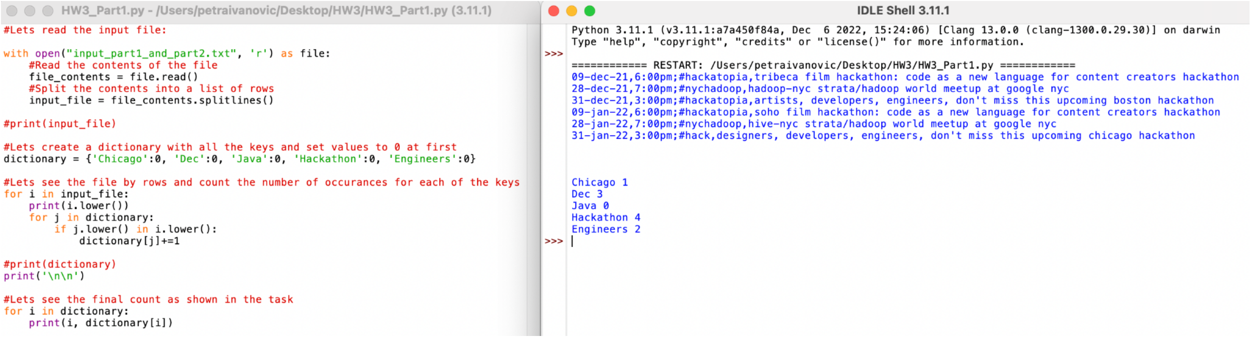
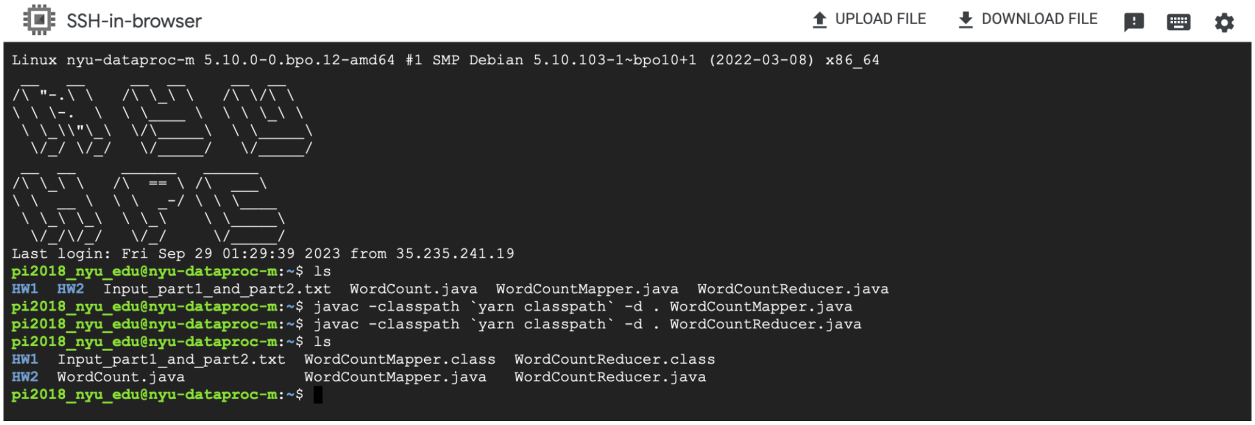
Part 1:


HW3_Part1.py - /Users/petraivanovic/Desktop/HW3/HW3_Part1.py (3.11.1)
IDLE Shell 3.11.1

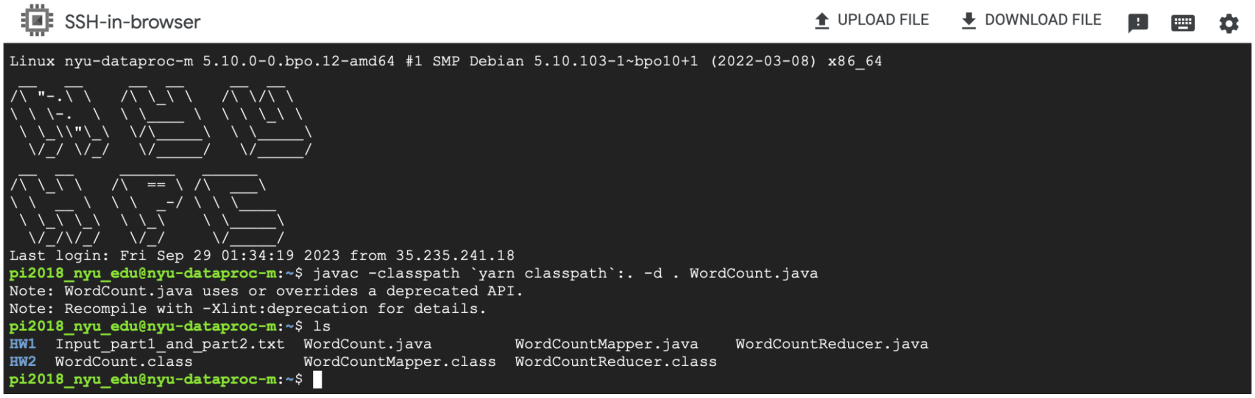
```python
#Lets read the input file:

with open("input_part1_and_part2.txt", 'r') as file:
    #Read the contents of the file
    file_contents = file.read()
    #Split the contents into a list of rows
    input_file = file_contents.splitlines()

#print(input_file)

#Lets create a dictionary with all the keys and set values to 0 at first
dictionary = {'Chicago':0, 'Dec':0, 'Java':0, 'Hackathon':0, 'Engineers':0}

#Lets see the file by rows and count the number of occurances for each of the keys
for i in input_file:
    print(i.lower())
    for j in dictionary:
        if j.lower() in i.lower():
            dictionary[j]+=1

#print(dictionary)
print('\n\n')

#Lets see the final count as shown in the task
for i in dictionary:
    print(i, dictionary[i])
```

```
Python 3.11.1 (v3.11.1:a7a450f84a, Dec  6 2022, 15:24:06) [Clang 13.0.0 (clang-1300.0.29.30)] on darwin
Type "help", "copyright", "credits" or "license()" for more information.
>>> 
============ RESTART: /Users/petraivanovic/Desktop/HW3/HW3_Part1.py ============
09-dec-21,6:00pm;#hackatopia,tribeca film hackathon: code as a new language for content creators hackathon
28-dec-21,7:00pm;#nychadoop,hadoop-nyc strata/hadoop world meetup at google nyc
31-dec-21,3:00pm;#hackatopia,artists, developers, engineers, don't miss this upcoming boston hackathon
09-jan-22,6:00pm;#hackatopia,soho film hackathon: code as a new language for content creators hackathon
