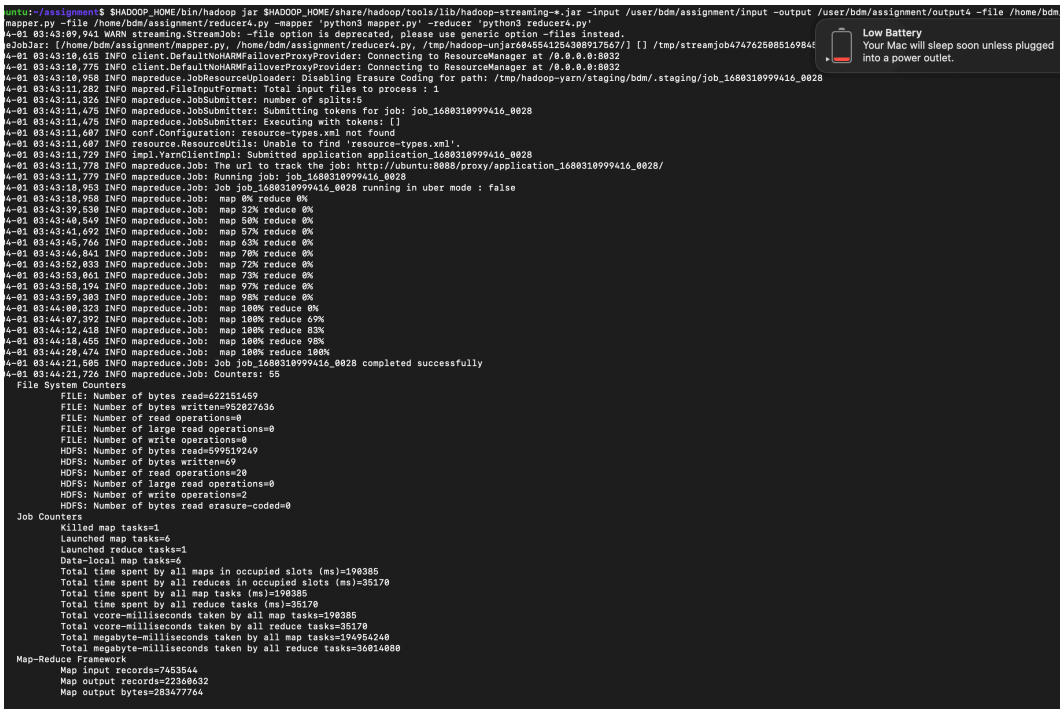
Question 3

```
untu:~/assignment$ $HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar -input /user/bdm/assignment/input -output /user/bdm/assignment/output3 -file /home/bdm/
mapper.py -file /home/bdm/assignment/reducer3.py -mapper 'python3 mapper.py' -reducer 'python3 reducer3.py'
4-01 02:54:12,128 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
eJobJar: [/home/bdm/assignment/mapper.py, /home/bdm/assignment/reducer3.py, /tmp/hadoop-unjar4896457725879384242/] [] /tmp/streamjob5846737037215985124.jar tmpDir=null
4-01 02:54:12,881 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
4-01 02:54:13,119 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
4-01 02:54:13,321 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/bdm/.staging/job_1680310999416_0017
4-01 02:54:14,098 INFO mapred.FileInputFormat: Total input files to process : 1
4-01 02:54:14,178 INFO mapreduce.JobSubmitter: number of splits:5
4-01 02:54:14,342 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1680310999416_0017
4-01 02:54:14,342 INFO mapreduce.JobSubmitter: Executing with tokens: []
4-01 02:54:14,484 INFO conf.Configuration: resource-types.xml not found
4-01 02:54:14,484 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
4-01 02:54:14,615 INFO impl.YarnClientImpl: Submitted application application_1680310999416_0017
4-01 02:54:14,651 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1680310999416_0017/
4-01 02:54:14,653 INFO mapreduce.Job: Running job: job_1680310999416_0017
4-01 02:54:21,879 INFO mapreduce.Job: Job job_1680310999416_0017 running in uber mode : false
4-01 02:54:21,884 INFO mapreduce.Job:  map 0% reduce 0%
4-01 02:54:42,653 INFO mapreduce.Job:  map 32% reduce 0%
4-01 02:54:43,673 INFO mapreduce.Job:  map 51% reduce 0%
4-01 02:54:44,863 INFO mapreduce.Job:  map 57% reduce 0%
4-01 02:54:49,034 INFO mapreduce.Job:  map 69% reduce 0%
4-01 02:54:50,058 INFO mapreduce.Job:  map 72% reduce 0%
4-01 02:54:55,198 INFO mapreduce.Job:  map 80% reduce 0%
4-01 02:54:56,215 INFO mapreduce.Job:  map 84% reduce 0%
4-01 02:54:57,260 INFO mapreduce.Job:  map 87% reduce 0%
4-01 02:54:58,370 INFO mapreduce.Job:  map 100% reduce 0%
4-01 02:55:06,503 INFO mapreduce.Job:  map 100% reduce 72%
4-01 02:55:12,544 INFO mapreduce.Job:  map 100% reduce 85%
4-01 02:55:16,571 INFO mapreduce.Job:  map 100% reduce 100%
4-01 02:55:17,684 INFO mapreduce.Job: Job job_1680310999416_0017 completed successfully
4-01 02:55:17,770 INFO mapreduce.Job: Counters: 55
  File System Counters
        FILE: Number of bytes read=575119970
        FILE: Number of bytes written=880190854
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=599519249
        HDFS: Number of bytes written=349
        HDFS: Number of read operations=20
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
  Job Counters
        Killed map tasks=1
        Launched map tasks=6
        Launched reduce tasks=1
        Data-local map tasks=6
        Total time spent by all maps in occupied slots (ms)=168902
        Total time spent by all reduces in occupied slots (ms)=29706
        Total time spent by all map tasks (ms)=168902
        Total time spent by all reduce tasks (ms)=29706
        Total vcore-milliseconds taken by all map tasks=168902
        Total vcore-milliseconds taken by all reduce tasks=29706
        Total megabyte-milliseconds taken by all map tasks=172956648
        Total megabyte-milliseconds taken by all reduce tasks=30418944
  Map-Reduce Framework
        Map input records=7453544
        Map output records=22360632
        Map output bytes=258672477
        Map output materialized bytes=303393771
        Input split bytes=570
```

