CS580K: Mini Project 2
B00814281:Prathamesh Walke

Task 1

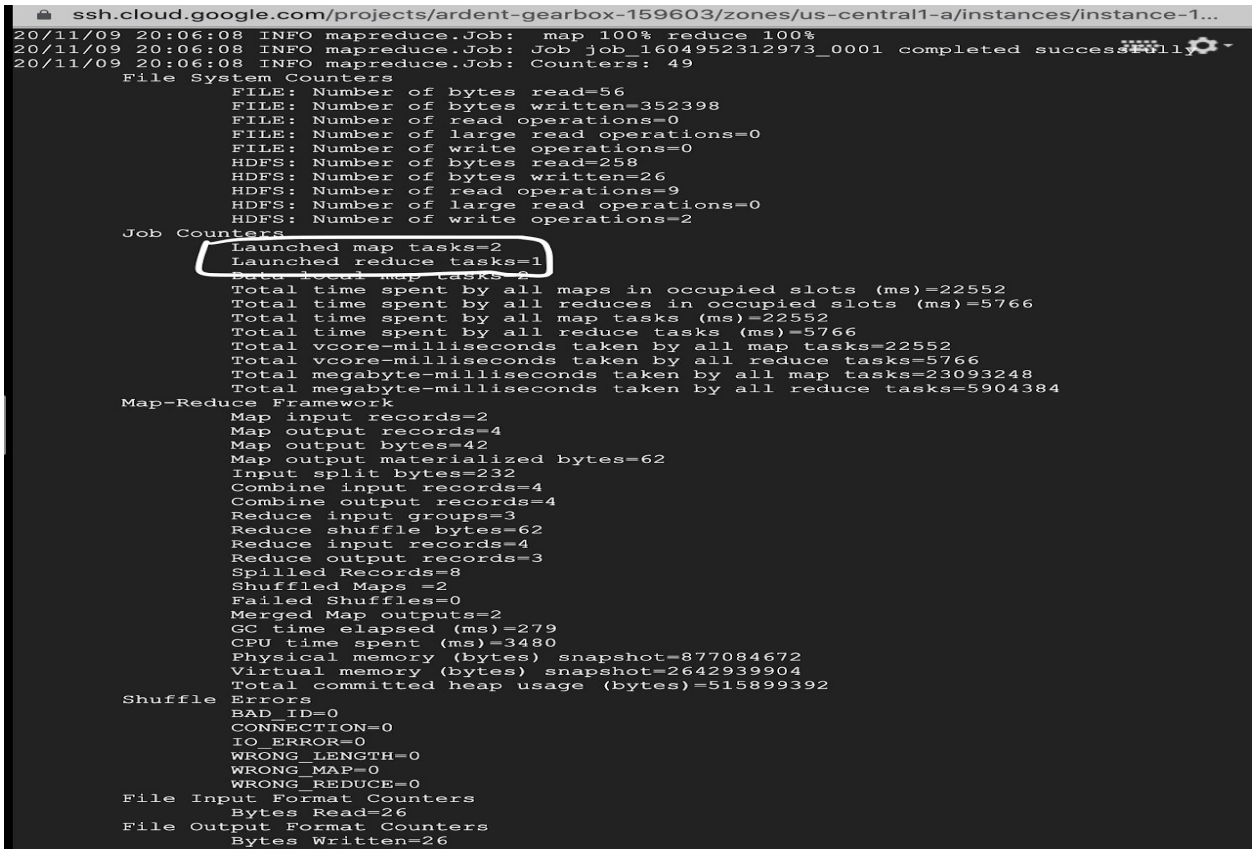
Q.1

1. We first create a network in docker for the containers to interact with each other.
2. We then start the hadoop clusters, which gives us one master and two slave containers as is mentioned in the “start-container.sh”
3. Launch the hdfs and YARN from the Hadoop master container root directory.
4. Create input files with text in it, that is some strings in it.
5. Move these to HDFS 6. Finally execute the wordcount program.