28-jan-22,7:00pm;#nychadoop,hive-nyc strata/hadoop world meetup at google nyc
31-jan-22,3:00pm;#hack,designers, developers, engineers, don't miss this upcoming chicago hackathon



Chicago 1
Dec 3
Java 0
Hackathon 4
Engineers 2
>>> 
```

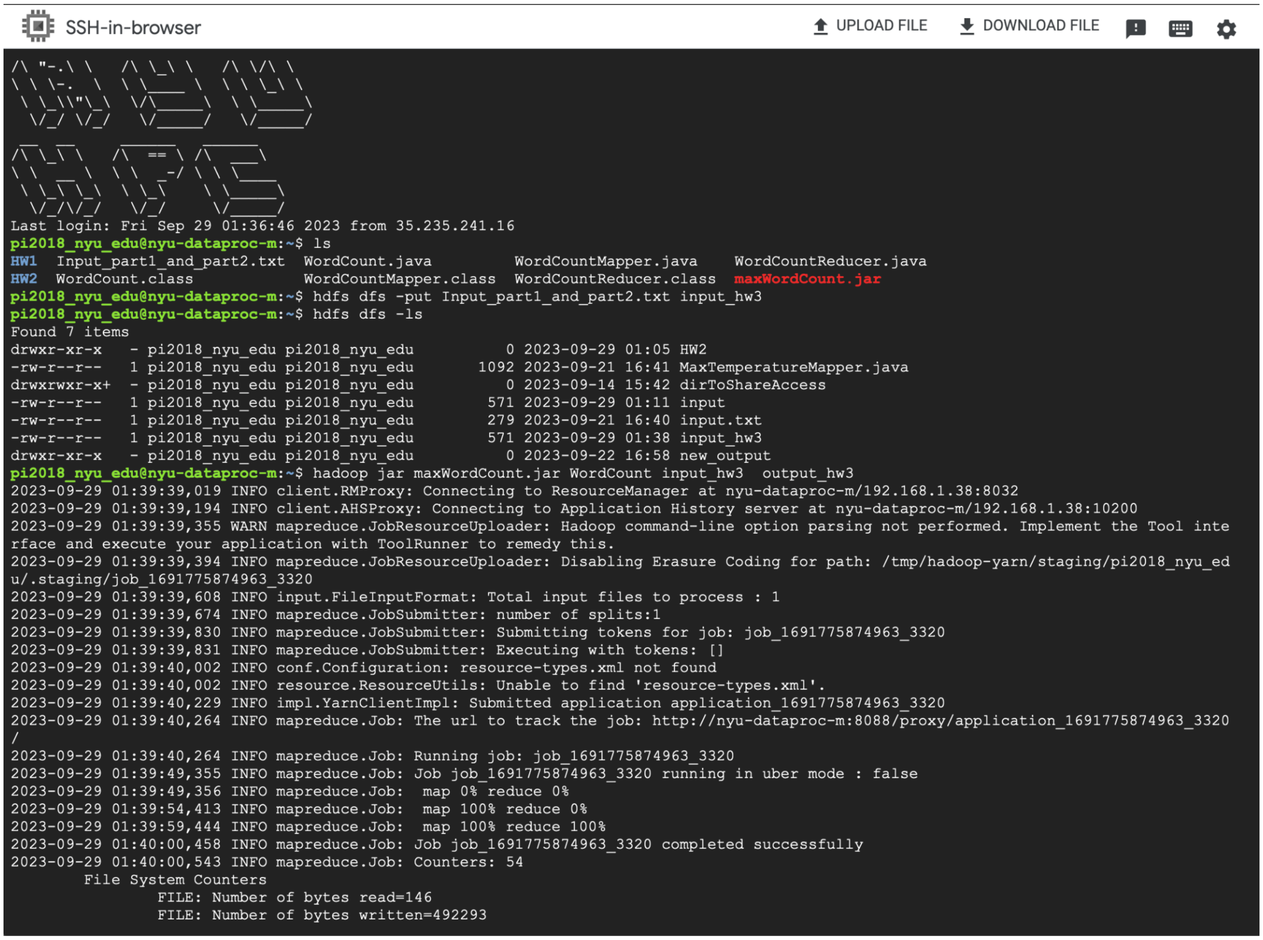
Part 2:



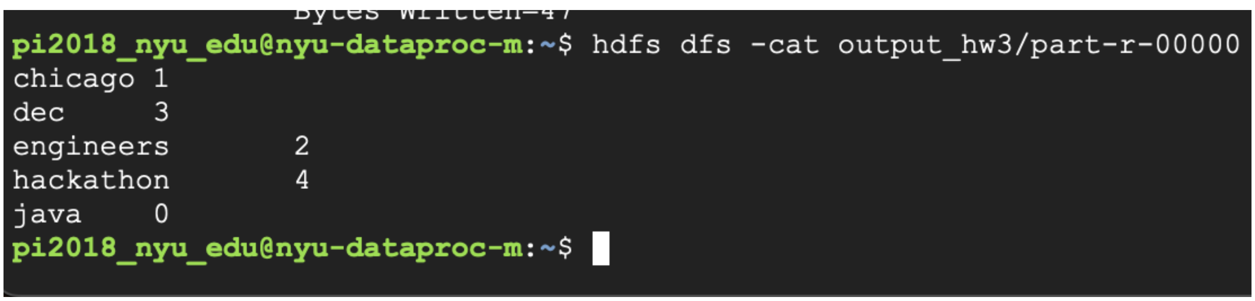
```
Linux nyu-dataproc-m 5.10.0-0.bpo.12-amd64 #1 SMP Debian 5.10.103-1~bpo10+1 (2022-03-08) x86_64
 __   __     __  __     __  __
/\ "-.\ \   /\ \_\ \   /\ \/\ \
\ \ \-.  \  \ \____ \  \ \ \_\ \
 \ \_\\"\_\  \/\_____\  \ \_____\
  \/_/ \/_/   \/_____/   \/_____/
 __  __     ______   ______
/\ \_\ \   /\  == \ /\  ___\
\ \  __ \  \ \  _-/ \ \ \____
 \ \_\ \_\  \ \_\    \ \_____\
  \/_/\/_/   \/_/     \/_____/
Last login: Fri Sep 29 01:29:39 2023 from 35.235.241.19
pi2018_nyu_edu@nyu-dataproc-m:~$ ls
HW1  HW2  Input_part1_and_part2.txt  WordCount.java  WordCountMapper.java  WordCountReducer.java
pi2018_nyu_edu@nyu-dataproc-m:~$ javac -classpath `yarn classpath` -d . WordCountMapper.java
pi2018_nyu_edu@nyu-dataproc-m:~$ javac -classpath `yarn classpath` -d . WordCountReducer.java
pi2018_nyu_edu@nyu-dataproc-m:~$ ls
HW1  Input_part1_and_part2.txt  WordCountMapper.class  WordCountReducer.class
HW2  WordCount.java             WordCountMapper.java   WordCountReducer.java
pi2018_nyu_edu@nyu-dataproc-m:~$ 
```



```
Linux nyu-dataproc-m 5.10.0-0.bpo.12-amd64 #1 SMP Debian 5.10.103-1~bpo10+1 (2022-03-08) x86_64
 __   __     __  __     __  __
/\ "-.\ \   /\ \_\ \   /\ \/\ \
\ \ \-.  \  \ \____ \  \ \ \_\ \
 \ \_\\"\_\  \/\_____\  \ \_____\
  \/_/ \/_/   \/_____/   \/_____/
 __  __     ______   ______
/\ \_\ \   /\  == \ /\  ___\
\ \  __ \  \ \  _-/ \ \ \____
 \ \_\ \_\  \ \_\    \ \_____\