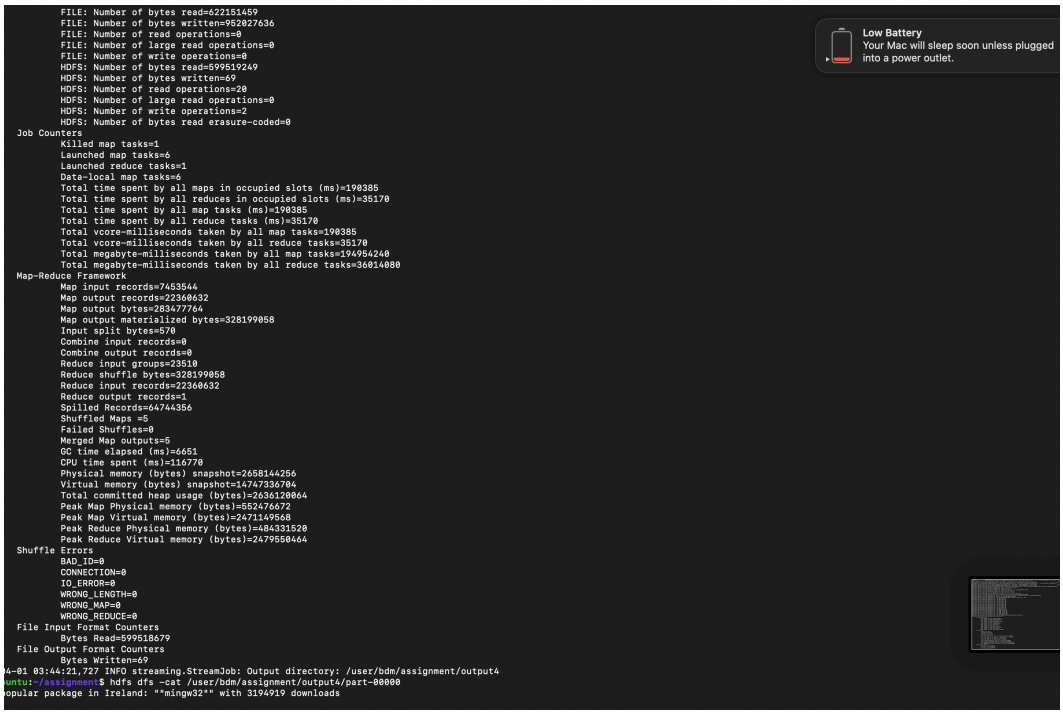
```
        HDFS: Number of bytes read erasure-coded=0
  Job Counters
        Killed map tasks=1
        Launched map tasks=6
        Launched reduce tasks=1
        Data-local map tasks=6
        Total time spent by all maps in occupied slots (ms)=168902
        Total time spent by all reduces in occupied slots (ms)=29706
        Total time spent by all map tasks (ms)=168902
        Total time spent by all reduce tasks (ms)=29706
        Total vcore-milliseconds taken by all map tasks=168902
        Total vcore-milliseconds taken by all reduce tasks=29706
        Total megabyte-milliseconds taken by all map tasks=172956648
        Total megabyte-milliseconds taken by all reduce tasks=30418944
  Map-Reduce Framework
        Map input records=7453544
        Map output records=22360632
        Map output bytes=258672477
        Map output materialized bytes=303393771
        Input split bytes=570
        Combine input records=0
        Combine output records=0
        Reduce input groups=20035
        Reduce shuffle bytes=303393771
        Reduce input records=22360632
        Reduce output records=11
        Spilled Records=64744356
        Shuffled Maps =5
        Failed Shuffles=0
        Merged Map outputs=5
        GC time elapsed (ms)=3212
        CPU time spent (ms)=100300
        Physical memory (bytes) snapshot=2329423872
        Virtual memory (bytes) snapshot=14756061184
        Total committed heap usage (bytes)=2153775104
        Peak Map Physical memory (bytes)=481894400
        Peak Map Virtual memory (bytes)=2471411712
        Peak Reduce Physical memory (bytes)=503808000
        Peak Reduce Virtual memory (bytes)=2481295360
  Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
  File Input Format Counters
        Bytes Read=599518679
  File Output Format Counters
        Bytes Written=349
4-01 02:55:17,771 INFO streaming.StreamJob: Output directory: /user/bdm/assignment/output3
untu:~/assignment$ hdfs dfs -cat /user/bdm/assignment/output3/part-00000
 most popular packages:
*": 3225650 downloads
ingw32"": 3194919 downloads
S"": 3061236 downloads
inux-gnu"": 778523 downloads
arwin17.0"": 648165 downloads
8"": 569535 downloads
arwin20"": 328304 downloads
N"": 282214 downloads
R"": 254392 downloads
DE"": 236903 downloads
```

