

# CS1699 course project

ZiYi Huang

The URL to GitHub repository (code and jar file) :

<https://github.com/YhzyY/cs1699project>

The URL to docker repository (image):

<https://hub.docker.com/repository/docker/ziyihuang/cs1699project>

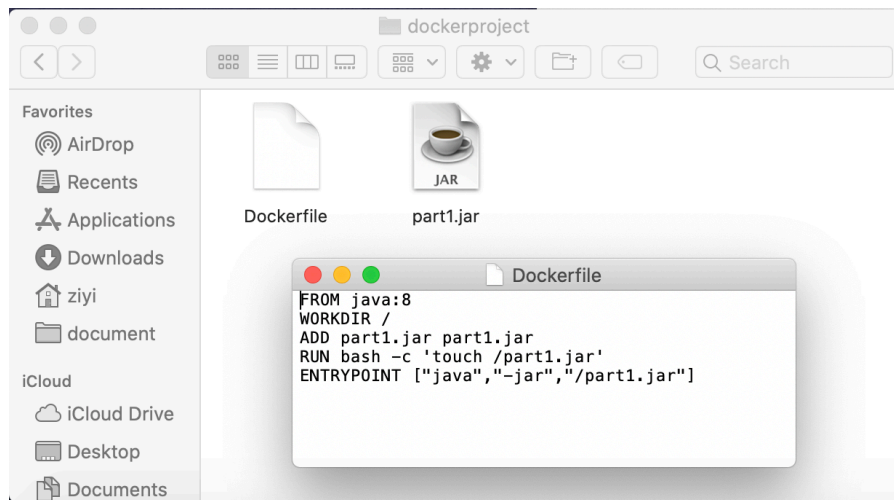
Whole project:

```
CS1699 — -bash — 124x49
^CZiYideMacBook-Pro:dockerproject ziyi$ docker run -it --rm -v /Users/ziyi/document:/Users/ziyi/document 1699project bash
the files you want to upload:
/Users/ziyi/document/cs1699/project/data/Hugo
/Users/ziyi/document/cs1699/project/data/Hugo/NotreDame_De_Paris.txt
/Users/ziyi/document/cs1699/project/data/Hugo/Miserables.txt
null null
null Blob{bucket=dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2, name=output3/part-r-00000, generation=1575148713211164, size=1584239, content-type=application/octet-stream, metadata=null}
null Blob{bucket=dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2, name=output3/part-r-00000, generation=1575148713211164, size=1584239, content-type=application/octet-stream, metadata=null}
Blob{bucket=dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2, name=output4/part-r-00000, generation=1575148793830607, size=331747, content-type=application/octet-stream, metadata=null} Blob{bucket=dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2, name=output3/part-r-00000, generation=1575148713211164, size=1584239, content-type=application/octet-stream, metadata=null}
jobs finished
Type "search" to search a term; Type "TopN" to find topN term; type "exit" to exit the system
search
the word you want to search:
ttt
no such word in files
Type "search" to search a term; Type "TopN" to find topN term; type "exit" to exit the system
search
the word you want to search:
this
files/NotreDame_De_Paris.txt : 782
files/Miserables.txt : 2975

Type "search" to search a term; Type "TopN" to find topN term; type "exit" to exit the system
topN
the value of N:
10
50297 the
27047 of
18781 and
18108 a
17741 to
13559 in
10778 was
9656 that
8451 he
8032 his
Type "search" to search a term; Type "TopN" to find topN term; type "exit" to exit the system
topN
the value of N:
-1
N should be between 1 and 37199
Type "search" to search a term; Type "TopN" to find topN term; type "exit" to exit the system
exit
exit the system
```

Docker:

Dockerfile and .jar file (.jar file can be found in the above GitHub repository)