Q.2

The command transfers the contents of the input folder from localhost to the HDFS, the contents are broken into chunks and stored in the disk.

Q.3



2 mappers and 1 reducer are launched for executing the above wordcount program

Q.4

Time spend by mappers separately is 22552 ms
Time spent by reducers separately is 5766 ms

Q.5

The output folder has two files namely _SUCCESS and “part-r-00000”
The content in part-r-00000 is the output given by mapreduce program and has following content :
Docker 1
Hadoop 1
Hello   2

```
root@hadoop-master:~# hdfs dfs -cat output/part-r-00000
Docker  1
Hadoop  1
Hello   2
root@hadoop-master:~# hdfs dfs -ls output/
Found 2 items
-rw-r--r--   2 root supergroup          0 2020-11-09 20:06 output/_SUCCESS
-rw-r--r--   2 root supergroup         26 2020-11-09 20:06 output/part-r-00000
root@hadoop-master:~#
```

Task 2

Q.6

```
root@CS580-pwalke1:/home/pwalke1/CS580k/project2/hadoop-cluster-docker# ./resize-cluster.sh 5


build docker hadoop image

Sending build context to Docker daemon  328.7kB
Step 1/15 : FROM ubuntu:14.04
 ---> df043b4f0cf1
Step 2/15 : MAINTAINER KiwenLau <kiwenlau@gmail.com>
 ---> Using cache
 ---> 45bc2a967d11
Step 3/15 : WORKDIR /root
 ---> Using cache
 ---> 0b3b1e5c0aec
Step 4/15 : RUN apt-get update && apt-get install -y openssh-server openjdk-7-jdk wget
 ---> Using cache
 ---> e011f25d9296
Step 5/15 : RUN wget https://github.com/kiwenlau/compile-hadoop/releases/download/2.7.2/hadoop-2.7.2.tar.gz &&     tar -xzvf hadoop-2.7.2.tar.gz &&     mv hadoop-2.7.2 /usr/local/hadoop &&     rm hadoop-2
.7.2.tar.gz
 ---> Using cache
 ---> ed48e79d6e84
Step 6/15 : ENV JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
 ---> Using cache
 ---> b4be4a5debc2
Step 7/15 : ENV HADOOP_HOME=/usr/local/hadoop
 ---> Using cache
 ---> 1191bd234102
Step 8/15 : ENV PATH=$PATH:/usr/local/hadoop/bin:/usr/local/hadoop/sbin
 ---> Using cache
 ---> 24a3c2c56efe
Step 9/15 : RUN ssh-keygen -t rsa -f ~/.ssh/id_rsa -P '' &&     cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
 ---> Using cache
 ---> 1aa29508c531
Step 10/15 : RUN mkdir -p ~/hdfs/namenode &&     mkdir -p ~/hdfs/datanode &&     mkdir $HADOOP_HOME/logs
 ---> Using cache
 ---> 35023fee5cdf
Step 11/15 : COPY config/* /tmp/
 ---> Using cache
 ---> 810e515b24fe
Step 12/15 : RUN mv /tmp/ssh_config ~/.ssh/config &&     mv /tmp/hadoop-env.sh /usr/local/hadoop/etc/hadoop/hadoop-env.sh &&     mv /tmp/hdfs-site.xml $HADOOP_HOME/etc/hadoop/hdfs-site.xml &&     mv /tmp/
core-site.xml $HADOOP_HOME/etc/hadoop/core-site.xml &&     mv /tmp/mapred-site.xml $HADOOP_HOME/etc/hadoop/mapred-site.xml &&     mv /tmp/yarn-site.xml $HADOOP_HOME/etc/hadoop/yarn-site.xml &&     mv /tmp
/slaves $HADOOP_HOME/etc/hadoop/slaves &&     mv /tmp/start-hadoop.sh ~/start-hadoop.sh &&     mv /tmp/run-wordcount.sh ~/run-wordcount.sh
 ---> Using cache
 ---> 31a8b4de0cf7
Step 13/15 : RUN chmod +x ~/start-hadoop.sh &&     chmod +x ~/run-wordcount.sh &&     chmod +x $HADOOP_HOME/sbin/start-dfs.sh &&     chmod +x $HADOOP_HOME/sbin/start-yarn.sh
 ---> Using cache
 ---> f36a0e980057
Step 14/15 : RUN /usr/local/hadoop/bin/hdfs namenode -format
 ---> Using cache
 ---> be2e2368a053
Step 15/15 : CMD [ "sh", "-c", "service ssh start; bash"]
 ---> Using cache
 ---> dab61f511bab
Successfully built dab61f511bab
Successfully tagged kiwenlau/hadoop:1.0

root@CS580-pwalke1:/home/pwalke1/CS580k/project2/hadoop-cluster-docker# ./start-container.sh
start hadoop-master container...
start hadoop-slave1 container...
start hadoop-slave2 container...
start hadoop-slave3 container...
start hadoop-slave4 container...
root@hadoop-master:~#
```