Task 3

Question 4




```
untu:~/assignment$ $HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar -input /user/bdm/assignment/input -output /user/bdm/assignment/output4 -file /home/bdm
/mapper.py -file /home/bdm/assignment/reducer4.py -mapper 'python3 mapper.py' -reducer 'python3 reducer4.py'
4-01 03:43:09,941 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
eJobJar: [/home/bdm/assignment/mapper.py, /home/bdm/assignment/reducer4.py, /tmp/hadoop-unjar6045541254308917567/] [] /tmp/streamjob4747625085169845
4-01 03:43:10,615 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
4-01 03:43:10,775 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
4-01 03:43:10,958 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/bdm/.staging/job_1680310999416_0028
4-01 03:43:11,282 INFO mapred.FileInputFormat: Total input files to process : 1
4-01 03:43:11,326 INFO mapreduce.JobSubmitter: number of splits:5
4-01 03:43:11,475 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1680310999416_0028
4-01 03:43:11,475 INFO mapreduce.JobSubmitter: Executing with tokens: []
4-01 03:43:11,607 INFO conf.Configuration: resource-types.xml not found
4-01 03:43:11,607 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
4-01 03:43:11,729 INFO impl.YarnClientImpl: Submitted application application_1680310999416_0028
4-01 03:43:11,778 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1680310999416_0028/
4-01 03:43:11,779 INFO mapreduce.Job: Running job: job_1680310999416_0028
4-01 03:43:18,953 INFO mapreduce.Job: Job job_1680310999416_0028 running in uber mode : false
4-01 03:43:18,958 INFO mapreduce.Job:  map 0% reduce 0%
4-01 03:43:39,638 INFO mapreduce.Job:  map 32% reduce 0%
4-01 03:43:40,549 INFO mapreduce.Job:  map 50% reduce 0%
4-01 03:43:41,692 INFO mapreduce.Job:  map 57% reduce 0%
4-01 03:43:45,766 INFO mapreduce.Job:  map 63% reduce 0%
4-01 03:43:46,841 INFO mapreduce.Job:  map 70% reduce 0%
4-01 03:43:52,033 INFO mapreduce.Job:  map 72% reduce 0%
4-01 03:43:53,061 INFO mapreduce.Job:  map 73% reduce 0%
4-01 03:43:58,194 INFO mapreduce.Job:  map 97% reduce 0%
4-01 03:43:59,303 INFO mapreduce.Job:  map 98% reduce 0%
4-01 03:44:00,323 INFO mapreduce.Job:  map 100% reduce 0%
4-01 03:44:07,392 INFO mapreduce.Job:  map 100% reduce 69%
4-01 03:44:12,418 INFO mapreduce.Job:  map 100% reduce 83%
4-01 03:44:18,455 INFO mapreduce.Job:  map 100% reduce 98%
4-01 03:44:20,474 INFO mapreduce.Job:  map 100% reduce 100%
4-01 03:44:21,505 INFO mapreduce.Job: Job job_1680310999416_0028 completed successfully
4-01 03:44:21,726 INFO mapreduce.Job: Counters: 55
  File System Counters
        FILE: Number of bytes read=622151459
        FILE: Number of bytes written=952027636
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=599519249
        HDFS: Number of bytes written=69
        HDFS: Number of read operations=20
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
  Job Counters
        Killed map tasks=1
        Launched map tasks=6
        Launched reduce tasks=1
        Data-local map tasks=6
        Total time spent by all maps in occupied slots (ms)=190385
        Total time spent by all reduces in occupied slots (ms)=35170
        Total time spent by all map tasks (ms)=190385
        Total time spent by all reduce tasks (ms)=35170
        Total vcore-milliseconds taken by all map tasks=190385
        Total vcore-milliseconds taken by all reduce tasks=35170
        Total megabyte-milliseconds taken by all map tasks=194954240
        Total megabyte-milliseconds taken by all reduce tasks=36014080
  Map-Reduce Framework
        Map input records=7453544
        Map output records=22360632
        Map output bytes=283477764
```




```
        FILE: Number of bytes read=622151459
        FILE: Number of bytes written=952027636
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=599519249
        HDFS: Number of bytes written=69
        HDFS: Number of read operations=20
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
  Job Counters
        Killed map tasks=1
        Launched map tasks=6
        Launched reduce tasks=1
        Data-local map tasks=6
        Total time spent by all maps in occupied slots (ms)=190385
        Total time spent by all reduces in occupied slots (ms)=35170
        Total time spent by all map tasks (ms)=190385
        Total time spent by all reduce tasks (ms)=35170
        Total vcore-milliseconds taken by all map tasks=190385
        Total vcore-milliseconds taken by all reduce tasks=35170
        Total megabyte-milliseconds taken by all map tasks=194954240
        Total megabyte-milliseconds taken by all reduce tasks=36014080
  Map-Reduce Framework
        Map input records=7453544
        Map output records=22360632
        Map output bytes=283477764
        Map output materialized bytes=328199058
        Input split bytes=570
        Combine input records=0
        Combine output records=0
        Reduce input groups=23510
        Reduce shuffle bytes=328199058
        Reduce input records=22360632
        Reduce output records=1
        Spilled Records=44744356
        Shuffled Maps =5
        Failed Shuffles=0
        Merged Map outputs=5
        GC time elapsed (ms)=6651
        CPU time spent (ms)=116770
        Physical memory (bytes) snapshot=2658144256
        Virtual memory (bytes) snapshot=14747336704
        Total committed heap usage (bytes)=2636120064
        Peak Map Physical memory (bytes)=552476672
        Peak Map Virtual memory (bytes)=2471149568
        Peak Reduce Physical memory (bytes)=484331520
        Peak Reduce Virtual memory (bytes)=2479550464
  Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
  File Input Format Counters
        Bytes Read=599518679
  File Output Format Counters
        Bytes Written=69
4-01 03:44:21,727 INFO streaming.StreamJob: Output directory: /user/bdm/assignment/output4
untu:~/assignment$ hdfs dfs -cat /user/bdm/assignment/output4/part-00000
opular package in Ireland: ""mingw32"" with 3194919 downloads
```