  \/_/\/_/   \/_/     \/_____/
Last login: Fri Sep 29 01:34:19 2023 from 35.235.241.18
pi2018_nyu_edu@nyu-dataproc-m:~$ javac -classpath `yarn classpath`:. -d . WordCount.java
Note: WordCount.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
pi2018_nyu_edu@nyu-dataproc-m:~$ ls
HW1  Input_part1_and_part2.txt  WordCount.java         WordCountMapper.java    WordCountReducer.java
HW2  WordCount.class            WordCountMapper.class  WordCountReducer.class
pi2018_nyu_edu@nyu-dataproc-m:~$ 
```



```
/\ "-.\ \   /\ \_\ \   /\ \/\ \
\ \ \-.  \  \ \____ \  \ \ \_\ \
 \ \_\\"\_\  \/\_____\  \ \_____\
  \/_/ \/_/   \/_____/   \/_____/

 __  __     ______   ______
/\ \_\ \   /\  == \ /\  ___\
\ \  __ \  \ \  _-/ \ \ \____
 \ \_\ \_\  \ \_\    \ \_____\
  \/_/\/_/   \/_/     \/_____/
Last login: Fri Sep 29 01:36:46 2023 from 35.235.241.16
pi2018_nyu_edu@nyu-dataproc-m:~$ ls
HW1  Input_part1_and_part2.txt  WordCount.java          WordCountMapper.java    WordCountReducer.java
HW2  WordCount.class            WordCountMapper.class  WordCountReducer.class  maxWordCount.jar
pi2018_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -put Input_part1_and_part2.txt input_hw3
pi2018_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -ls
Found 7 items
drwxr-xr-x   - pi2018_nyu_edu pi2018_nyu_edu          0 2023-09-29 01:05 HW2
-rw-r--r--   1 pi2018_nyu_edu pi2018_nyu_edu       1092 2023-09-21 16:41 MaxTemperatureMapper.java
drwxrwxr-x+  - pi2018_nyu_edu pi2018_nyu_edu          0 2023-09-14 15:42 dirToShareAccess
-rw-r--r--   1 pi2018_nyu_edu pi2018_nyu_edu        571 2023-09-29 01:11 input
-rw-r--r--   1 pi2018_nyu_edu pi2018_nyu_edu        279 2023-09-21 16:40 input.txt