1 master and 4 slaves are launched separately

Q.7

Master Node – Master node in a hadoop cluster is responsible for storing data in HDFS and executing parallel computation of the stored data using MapReduce. Master Node has 3 nodes – NameNode, Secondary NameNode and JobTracker. JobTracker monitors the parallel processing of data using MapReduce while the NameNode handles the data storage function with HDFS. NameNode keeps a track of all the information on files (i.e. the metadata on files) such as the access time of the file, which user is accessing a file on current time and which file is saved in which hadoop cluster. The secondary NameNode keeps a backup of the NameNode data.

Slave/Worker Node- This component in a hadoop cluster is responsible for storing the data and performing computations. Every slave/worker node runs both a TaskTracker and a DataNode service to communicate with the Master node in the cluster. The DataNode service is secondary to the NameNode and the TaskTracker service is secondary to the JobTracker.

Q.8

```
[root@hadoop-master:~# ./run-wordcount.sh
mkdir: cannot create directory 'input': File exists
rm: `output': No such file or directory
rm: `input': No such file or directory
20/11/15 17:49:45 INFO client.RMProxy: Connecting to ResourceManager at hadoop-master/172.18.0.2:8032
20/11/15 17:49:46 INFO input.FileInputFormat: Total input paths to process : 3
20/11/15 17:49:46 INFO mapreduce.JobSubmitter: number of splits:3
20/11/15 17:49:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1605462562448_0001
20/11/15 17:49:47 INFO impl.YarnClientImpl: Submitted application application_1605462562448_0001
20/11/15 17:49:47 INFO mapreduce.Job: The url to track the job: http://hadoop-master:8088/proxy/application_1605462562448_0001/
20/11/15 17:49:47 INFO mapreduce.Job: Running job: job_1605462562448_0001
20/11/15 17:49:55 INFO mapreduce.Job: Job job_1605462562448_0001 running in uber mode : false
20/11/15 17:49:55 INFO mapreduce.Job:  map 0% reduce 0%
20/11/15 17:50:08 INFO mapreduce.Job:  map 33% reduce 0%
20/11/15 17:50:10 INFO mapreduce.Job:  map 100% reduce 0%
20/11/15 17:50:17 INFO mapreduce.Job:  map 100% reduce 100%
20/11/15 17:50:17 INFO mapreduce.Job: Job job_1605462562448_0001 completed successfully
20/11/15 17:50:18 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=2418
                FILE: Number of bytes written=474573
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2990
                HDFS: Number of bytes written=1629
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=3
                Launched reduce tasks=1
                Data-local map tasks=2
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=36033
                Total time spent by all reduces in occupied slots (ms)=5542
                Total time spent by all map tasks (ms)=36033
                Total time spent by all reduce tasks (ms)=5542
                Total vcore-milliseconds taken by all map tasks=36033
                Total vcore-milliseconds taken by all reduce tasks=5542
                Total megabyte-milliseconds taken by all map tasks=36897792
                Total megabyte-milliseconds taken by all reduce tasks=5675008
        Map-Reduce Framework
                Map input records=3
                Map output records=497
                Map output bytes=4631
                Map output materialized bytes=2430
                Input split bytes=347
                Combine input records=497
                Combine output records=195
                Reduce input groups=194
                Reduce shuffle bytes=2430
                Reduce input records=195
                Reduce output records=194
                Spilled Records=390
                Shuffled Maps =3
                Failed Shuffles=0
                Merged Map outputs=3
                GC time elapsed (ms)=271
                CPU time spent (ms)=1920
                Physical memory (bytes) snapshot=1097105408
                Virtual memory (bytes) snapshot=3498110976
                Total committed heap usage (bytes)=719847424
```