Task 4

Question 5

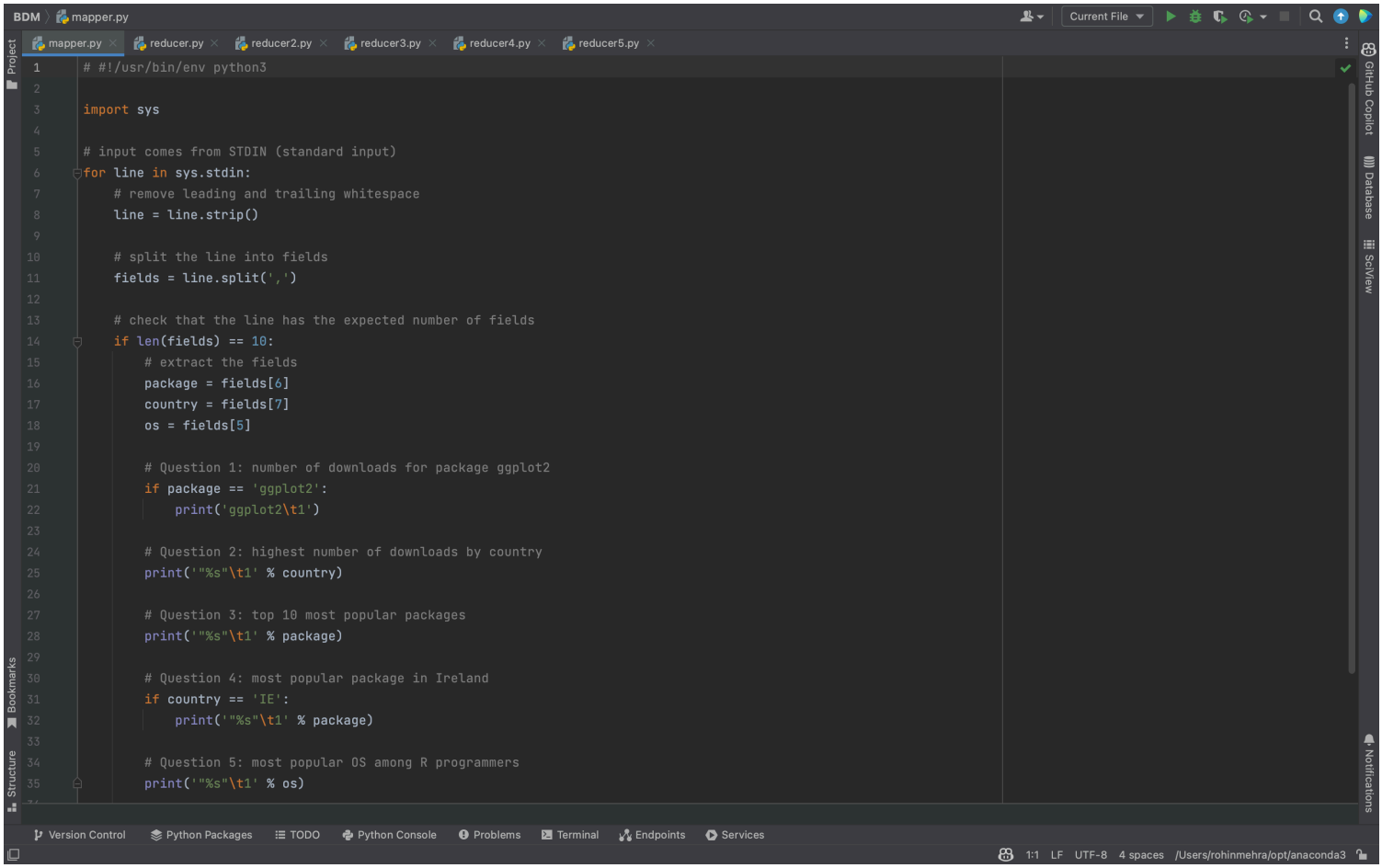
```
untu:~/assignment$ $HADOOP_HOME/bin/hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-*.jar -input /user/bdm/assignment/input -output /user/bdm/assignment/output5 -file /home/bdm/
/mapper.py -file /home/bdm/assignment/reducer5.py -mapper 'python3 mapper.py' -reducer 'python3 reducer5.py'
4-01 03:34:30,353 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
eJobJar: [/home/bdm/assignment/mapper.py, /home/bdm/assignment/reducer5.py, /tmp/hadoop-unjar5579872319461813851/] [] /tmp/streamjob8726752793092545136.jar tmpDir=null
4-01 03:34:31,095 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
4-01 03:34:31,354 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
4-01 03:34:31,572 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/bdm/.staging/job_1680310999416_0026
4-01 03:34:31,995 INFO mapred.FileInputFormat: Total input files to process : 1
4-01 03:34:32,050 INFO mapreduce.JobSubmitter: number of splits:5
4-01 03:34:32,222 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1680310999416_0026
4-01 03:34:32,222 INFO mapreduce.JobSubmitter: Executing with tokens: []
4-01 03:34:32,360 INFO conf.Configuration: resource-types.xml not found
4-01 03:34:32,360 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
4-01 03:34:32,472 INFO impl.YarnClientImpl: Submitted application application_1680310999416_0026
4-01 03:34:32,503 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1680310999416_0026/
4-01 03:34:32,505 INFO mapreduce.Job: Running job: job_1680310999416_0026
4-01 03:34:39,706 INFO mapreduce.Job: Job job_1680310999416_0026 running in uber mode : false
4-01 03:34:39,711 INFO mapreduce.Job:  map 0% reduce 0%
4-01 03:35:00,267 INFO mapreduce.Job:  map 20% reduce 0%
4-01 03:35:01,282 INFO mapreduce.Job:  map 50% reduce 0%
4-01 03:35:03,436 INFO mapreduce.Job:  map 57% reduce 0%
4-01 03:35:06,523 INFO mapreduce.Job:  map 68% reduce 0%
4-01 03:35:07,551 INFO mapreduce.Job:  map 71% reduce 0%
4-01 03:35:12,656 INFO mapreduce.Job:  map 84% reduce 0%
4-01 03:35:13,678 INFO mapreduce.Job:  map 88% reduce 0%
4-01 03:35:14,743 INFO mapreduce.Job:  map 100% reduce 0%
4-01 03:35:24,824 INFO mapreduce.Job:  map 100% reduce 74%
4-01 03:35:29,863 INFO mapreduce.Job:  map 100% reduce 87%
4-01 03:35:34,911 INFO mapreduce.Job:  map 100% reduce 100%
4-01 03:35:34,934 INFO mapreduce.Job: Job job_1680310999416_0026 completed successfully
4-01 03:35:35,129 INFO mapreduce.Job: Counters: 54
  File System Counters
        FILE: Number of bytes read=622151459
        FILE: Number of bytes written=952027630
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=599519249
        HDFS: Number of bytes written=73
        HDFS: Number of read operations=20
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
  Job Counters
        Launched map tasks=5
        Launched reduce tasks=1
        Data-local map tasks=5
        Total time spent by all maps in occupied slots (ms)=152851
