ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY
ΜΕΤΑΠΤΥΧΙΑΚΟ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ
MSc IN DATA SCIENCE

Course: Large Scale Data Management

Professor: Panagiotis Liakos

Student name: Vasileios Ilias Drouzas

Student ID: f3352301

1st Assignment

In this project, we will be using the Hadoop Map-Reduce framework. In Part I, we will build a map-reduce application while in Part II we will develop our own map-reduce application.

PART I:

For this part, we will use a book from Agatha Christie, "The mystery of the Blue train". Here, we will be using the Vagrantfile that is provided, but we will do two minor changes, in order to run our own text file:

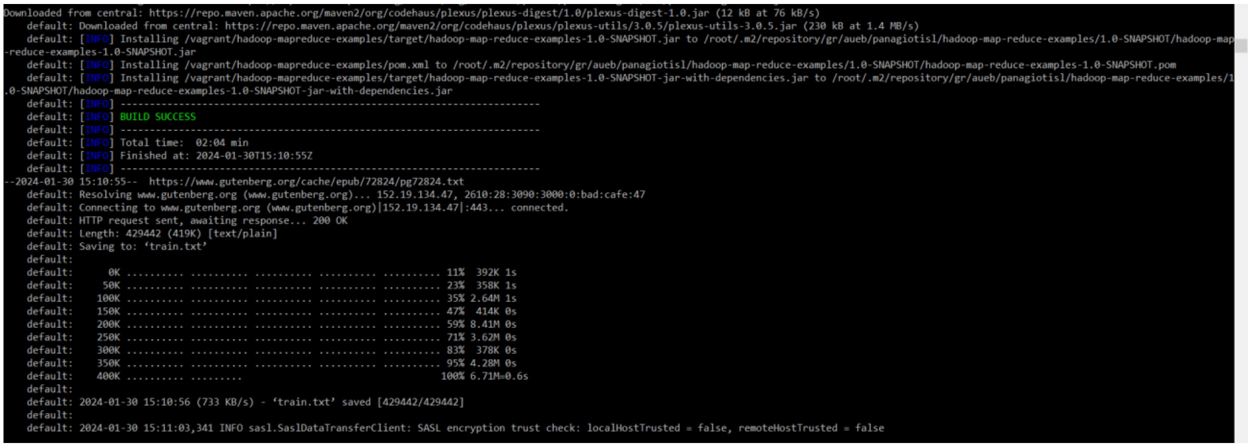
A) we will alter in the vagrant file the link to our text file and the name of the file to *train.txt*.

B) we will change the name in the java executable source file **Driver.java** to *train.txt* to make sure we will produce the correct .jar file.

The next steps are the following:

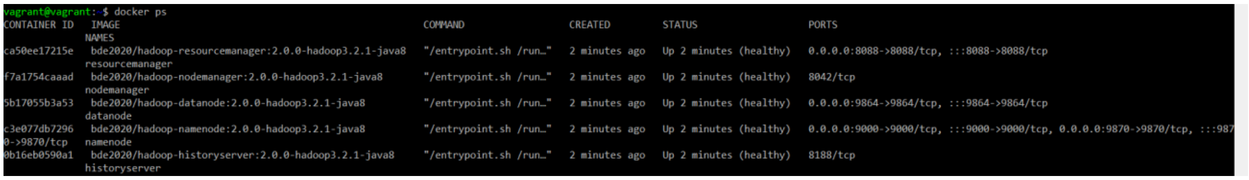
(Note: Here, not all the output is presented (only the most important aspects). The whole output is on 'output.txt').