3 mappers and 1 reducers are launched for executing the above program

Q.9

Total time spent by all maps in occupied slots is 36033 ms
Total time spent by all reduces in occupied slots is 5542 ms

Q.10

```
moment; 2
mornings        2
much    1
my      17
myself  2
neglect 2
never   2
noticed 2
now.    2
of      27
often   1
overspreads     1
own     1
paper   1
plants  2
possession      2
power,  1
presence        1
present 2
sanctuary,      2
seem    1
sense   2
serenity        2
should  2
single  2
sink    1
so      5
soul    2
soul,   3
souls   2
splendour       1
spot,   2
spring  2
stalks, 2
steal   2
stray   2
stream; 2
strength        1
strikes 2
stroke  2
sun     2
surface 2
sustains        1
sweet   2
taken   2
talents.        2
tall    2
teems   2
than    2
that    7
the     42
then    2
then,   1
these   4
think   1
this    2
thousand        2
throw   2
to      3
too     1
tranquil        2
trees,  2
trickling       2
```

The two most frequently occurring words are
“the” 42 times
“of” 27 times

Task 3

Q11

In the map function we tokenize the string and give each token a count of 1. This count of each token will be merged by the reducer.

Q12

Keeping the unique keys intact and combining the replica keys the created keys are arranged and finally are stored with the sum of their occurrences.

Q13

```
20/11/15 18:05:32 INFO mapreduce.Job: Running job: job_1605462562448_0002
20/11/15 18:05:39 INFO mapreduce.Job: Job job_1605462562448_0002 running in uber mode : false
20/11/15 18:05:39 INFO mapreduce.Job:  map 0% reduce 0%
20/11/15 18:05:47 INFO mapreduce.Job:  map 100% reduce 0%
20/11/15 18:05:55 INFO mapreduce.Job:  map 100% reduce 100%
20/11/15 18:05:55 INFO mapreduce.Job: Job job_1605462562448_0002 completed successfully
20/11/15 18:05:55 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=56
                FILE: Number of bytes written=351699
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=258
                HDFS: Number of bytes written=26
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=10844
                Total time spent by all reduces in occupied slots (ms)=4627
                Total time spent by all map tasks (ms)=10844
                Total time spent by all reduce tasks (ms)=4627
                Total vcore-milliseconds taken by all map tasks=10844
                Total vcore-milliseconds taken by all reduce tasks=4627
                Total megabyte-milliseconds taken by all map tasks=11104256
                Total megabyte-milliseconds taken by all reduce tasks=4738048
        Map-Reduce Framework
                Map input records=2
                Map output records=4
                Map output bytes=42
                Map output materialized bytes=62
                Input split bytes=232
                Combine input records=4
                Combine output records=4
                Reduce input groups=3
                Reduce shuffle bytes=62
                Reduce input records=4
                Reduce output records=3
                Spilled Records=8
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=130
                CPU time spent (ms)=1550
                Physical memory (bytes) snapshot=821354496
                Virtual memory (bytes) snapshot=2628014080
                Total committed heap usage (bytes)=516423680
        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=26
        File Output Format Counters
                Bytes Written=26
```