        Total time spent by all reduces in occupied slots (ms)=28556
        Total time spent by all map tasks (ms)=152851
        Total time spent by all reduce tasks (ms)=28556
        Total vcore-milliseconds taken by all map tasks=152851
        Total vcore-milliseconds taken by all reduce tasks=28556
        Total megabyte-milliseconds taken by all map tasks=156519424
        Total megabyte-milliseconds taken by all reduce tasks=29241344
  Map-Reduce Framework
        Map input records=7453544
        Map output records=22360632
        Map output bytes=283477764
        Map output materialized bytes=328199058
        Input split bytes=570
        Combine input records=0
        Combine output records=0
```

```
  File System Counters
        FILE: Number of bytes read=622151459
        FILE: Number of bytes written=952027630
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=599519249
        HDFS: Number of bytes written=73
        HDFS: Number of read operations=20
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
  Job Counters
        Launched map tasks=5
        Launched reduce tasks=1
        Data-local map tasks=5
        Total time spent by all maps in occupied slots (ms)=152851
        Total time spent by all reduces in occupied slots (ms)=28556
        Total time spent by all map tasks (ms)=152851
        Total time spent by all reduce tasks (ms)=28556
        Total vcore-milliseconds taken by all map tasks=152851
        Total vcore-milliseconds taken by all reduce tasks=28556
        Total megabyte-milliseconds taken by all map tasks=156519424
        Total megabyte-milliseconds taken by all reduce tasks=29241344
  Map-Reduce Framework
        Map input records=7453544
        Map output records=22360632
        Map output bytes=283477764
        Map output materialized bytes=328199058
        Input split bytes=570
        Combine input records=0
        Combine output records=0
        Reduce input groups=23510
        Reduce shuffle bytes=328199058
        Reduce input records=22360632
        Reduce output records=1
        Spilled Records=64744356
        Shuffled Maps =5
        Failed Shuffles=0
        Merged Map outputs=5
        GC time elapsed (ms)=4954
        CPU time spent (ms)=109220
        Physical memory (bytes) snapshot=2621042688
        Virtual memory (bytes) snapshot=14748766208
        Total committed heap usage (bytes)=2527068160
        Peak Map Physical memory (bytes)=480972800
        Peak Map Virtual memory (bytes)=2470150144
        Peak Reduce Physical memory (bytes)=666948736
        Peak Reduce Virtual memory (bytes)=2478886912
  Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
  File Input Format Counters
        Bytes Read=599518679
  File Output Format Counters
        Bytes Written=73
4-01 03:35:35,129 INFO streaming.StreamJob: Output directory: /user/bdm/assignment/output5
untu:~/assignment$ hdfs dfs -cat /user/bdm/assignment/output5/part-00000
opular OS among R programmers: ""mingw32"" with 3194919 downloads
```