1) Running `vagrant up` we get a successful build at 02:04 min.



```
Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-digest/1.0/plexus-digest-1.0.jar (12 kB at 76 kB/s)
   default: Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-utils/3.0.5/plexus-utils-3.0.5.jar (230 kB at 1.4 MB/s)
   default: [INFO] Installing /vagrant/hadoop-mapreduce-examples/target/hadoop-map-reduce-examples-1.0-SNAPSHOT.jar to /root/.m2/repository/gr/aueb/panagiotisl/hadoop-map-reduce-examples/1.0-SNAPSHOT/hadoop-map
-reduce-examples-1.0-SNAPSHOT.jar
   default: [INFO] Installing /vagrant/hadoop-mapreduce-examples/pom.xml to /root/.m2/repository/gr/aueb/panagiotisl/hadoop-map-reduce-examples/1.0-SNAPSHOT/hadoop-map-reduce-examples-1.0-SNAPSHOT.pom
   default: [INFO] Installing /vagrant/hadoop-mapreduce-examples/target/hadoop-map-reduce-examples-1.0-SNAPSHOT-jar-with-dependencies.jar to /root/.m2/repository/gr/aueb/panagiotisl/hadoop-map-reduce-examples/1
.0-SNAPSHOT/hadoop-map-reduce-examples-1.0-SNAPSHOT-jar-with-dependencies.jar
   default: [INFO] ------------------------------------------------------------------------
   default: [INFO] BUILD SUCCESS
   default: [INFO] ------------------------------------------------------------------------
   default: [INFO] Total time:  02:04 min
   default: [INFO] Finished at: 2024-01-30T15:10:55Z
   default: [INFO] ------------------------------------------------------------------------
--2024-01-30 15:10:55--  https://www.gutenberg.org/cache/epub/72824/pg72824.txt
   default: Resolving www.gutenberg.org (www.gutenberg.org)... 152.19.134.47, 2610:28:3090:3000:0:bad:cafe:47
   default: Connecting to www.gutenberg.org (www.gutenberg.org)|152.19.134.47|:443... connected.
   default: HTTP request sent, awaiting response... 200 OK
   default: Length: 429442 (419K) [text/plain]
   default: Saving to: ‘train.txt’
   default:
   default:      0K .......... .......... .......... .......... .......... 11%  392K 1s
   default:     50K .......... .......... .......... .......... .......... 23%  358K 1s
   default:    100K .......... .......... .......... .......... .......... 35% 2.64M 1s
   default:    150K .......... .......... .......... .......... .......... 47%  414K 0s
   default:    200K .......... .......... .......... .......... .......... 59% 8.41M 0s
   default:    250K .......... .......... .......... .......... .......... 71% 3.62M 0s
   default:    300K .......... .......... .......... .......... .......... 83%  378K 0s
   default:    350K .......... .......... .......... .......... .......... 95% 4.28M 0s
   default:    400K .......... .........                                  100% 6.71M=0.6s
   default:
   default: 2024-01-30 15:10:56 (733 KB/s) - ‘train.txt’ saved [429442/429442]
   default:
   default: 2024-01-30 15:11:03,341 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
```