2 mappers and 1 reducers are launched for executing the above program

Q.14

Total time spent by all maps in occupied slots is 10844 ms
Total time spent by all reduces in occupied slots is 4627 ms

Task 4

```
                Total time spent by all map tasks (ms)=22729
                Total time spent by all reduce tasks (ms)=4169
                Total vcore-milliseconds taken by all map tasks=22729
                Total vcore-milliseconds taken by all reduce tasks=4169
                Total megabyte-milliseconds taken by all map tasks=23274496
                Total megabyte-milliseconds taken by all reduce tasks=4269056
        Map-Reduce Framework
                Map input records=3
                Map output records=18
                Map output bytes=144
                Map output materialized bytes=28
                Input split bytes=347
                Combine input records=18
                Combine output records=1
                Reduce input groups=1
                Reduce shuffle bytes=28
                Reduce input records=1
                Reduce output records=1
                Spilled Records=2
                Shuffled Maps =3
                Failed Shuffles=0
                Merged Map outputs=3
                GC time elapsed (ms)=172
                CPU time spent (ms)=1800
                Physical memory (bytes) snapshot=1150484480
                Virtual memory (bytes) snapshot=3502403584
                Total committed heap usage (bytes)=715128832
        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=2643
        File Output Format Counters
                Bytes Written=7

input file1.txt:
Hello Hadoop

input file2.txt:
Hello Docker

input text.txt:
A wonderful serenity has taken possession of my entire soul, like these sweet mornings of spring which I enjoy with my whole heart. I am alone, and feel the charm of existence in this spot, which was crea
ted for the bliss of souls like mine. I am so happy, my dear friend, so absorbed in the exquisite sense of mere tranquil existence, that I neglect my talents. I should be incapable of drawing a single str
oke at the present moment; and yet I feel that I never was a greater artist than now. When, while the lovely valley teems with vapour around me, and the meridian sun strikes the upper surface of the impen
etrable foliage of my trees, and but a few stray gleams steal into the inner sanctuary, I throw myself down among the tall grass by the trickling stream; and, as I lie close to the earth, a thousand unkno
wn plants are noticed by me: when I hear the buzz of the little world among the stalks, and grow familiar with the countless indescribable forms of the insects and flies, then I feel the presence of the A
lmighty, who formed us in his own image, and the breath of that universal love which bears and sustains us, as it floats around us in an eternity of bliss; and then, my friend, when darkness overspreads m
y eyes, and heaven and earth seem to dwell in my soul and absorb its power, like the form of a beloved mistress, then I often think with longing, Oh, would I could describe these conceptions, could impres
s upon paper all that is living so full and warm within me, that it might be the mirror of my soul, as my soul is the mirror of the infinite God! O my friend -- but it is too much for my strength -- I sin
k under the weight of the splendour of these visions! A wonderful serenity has taken possession of my entire soul, like these sweet mornings of spring which I enjoy with my whole heart. I am alone, and fe
el the charm of existence in this spot, which was created for the bliss of souls like mine. I am so happy, my dear friend, so absorbed in the exquisite sense of mere tranquil existence, that I neglect my
talents. I should be incapable of drawing a single stroke at the present moment; and yet I feel that I never was a greater artist than now. When, while the lovely valley teems with vapour around me, and t
he meridian sun strikes the upper surface of the impenetrable foliage of my trees, and but a few stray gleams steal into the inner sanctuary, I throw myself down among the tall grass by the trickling stre
am; and, as I lie close to the earth, a thousand unknown plants are noticed by me: when I hear the buzz of the little world among the stalks, and grow familiar with the

wordcount output:
and     18
root@hadoop-master:~#
```

The word and is repeated 18 times

Task 5

The program aims to get maximum temperature corresponding to the year.
In the input file we give years and corresponding temperatures that are noted
Eg:
1901 34
1901 45
1956 45
1956 47