-rw-r--r--   1 pi2018_nyu_edu pi2018_nyu_edu        571 2023-09-29 01:38 input_hw3
drwxr-xr-x   - pi2018_nyu_edu pi2018_nyu_edu          0 2023-09-22 16:58 new_output
pi2018_nyu_edu@nyu-dataproc-m:~$ hadoop jar maxWordCount.jar WordCount input_hw3  output_hw3
2023-09-29 01:39:39,019 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.38:8032
2023-09-29 01:39:39,194 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.38:10200
2023-09-29 01:39:39,355 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool inte
rface and execute your application with ToolRunner to remedy this.
2023-09-29 01:39:39,394 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/pi2018_nyu_ed
u/.staging/job_1691775874963_3320
2023-09-29 01:39:39,608 INFO input.FileInputFormat: Total input files to process : 1
2023-09-29 01:39:39,674 INFO mapreduce.JobSubmitter: number of splits:1
2023-09-29 01:39:39,830 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1691775874963_3320
2023-09-29 01:39:39,831 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-09-29 01:39:40,002 INFO conf.Configuration: resource-types.xml not found
2023-09-29 01:39:40,002 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-09-29 01:39:40,229 INFO impl.YarnClientImpl: Submitted application application_1691775874963_3320
2023-09-29 01:39:40,264 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1691775874963_3320
/
2023-09-29 01:39:40,264 INFO mapreduce.Job: Running job: job_1691775874963_3320
2023-09-29 01:39:49,355 INFO mapreduce.Job: Job job_1691775874963_3320 running in uber mode : false
2023-09-29 01:39:49,356 INFO mapreduce.Job:  map 0% reduce 0%
2023-09-29 01:39:54,413 INFO mapreduce.Job:  map 100% reduce 0%