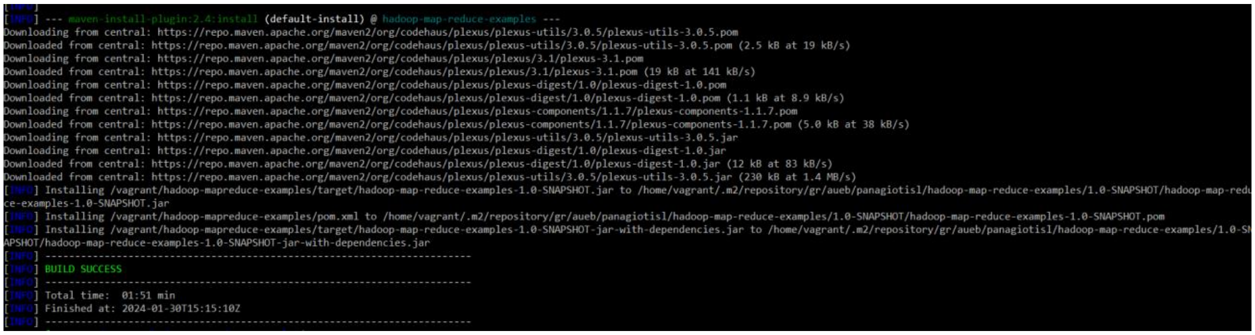
2) We will execute `vagrant ssh` to get to the environment of the vagrant and then with `docker ps` we can check if everything went as normal:



```
vagrant@vagrant:~$ docker ps
CONTAINER ID   IMAGE                                                COMMAND                  CREATED         STATUS                   PORTS
               NAMES
ca50ee17215e   bde2020/hadoop-resourcemanager:2.0.0-hadoop3.2.1-java8   "/entrypoint.sh /run…"   2 minutes ago   Up 2 minutes (healthy)   0.0.0.0:8088->8088/tcp, :::8088->8088/tcp
               resourcemanager
f7a1754caaad   bde2020/hadoop-nodemanager:2.0.0-hadoop3.2.1-java8       "/entrypoint.sh /run…"   2 minutes ago   Up 2 minutes (healthy)   8042/tcp
               nodemanager
5b17055b3a53   bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8          "/entrypoint.sh /run…"   2 minutes ago   Up 2 minutes (healthy)   0.0.0.0:9864->9864/tcp, :::9864->9864/tcp
               datanode
c3e077db7296   bde2020/hadoop-namenode:2.0.0-hadoop3.2.1-java8          "/entrypoint.sh /run…"   2 minutes ago   Up 2 minutes (healthy)   0.0.0.0:9000->9000/tcp, :::9000->9000/tcp, 0.0.0.0:9870->9870/tcp, :::987
0->9870/tcp    namenode
0b16eb0590a1   bde2020/hadoop-historyserver:2.0.0-hadoop3.2.1-java8     "/entrypoint.sh /run…"   2 minutes ago   Up 2 minutes (healthy)   8188/tcp
               historyserver
```

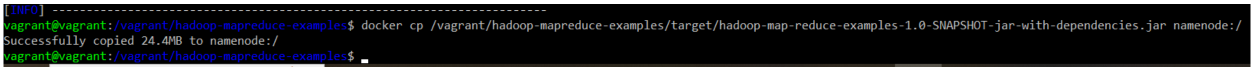
We get all the expected containers and they are all set up and running, so we are ready to go!

3) Next, we will change directory and install maven:

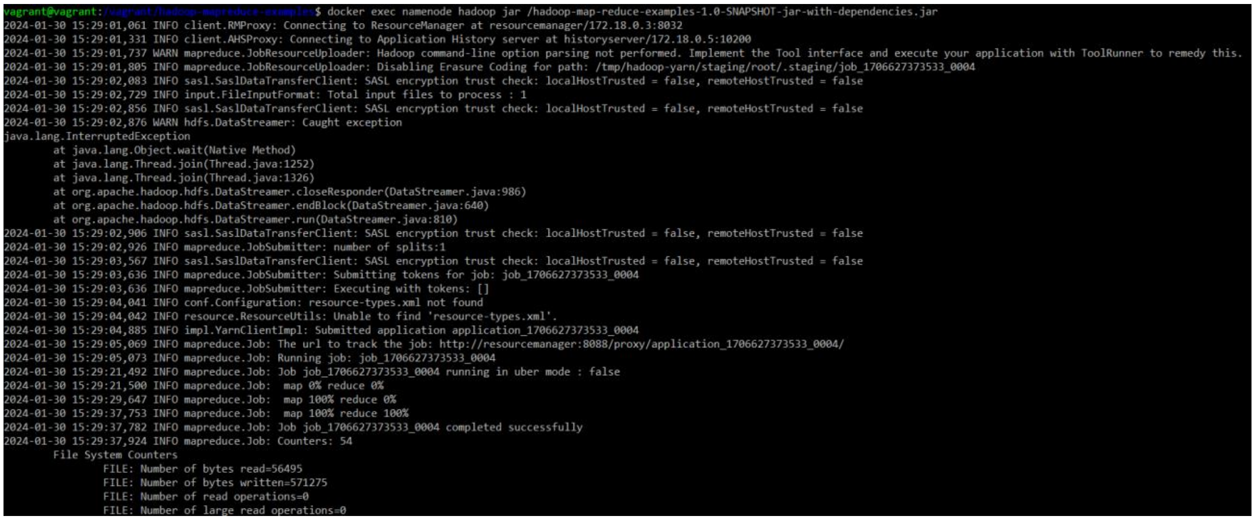


```
[INFO]
[INFO] --- maven-install-plugin:2.4:install (default-install) @ hadoop-map-reduce-examples ---
Downloading from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-utils/3.0.5/plexus-utils-3.0.5.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-utils/3.0.5/plexus-utils-3.0.5.pom (2.5 kB at 19 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus/3.1/plexus-3.1.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus/3.1/plexus-3.1.pom (19 kB at 141 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-digest/1.0/plexus-digest-1.0.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-digest/1.0/plexus-digest-1.0.pom (1.1 kB at 8.9 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-components/1.1.7/plexus-components-1.1.7.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-components/1.1.7/plexus-components-1.1.7.pom (5.0 kB at 38 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-utils/3.0.5/plexus-utils-3.0.5.jar
Downloading from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-digest/1.0/plexus-digest-1.0.jar
Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-digest/1.0/plexus-digest-1.0.jar (12 kB at 83 kB/s)
Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-utils/3.0.5/plexus-utils-3.0.5.jar (230 kB at 1.4 MB/s)
[INFO] Installing /vagrant/hadoop-mapreduce-examples/target/hadoop-map-reduce-examples-1.0-SNAPSHOT.jar to /home/vagrant/.m2/repository/gr/aueb/panagiotisl/hadoop-map-reduce-examples/1.0-SNAPSHOT/hadoop-map-redu
ce-examples-1.0-SNAPSHOT.jar
[INFO] Installing /vagrant/hadoop-mapreduce-examples/pom.xml to /home/vagrant/.m2/repository/gr/aueb/panagiotisl/hadoop-map-reduce-examples/1.0-SNAPSHOT/hadoop-map-reduce-examples-1.0-SNAPSHOT.pom
[INFO] Installing /vagrant/hadoop-mapreduce-examples/target/hadoop-map-reduce-examples-1.0-SNAPSHOT-jar-with-dependencies.jar to /home/vagrant/.m2/repository/gr/aueb/panagiotisl/hadoop-map-reduce-examples/1.0-SN
APSHOT/hadoop-map-reduce-examples-1.0-SNAPSHOT-jar-with-dependencies.jar
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  01:51 min
[INFO] Finished at: 2024-01-30T15:15:10Z
[INFO] ------------------------------------------------------------------------
```

4) Now it’s time to copy our application inside a docker container and execute it!