Build docker image:

```
CS1699 — -bash — 124x48
[ZiyideMacBook-Pro:project ziyi$ cd dockerproject
[ZiyideMacBook-Pro:dockerproject ziyi$ ls
Dockerfile    part1.jar
[ZiyideMacBook-Pro:dockerproject ziyi$ ls
Dockerfile    part1.jar
[ZiyideMacBook-Pro:dockerproject ziyi$ docker build -t="1699project" .
Sending build context to Docker daemon  15.8MB
Step 1/5 : FROM java:8
----> d23bdf5b1b1b
Step 2/5 : WORKDIR /
----> Using cache
----> 6a73175cde39
Step 3/5 : ADD part1.jar part1.jar
----> 1738c9fb1588
Step 4/5 : RUN bash -c 'touch /part1.jar'
----> Running in d39dc36d93ea
Removing intermediate container d39dc36d93ea
----> fb40f354085d
Step 5/5 : ENTRYPOINT ["java","-jar","/part1.jar"]
----> Running in 0f1dc9ff2c55
Removing intermediate container 0f1dc9ff2c55
----> 3a93cead52ea
Successfully built 3a93cead52ea
Successfully tagged 1699project:latest
ZiyideMacBook-Pro:dockerproject ziyi$ docker image ls
REPOSITORY          TAG                 IMAGE ID            CREATED             SIZE
1699project          latest              3a93cead52ea        6 seconds ago      675MB
cs1699project        latest              250467566e43        14 hours ago       675MB
<none>               <none>              c3bc48e28a2e        14 hours ago       643MB
<none>               <none>              e93eb057ce4a        15 hours ago       704MB
<none>               <none>              7009c77711b6        15 hours ago       704MB
centos               latest              0f3e07c0138f        8 weeks ago        220MB
hello                latest              7e478924fb09        2 months ago       643MB
ziyihuang/hello      latest              7e478924fb09        2 months ago       643MB
<none>               <none>              c8c7ada422af        2 months ago       475MB
<none>               <none>              c411ee4dd38e        2 months ago       475MB
<none>               <none>              c190c1d07f04        2 months ago       475MB
<none>               <none>              7d28a14b323a        2 months ago       475MB
<none>               <none>              dc1835616e37        2 months ago       643MB
<none>               <none>              c11f6ccdeb8d        2 months ago       643MB
<none>               <none>              f6e5459d61a8        2 months ago       643MB
ziyihuang/cheers2019 latest              685ecff6e9de        2 months ago       4.01MB
<none>               <none>              4d4d64c66764        2 months ago       356MB
golang               1.11-alpine        e116d2efa2ab        3 months ago       312MB
openjdk              latest              e1e07dfba89c        3 months ago       470MB
openjdk              7                  d735a2057e60        6 months ago       475MB
java                 8                  d23bdf5b1b1b        2 years ago        643MB
ZiyideMacBook-Pro:dockerproject ziyi$ docker run -it --rm -v /Users/ziyi/document:/Users/ziyi/document 1699project bash
```



Run the docker image:

```
CS1699 — -bash — 124x49
^Cziyi@MacBook-Pro:dockerproject ziyi$ docker run -it --rm -v /Users/ziyi/document:/Users/ziyi/document 1699project bash
the files you want to upload:
/Users/ziyi/document/cs1699/project/data/Hugo
/Users/ziyi/document/cs1699/project/data/Hugo/NotreDame_De_Paris.txt
/Users/ziyi/document/cs1699/project/data/Hugo/Miserables.txt
null null
null Blob{bucket=dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2, name=output3/part-r-00000, generation=157514871321
1164, size=1584239, content-type=application/octet-stream, metadata=null}
null Blob{bucket=dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2, name=output3/part-r-00000, generation=157514871321
1164, size=1584239, content-type=application/octet-stream, metadata=null}
Blob{bucket=dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2, name=output4/part-r-00000, generation=1575148793830607,
size=331747, content-type=application/octet-stream, metadata=null} Blob{bucket=dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af2
3-us-west2, name=output3/part-r-00000, generation=1575148713211164, size=1584239, content-type=application/octet-stream, met
adata=null}
jobs finished
Type "search" to search a term; Type "TopN" to find topN term; type "exit" to exit the system
search
the word you want to search:
ttt
no such word in files
Type "search" to search a term; Type "TopN" to find topN term; type "exit" to exit the system
search
the word you want to search:
this
files/NotreDame_De_Paris.txt : 782
files/Miserables.txt : 2975
Type "search" to search a term; Type "TopN" to find topN term; type "exit" to exit the system
topN
the value of N:
10
50297 the
27047 of
18781 and
18108 a
17741 to
13559 in
10778 was
9656 that
8451 he
8032 his
Type "search" to search a term; Type "TopN" to find topN term; type "exit" to exit the system
topN
the value of N:
-1
N should be between 1 and 37199
Type "search" to search a term; Type "TopN" to find topN term; type "exit" to exit the system
exit
exit the system
```