Task 5

Mapper & Reducer Screenshots

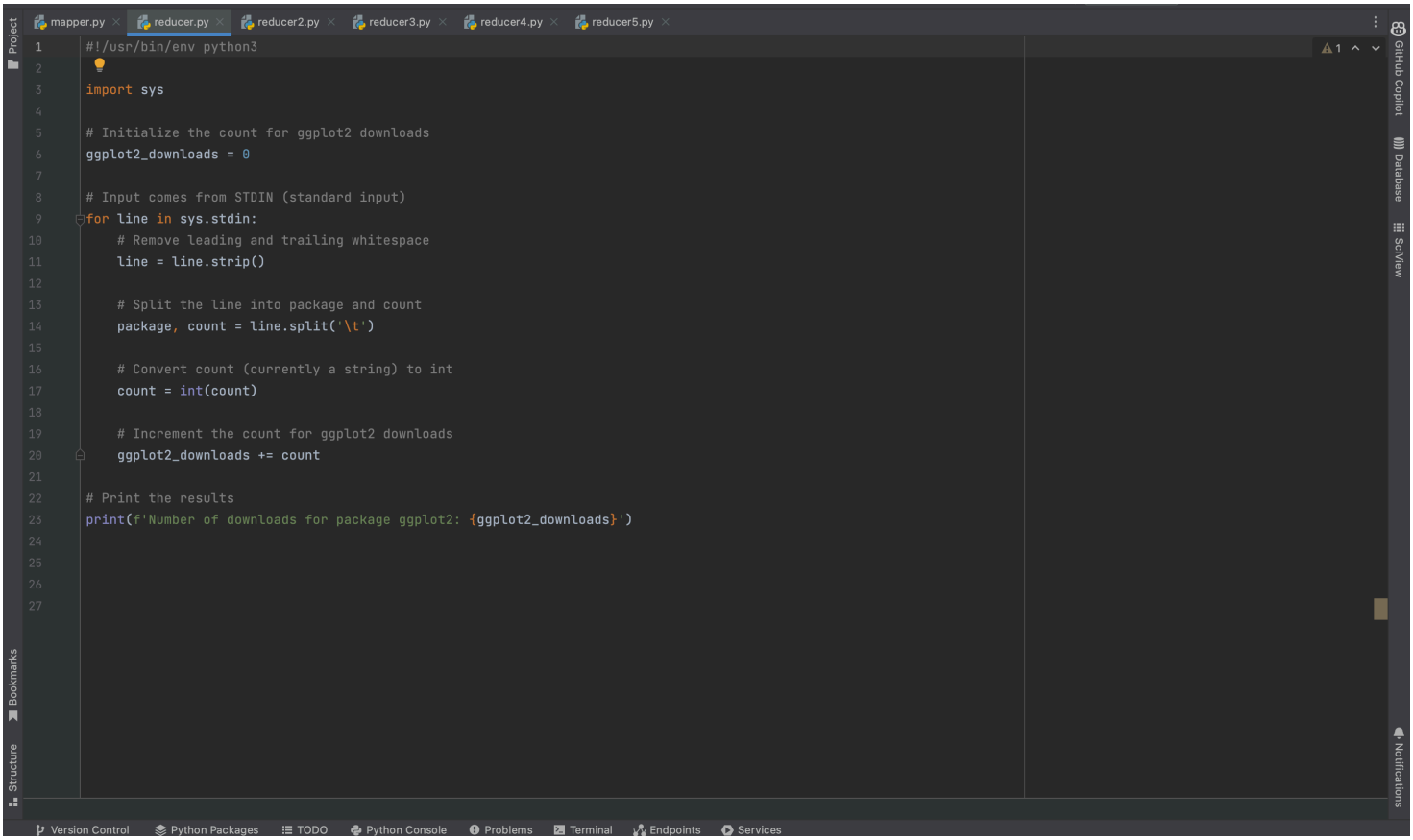
Mapper is common in structure for all tasks it's just that reducer changes for every operation desired.




```python
# #!/usr/bin/env python3

import sys

# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()

    # split the line into fields
    fields = line.split(',')

    # check that the line has the expected number of fields
    if len(fields) == 10:
        # extract the fields
        package = fields[6]
        country = fields[7]
        os = fields[5]

        # Question 1: number of downloads for package ggplot2
        if package == 'ggplot2':
            print('ggplot2\t1')

        # Question 2: highest number of downloads by country
        print('"%s"\t1' % country)

        # Question 3: top 10 most popular packages
        print('"%s"\t1' % package)

        # Question 4: most popular package in Ireland
        if country == 'IE':
            print('"%s"\t1' % package)

        # Question 5: most popular OS among R programmers
        print('"%s"\t1' % os)
```