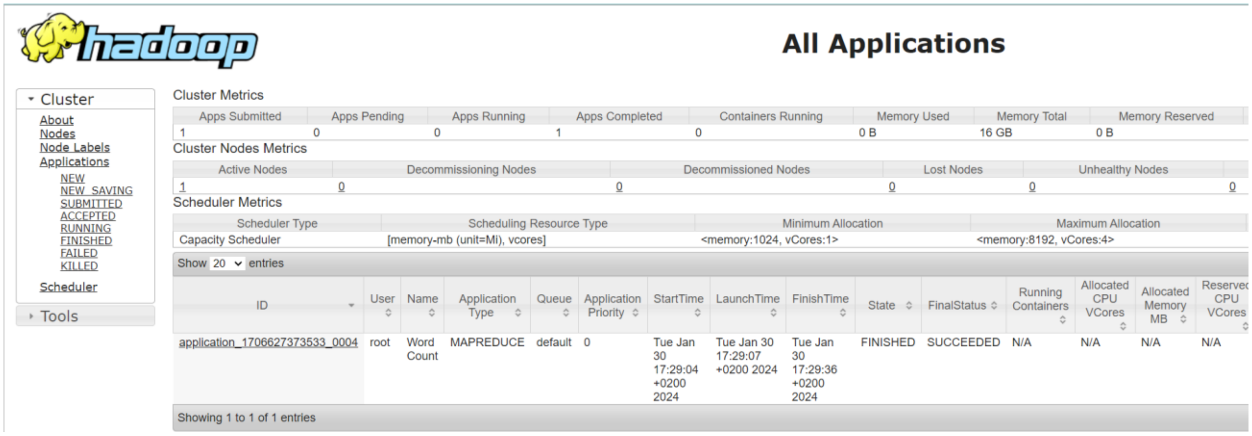


```
[INFO] ------------------------------------------------------------------------
vagrant@vagrant:/vagrant/hadoop-mapreduce-examples$ docker cp /vagrant/hadoop-mapreduce-examples/target/hadoop-map-reduce-examples-1.0-SNAPSHOT-jar-with-dependencies.jar namenode:/
Successfully copied 24.4MB to namenode:/
vagrant@vagrant:/vagrant/hadoop-mapreduce-examples$ _
```



```
vagrant@vagrant:/vagrant/hadoop-mapreduce-examples$ docker exec namenode hadoop jar /hadoop-map-reduce-examples-1.0-SNAPSHOT-jar-with-dependencies.jar
2024-01-30 15:29:01,061 INFO client.RMProxy: Connecting to ResourceManager at resourcemanager/172.18.0.3:8032
2024-01-30 15:29:01,331 INFO client.AHSProxy: Connecting to Application History server at historyserver/172.18.0.5:10200
2024-01-30 15:29:01,737 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2024-01-30 15:29:01,805 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/root/.staging/job_1706627373533_0004
2024-01-30 15:29:02,083 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-01-30 15:29:02,729 INFO input.FileInputFormat: Total input files to process : 1
2024-01-30 15:29:02,856 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-01-30 15:29:02,876 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:986)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:640)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:810)
2024-01-30 15:29:02,906 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-01-30 15:29:02,926 INFO mapreduce.JobSubmitter: number of splits:1
2024-01-30 15:29:03,567 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-01-30 15:29:03,636 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1706627373533_0004
2024-01-30 15:29:03,636 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-01-30 15:29:04,041 INFO conf.Configuration: resource-types.xml not found
2024-01-30 15:29:04,042 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-01-30 15:29:04,885 INFO impl.YarnClientImpl: Submitted application application_1706627373533_0004
2024-01-30 15:29:05,069 INFO mapreduce.Job: The url to track the job: http://resourcemanager:8088/proxy/application_1706627373533_0004/
2024-01-30 15:29:05,073 INFO mapreduce.Job: Running job: job_1706627373533_0004
2024-01-30 15:29:21,492 INFO mapreduce.Job: Job job_1706627373533_0004 running in uber mode : false
2024-01-30 15:29:21,500 INFO mapreduce.Job:  map 0% reduce 0%
2024-01-30 15:29:29,647 INFO mapreduce.Job:  map 100% reduce 0%
2024-01-30 15:29:37,753 INFO mapreduce.Job:  map 100% reduce 100%
2024-01-30 15:29:37,782 INFO mapreduce.Job: Job job_1706627373533_0004 completed successfully
2024-01-30 15:29:37,924 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=56495
                FILE: Number of bytes written=571275
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
```

Even though we got a (warning) exception in the DataStreamer, our execution is not interrupted and the application is finely produced. Now we can check whether our application is running in http://localhost:8088 :



All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved
1	0	0	1	0	0 B	16 GB	0 B