GCP:

The input path used by the above two job is

gs://dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2/files

The output path used by the inverted indices job is

gs://dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2/output3

The output path used by the topN job is

gs://dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2/output4

GCP inverted indices job:

(the jar file it used is JAR/jar16.jar in the GitHub repository )

[←](#) Job details [REFRESH](#) [CLONE](#)

✓ 4f1e6ba8-cc51-4779-8414-d942c9fd588c

Start time: Nov 30, 2019, 4:17:53 PM Elapsed time: 46 sec Status:

Output Configuration

☐ Line wrapping

Total time spent by all reduce tasks (ms)=7012  
Total vcore-milliseconds taken by all map tasks=22780  
Total vcore-milliseconds taken by all reduce tasks=7012  
Total megabyte-milliseconds taken by all map tasks=46653440  
Total megabyte-milliseconds taken by all reduce tasks=14360576  
Map-Reduce Framework  
Map input records=2  
Map output records=778013  
Map output bytes=23619667  
Map output materialized bytes=1681439  
Input split bytes=298  
Combine input records=778013  
Combine output records=43876  
Reduce input groups=32970  
Reduce shuffle bytes=1681439  
Reduce input records=43876  
Reduce output records=32970  
Spilled Records=87752  
Shuffled Maps =2  
Failed Shuffles=0  
Merged Map outputs=2  
GC time elapsed (ms)=2008  
CPU time spent (ms)=10490  
Physical memory (bytes) snapshot=1476657152  
Virtual memory (bytes) snapshot=10542686208  
Total committed heap usage (bytes)=1366294528  
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=4334239  
File Output Format Counters  
Bytes Written=1584239

Job output is complete

✓ 4f1e6ba8-cc51-4779-8414-d942c9fd588c

Start time: Nov 30, 2019, 4:17:53 PM Elapsed time: 46 sec Status:

Output [Configuration](#)

Edit

Region	us-west2
Cluster	cs1699project
Job type	Hadoop
Main class or jar	gs://dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2/JAR/jar16.jar
Jar files	
Properties	
Arguments	gs://dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2/files
	gs://dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2/output3

Labels

+ Add label

GCP topN job:

(the jar file it used is JAR/top13.jar in the GitHub repository )

[←](#) **Job details** [REFRESH](#) [CLONE](#)

✓ 5e161744-0c03-46f3-9a93-ba9e7e4ae8d3

Start time: Nov 30, 2019, 4:17:53 PM Elapsed time: 2 min 6 sec Status:

Output

Configuration

☐ Line wrapping

Total time spent by all reduce tasks (ms)=5559  
Total vcore-milliseconds taken by all map tasks=60437  
Total vcore-milliseconds taken by all reduce tasks=5559  
Total megabyte-milliseconds taken by all map tasks=123774976  
Total megabyte-milliseconds taken by all reduce tasks=11384832  
Map-Reduce Framework  
Map input records=31417  
Map output records=31417  
Map output bytes=387558  
Map output materialized bytes=450422  
Input split bytes=660  
Combine input records=0  
Combine output records=0  
Reduce input groups=515  
Reduce shuffle bytes=450422  
Reduce input records=31417  
Reduce output records=31417  
Spilled Records=62834  
Shuffled Maps =5  
Failed Shuffles=0  
Merged Map outputs=5  
GC time elapsed (ms)=1648  
CPU time spent (ms)=10210  
Physical memory (bytes) snapshot=2906820608  
Virtual memory (bytes) snapshot=21083021312  
Total committed heap usage (bytes)=2485649408  
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=645619  
File Output Format Counters  
Bytes Written=331747