The output would be maximum temperature for that year
1901 45
1956 47

~ — pwalke1@CS580-pwalke1: ~/CS580k/project2/hadoop-cluster-docker — ssh pwalke1@remote.cs.binghamton.edu ... ...dies-master/MapReduce/ColdAndHot Days — root@hadoop-master: ~ — ssh pwalke1@remote.cs.binghamton.edu +

```
20/11/15 21:34:47 INFO mapreduce.Job: Job job_1605475660950_0001 completed successfully
20/11/15 21:34:47 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=13195
                FILE: Number of bytes written=261167
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=9475
                HDFS: Number of bytes written=912
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=3695
                Total time spent by all reduces in occupied slots (ms)=5154
                Total time spent by all map tasks (ms)=3695
                Total time spent by all reduce tasks (ms)=5154
                Total vcore-milliseconds taken by all map tasks=3695
                Total vcore-milliseconds taken by all reduce tasks=5154
                Total megabyte-milliseconds taken by all map tasks=3783680
                Total megabyte-milliseconds taken by all reduce tasks=5277696
        Map-Reduce Framework
                Map input records=1199
                Map output records=1199
                Map output bytes=10791
                Map output materialized bytes=13195
                Input split bytes=115
                Combine input records=0
                Combine output records=0
                Reduce input groups=114
                Reduce shuffle bytes=13195
                Reduce input records=1199
                Reduce output records=114
                Spilled Records=2398
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=74
                CPU time spent (ms)=1240
                Physical memory (bytes) snapshot=513347584
                Virtual memory (bytes) snapshot=1756909568
                Total committed heap usage (bytes)=308281344
        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=9360
        File Output Format Counters
                Bytes Written=912

input text.txt:
1900 39
1900 14
```

~ — pwalke1@CS580-pwalke1: ~/CS580k/project2/hadoop-cluster-docker — ssh pwalke1@remote.cs.binghamton.edu ... ...dies-master/MapReduce/ColdAndHot Days — root@hadoop-master: ~ — ssh pwalke1@remote.cs.binghamton.edu +

```
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=9360
        File Output Format Counters
                Bytes Written=912

input text.txt:
1900 39
1900 14
1900 5
1900 11
1900 20
1900 20
1900 22
1900 15
1900 41
1900 42
1900 46
1900 6
1900 13
1900 13
1900 30
1900 45
1900 13
1900 16
1900 36
1900 29
1901 32
1901 40
1901 29
1901 48
1901 16
1901 11
1901 21
1901 6
1901 22
1902 49
1902 49
1902 13
1902 2
1902 49
1902 48
1902 1
1902 22
1902 2
1902 24
1902 39
1902 24
1903 13
1903 35
1903 35
1903 18
1903 5
1904 29
1904 23
1904 28
1904 46
1904 28
```

~ — pwalke1@CS580-pwalke1: ~/CS580k/project2/hadoop-cluster-docker — ssh pwalke1@remote.cs.binghamton.edu ... ...dies-master/MapReduce/ColdAndHot Days — root@hadoop-master: ~ — ssh pwalke1@remote.cs.binghamton.edu +

```
2012 5
2012 28
2012 21
2012 35
2012 33
2012 1
2012 45
2013 34
2013 26
2013 49

wordcount output:
1900    46
1901    48
1902    49
1903    35
1904    46
1905    35
1906    32
1907    49
1908    44
1909    38
1910    47
1911    48
1912    44
1913    43
1914    49
1915    49
1916    18
1917    35
1918    49
1919    42
1920    47
1921    47
1922    45
1923    41
1924    49
1925    48
1926    49
1927    47
1928    48
1929    35
1930    48
1931    37
1932    33
1933    43
1934    47
1935    40
1936    48
1937    44
1938    43
1939    48
1940    49
1941    49
1942    24
1943    45
1944    39
1945    47
1946    48
1947    41
1948    42
```

GitHub link to the code: https://github.com/pvwalke/CS580K