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	
1	0	0	0	0	0

Scheduler Metrics

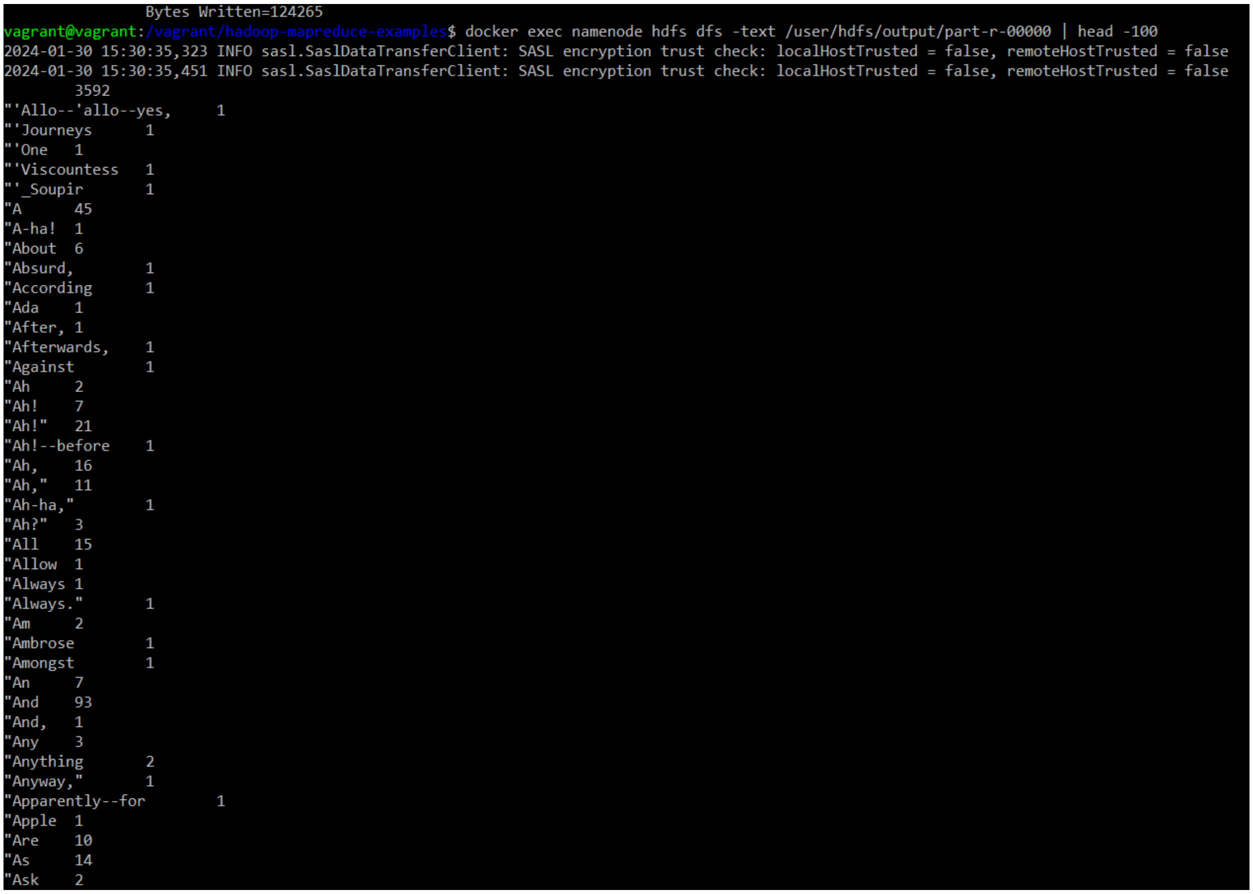
Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show 20 entries