Job output is complete

✔ 5e161744-0c03-46f3-9a93-ba9e7e4ae8d3

Start time: Nov 30, 2019, 4:17:53 PM Elapsed time: 2 min 6 sec Status:

Output Configuration

Edit

Region	us-west2
Cluster	cs1699project
Job type	Hadoop
Main class or jar	gs://dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2/TOPJAR/top13.jar
Jar files	
Properties	
Arguments	gs://dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2/files
	gs://dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2/output4
Labels	<div>+ Add label</div>









## GCP Storage:

dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2

[Objects](#) [Overview](#) [Permissions](#) [Bucket Lock](#)

[Upload files](#) [Upload folder](#) [Create folder](#) [Manage holds](#) [Delete](#)

[Buckets](#) / dataproc-effe9cb0-28c9-4f0d-94d4-13f79e57af23-us-west2

<input type="checkbox"/> Name	Size	Type	Storage class	Last modified	Public access <span>?</span>
<input type="checkbox"/>  JAR/	—	Folder	—	—	Per object
<input type="checkbox"/>  TOPJAR/	—	Folder	—	—	Per object
<input type="checkbox"/>  files/	—	Folder	—	—	Per object
<input type="checkbox"/>  google-cloud-dataproc-metainfo/	—	Folder	—	—	Per object
<input type="checkbox"/>  output/	—	Folder	—	—	Per object
<input type="checkbox"/>  output2/	—	Folder	—	—	Per object
<input type="checkbox"/>  <u>output3/</u>	—	Folder	—	—	Per object
<input type="checkbox"/>  output4/	—	Folder	—	—	Per object

GCP Output files are stored in GitHub



output3.txt

```
"A      files/Miserables.txt : 82; files/NotreDame_De_Paris.txt : 31;
"Abandoned    files/Miserables.txt : 1;
"Abbé    files/NotreDame_De_Paris.txt : 1;
"Abominable    files/Miserables.txt : 1;
"Abomination    files/NotreDame_De_Paris.txt : 1;
"About    files/Miserables.txt : 3;
"Above    files/Miserables.txt : 2;
"According    files/Miserables.txt : 1;
"Accordingly    files/NotreDame_De_Paris.txt : 1;
"Accused    files/NotreDame_De_Paris.txt : 3;
"Actually    files/NotreDame_De_Paris.txt : 1;
"Address    files/Miserables.txt : 1;
"Adieu    files/Miserables.txt : 3; files/NotreDame_De_Paris.txt : 1;
"Admirable    files/NotreDame_De_Paris.txt : 1;
"Admit    files/Miserables.txt : 2;
"Adorable    files/Miserables.txt : 1;
"Advocate    files/NotreDame_De_Paris.txt : 1;
"After    files/NotreDame_De_Paris.txt : 3; files/Miserables.txt : 12;
"Again    files/Miserables.txt : 1;
"Against    files/NotreDame_De_Paris.txt : 2;
"Agreed    files/Miserables.txt : 4;
"Ah      files/Miserables.txt : 130; files/NotreDame_De_Paris.txt : 39;
"Aie    files/Miserables.txt : 1;
"Aigle    files/Miserables.txt : 1;
"Ainsi    files/Miserables.txt : 1;
"Alarm    files/Miserables.txt : 1;
"Alas    files/Miserables.txt : 7; files/NotreDame_De_Paris.txt : 21;
"Alchemy    files/NotreDame_De_Paris.txt : 1;
"All      files/NotreDame_De_Paris.txt : 4; files/Miserables.txt : 20;
"Alone    files/Miserables.txt : 2; files/NotreDame_De_Paris.txt : 1;
"Aloud    files/Miserables.txt : 1;
"Already    files/NotreDame_De_Paris.txt : 2;
```

output4.txt

```
50297 the
27047 of
18781 and
18108 a
17741 to
13559 in
10778 was
9656 that
8451 he
8032 his
7555 had
7451 is
6841 which
6399 it
6280 with
5320 on
5221 The
5213 I
4822 not
4673 at
4279 you
3983 her
3783 this
3709 him
3704 for
3699 as
3697 one
3520 have
3359 He
3174 from
```