2023-09-29 01:39:59,444 INFO mapreduce.Job:  map 100% reduce 100%
2023-09-29 01:40:00,458 INFO mapreduce.Job: Job job_1691775874963_3320 completed successfully
2023-09-29 01:40:00,543 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=146
                FILE: Number of bytes written=492293
```



```
                        Bytes Written=47
pi2018_nyu_edu@nyu-dataproc-m:~$ hdfs dfs -cat output_hw3/part-r-00000
chicago 1
dec     3
engineers       2
hackathon       4
java    0
pi2018_nyu_edu@nyu-dataproc-m:~$ 
```

Part 3:

This article talks about the challenge Twitter faced that led them to optimize thair Hadoop clusters because they were reaching performance limits because of the storage I/O bottlenecks. Since Twitter is a huge platform with up to 10,000 nodes and 100PB of logical storage, simply getting more HDDs would not solve the problem - even with that scalability would be an issue. Because of this Twitter partnered with Inter engineers to solve this problem. After conducting numerous experiments, they realized that “selectively placing the temporary data contained in the YARN Temp directory” on fast SSDs could significantly improve performance with up to 50% reduction in runtime. This results were revolutionary and they led to not only better performance but also amazing cost savings which enabled Twitter to reduce the number of HDDs by 75% without negatively impacting their benchmarks. This article and the Twitter Hadoop case showed how their innovative thinking and usage of fast SSDs can incredibly increase computing power and create more efficient implementations of Hadoop clusters to deliver better performance and lower cost.