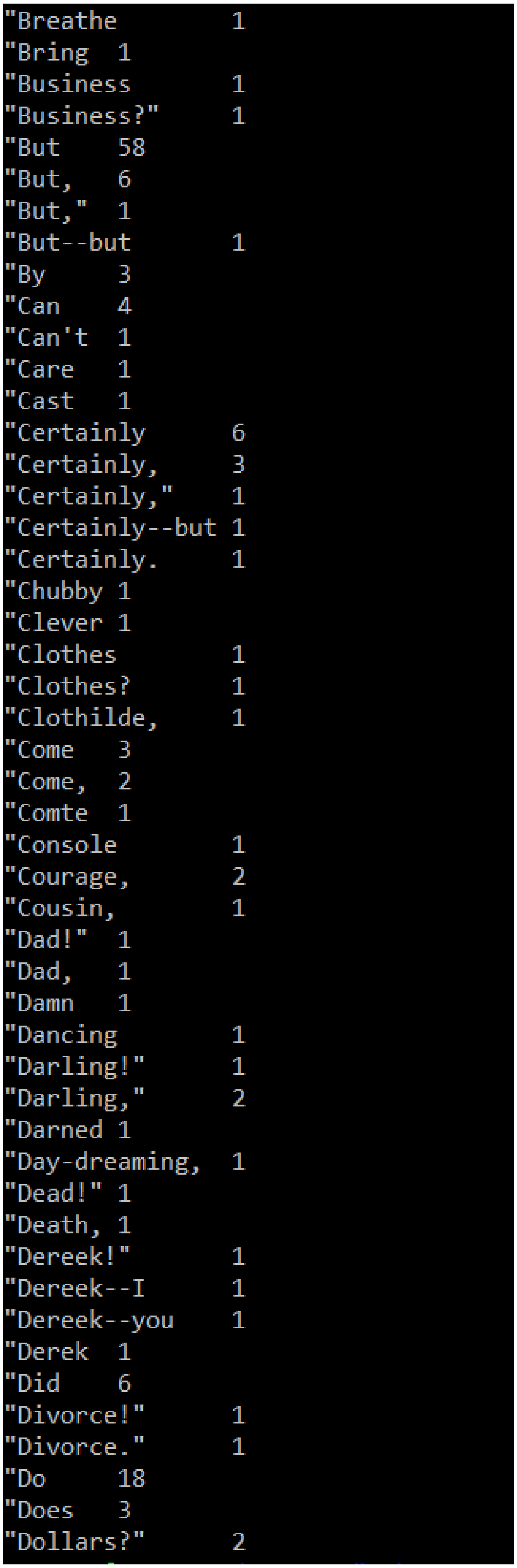
ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reservec CPU VCores
application_1706627373533_0004	root	Word Count	MAPREDUCE	default	0	Tue Jan 30 17:29:04 +0200 2024	Tue Jan 30 17:29:07 +0200 2024	Tue Jan 30 17:29:36 +0200 2024	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A

Showing 1 to 1 of 1 entries

Our application is up and running! We can also check the first 100 lines of the results:



```
                Bytes Written=124265
vagrant@vagrant:/vagrant/hadoop-mapreduce-examples$ docker exec namenode hdfs dfs -text /user/hdfs/output/part-r-00000 | head -100
2024-01-30 15:30:35,323 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2024-01-30 15:30:35,451 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
        3592
"'Allo--'allo--yes,     1
"'Journeys      1
"'One   1
"'Viscountess   1
"'_Soupir       1
"A      45
"A-ha!  1
"About  6
"Absurd,        1
"According      1
"Ada    1
"After, 1
"Afterwards,    1
"Against        1
"Ah     2
"Ah!    7
"Ah!"   21
"Ah!--before    1
"Ah,    16
"Ah,"   11
"Ah-ha,"        1
"Ah?"   3
"All    15
"Allow  1
"Always 1
"Always."       1
"Am     2
"Ambrose        1
"Amongst        1
"An     7
"And    93
"And,   1
"Any    3
"Anything       2
"Anyway,"       1
"Apparently--for        1
"Apple  1
"Are    10
"As     14
"Ask    2
```