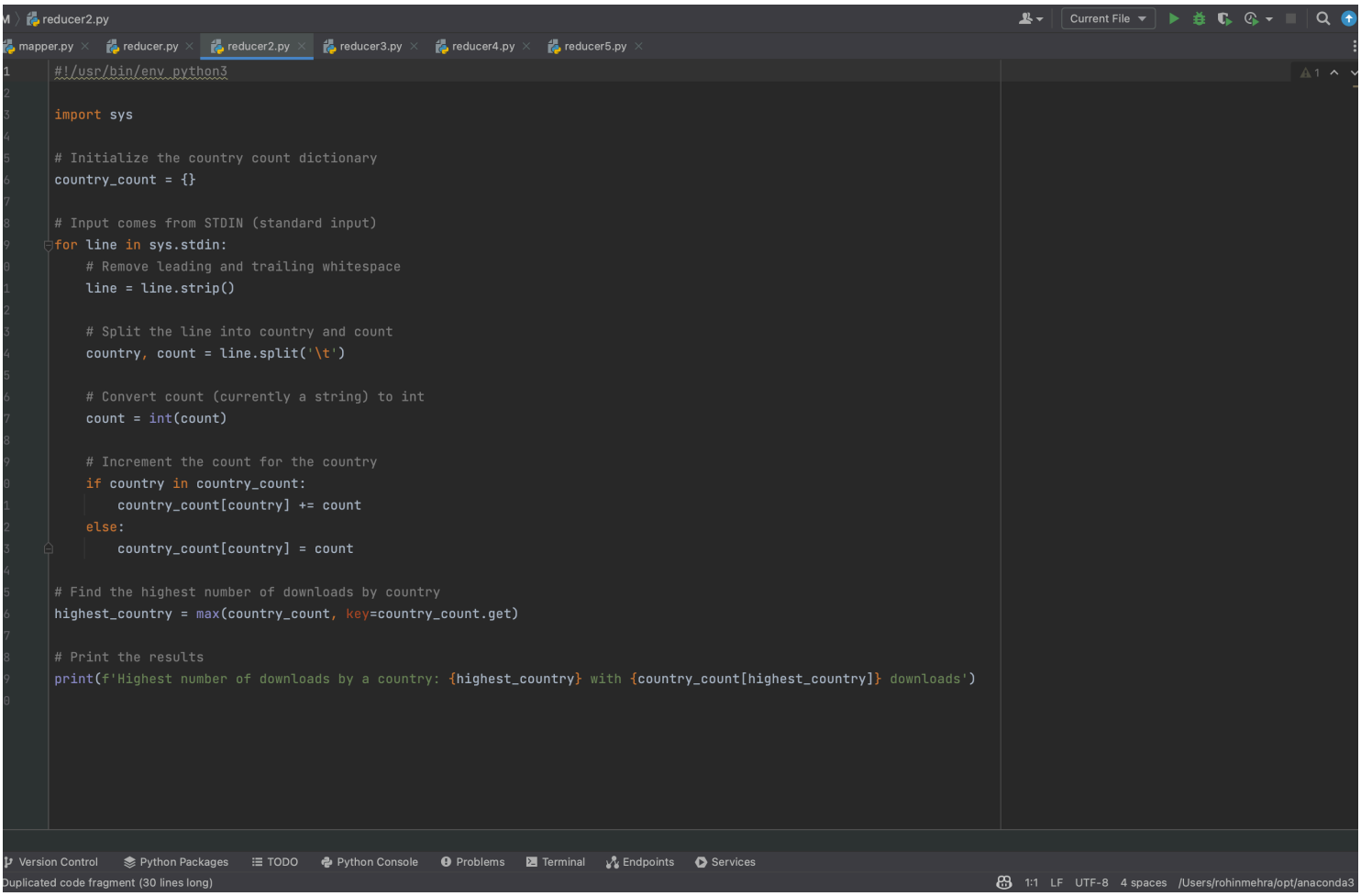
Mapper Code

For Question 1 reducer is as follows:



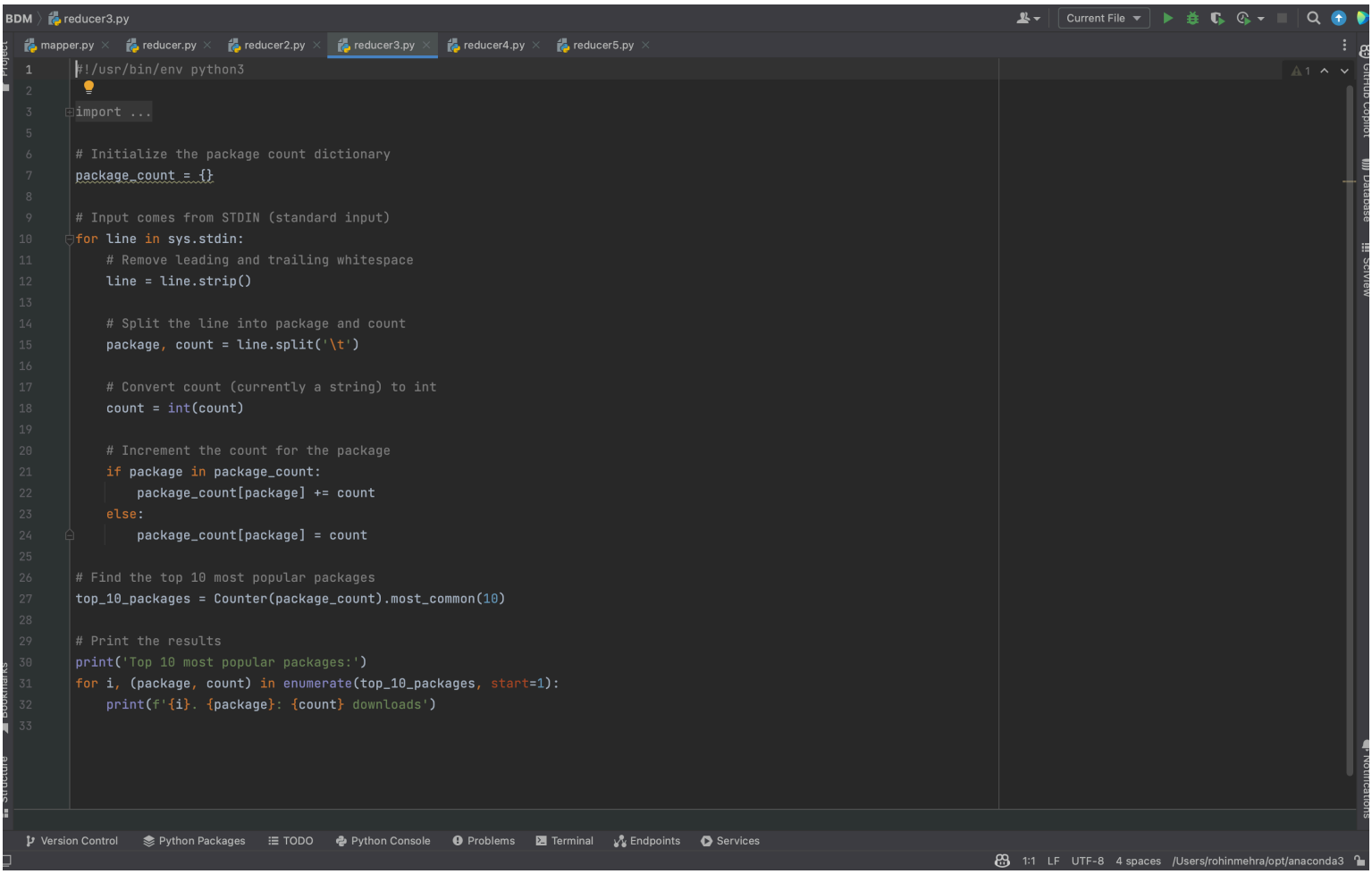
Reducer for Question 1

For Question 2 reducer is as follows:



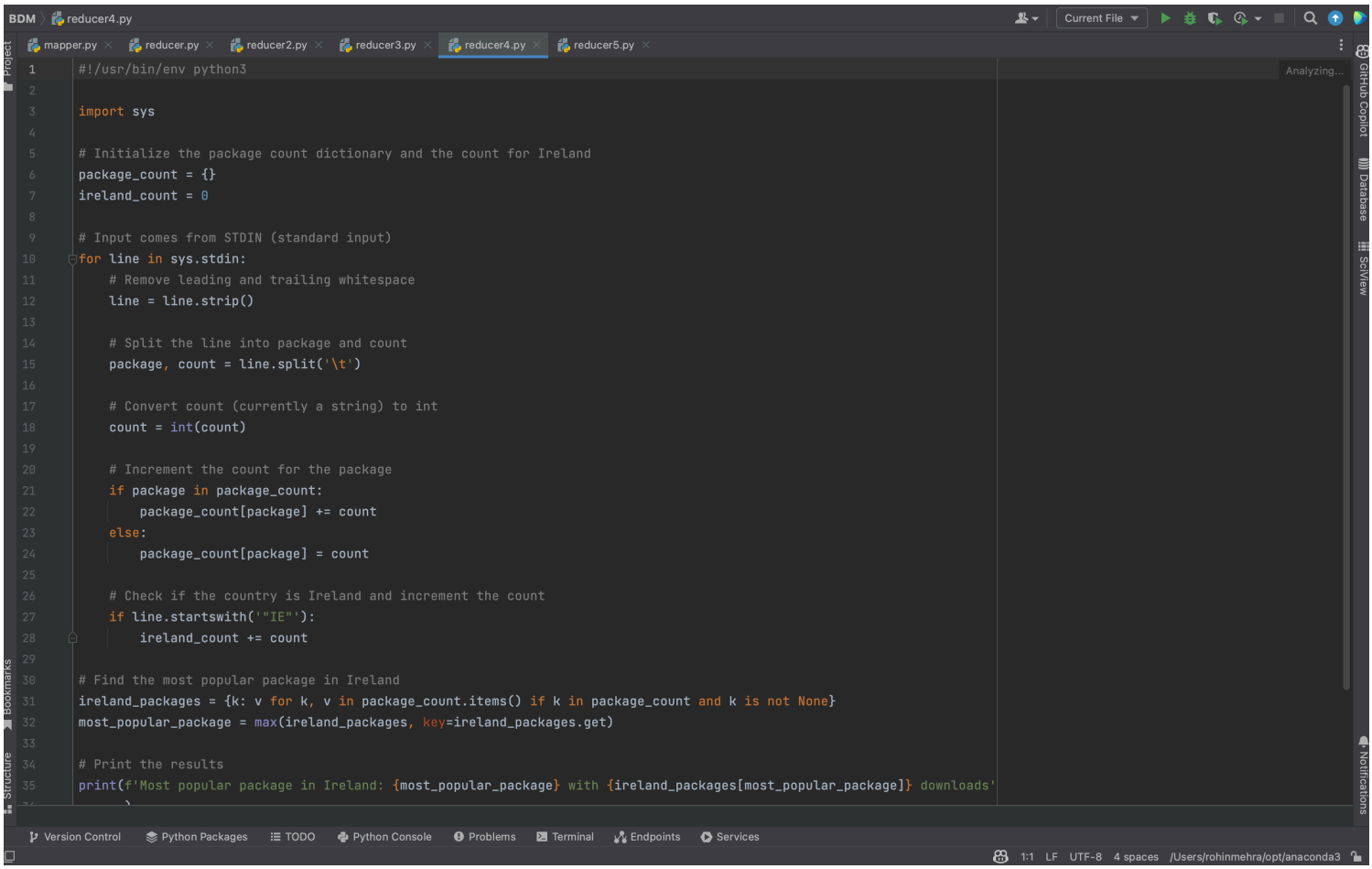
Reducer for Question 2

For Question 3 reducer is as follows:



Reducer for Question 3

For Question 4 reducer is as follows:



Reducer for Question 4

For Question 5 reducer is as follows:

reducer5.py

Current File

mapper.py | reducer.py | reducer2.py | reducer3.py | reducer4.py | reducer5.py

```
#!/usr/bin/env python3
import sys

# Initialize the OS count dictionary
os_count = {}

# Input comes from STDIN (standard input)
for line in sys.stdin:
    # Remove leading and trailing whitespace
    line = line.strip()

    # Split the line into OS and count
    os, count = line.split('\t')

    # Convert count (currently a string) to int
    count = int(count)

    # Increment the count for the OS
    if os in os_count:
        os_count[os] += count
    else:
        os_count[os] = count

# Find the most popular OS among R programmers
most_popular_os = max(os_count, key=os_count.get)

# Print the results
print(f'Most popular OS among R programmers: {most_popular_os} with {os_count[most_popular_os]} downloads')
```

Version Control | Python Packages | TODO | Python Console | Problems | Terminal | Endpoints | Services

Duplicated code fragment (30 lines long) 1:1 LF UTF-8 4 spaces /Users/rohinmehra/opt/anacon

Reducer for Question 5