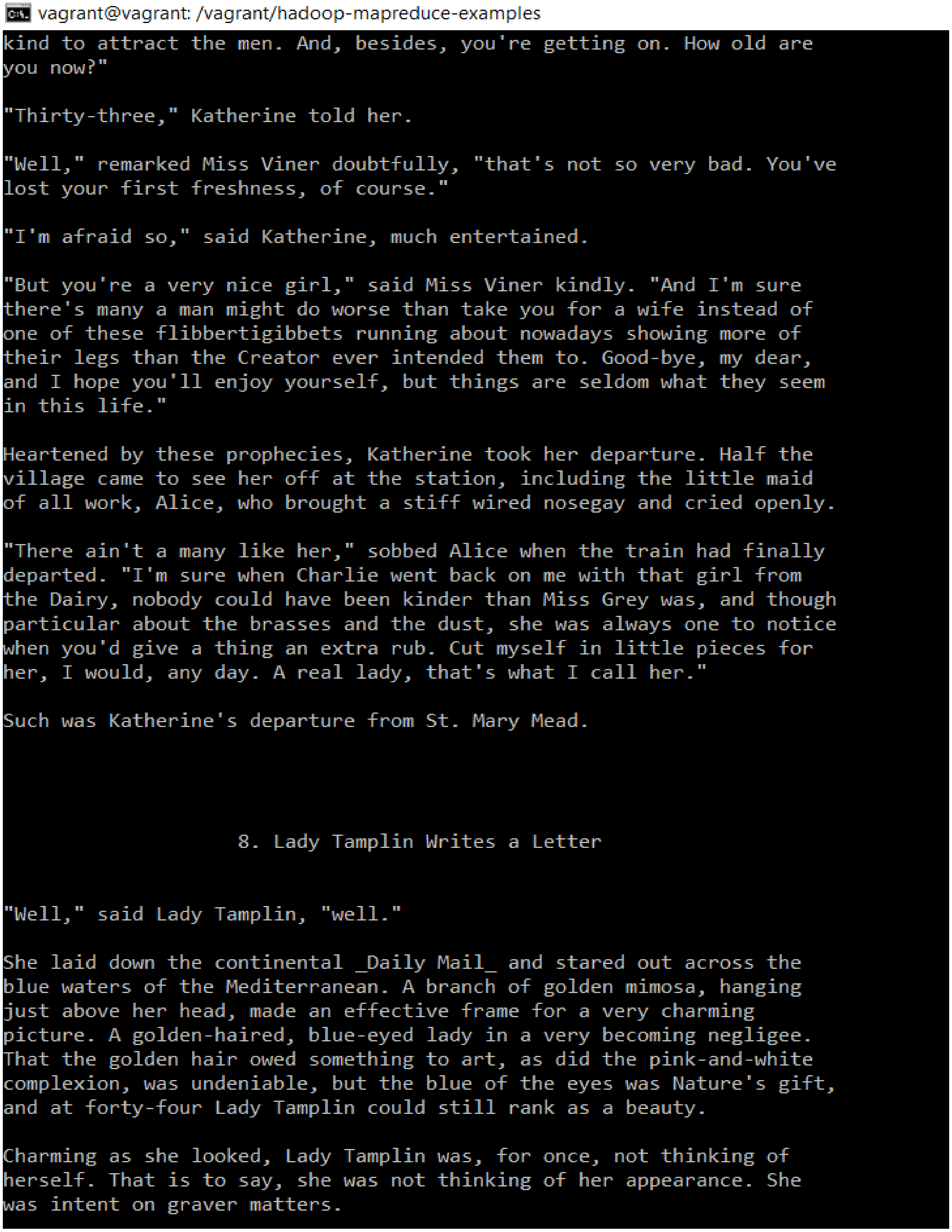


```
"Breathe        1
"Bring  1
"Business       1
"Business?"     1
"But    58
"But,   6
"But,"  1
"But--but       1
"By     3
"Can    4
"Can't  1
"Care   1
"Cast   1
"Certainly      6
"Certainly,     3
"Certainly,"    1
"Certainly--but 1
"Certainly.     1
"Chubby 1
"Clever 1
"Clothes        1
"Clothes?       1
"Clothilde,     1
"Come   3
"Come,  2
"Comte  1
"Console        1
"Courage,       2
"Cousin,        1
"Dad!"  1
"Dad,   1
"Damn   1
"Dancing        1
"Darling!"      1
"Darling,"      2
"Darned 1
"Day-dreaming,  1
"Dead!" 1
"Death, 1
"Dereek!"       1
"Dereek--I      1
"Dereek--you    1
"Derek  1
"Did    6
"Divorce!"      1
"Divorce."      1
"Do     18
"Does   3
"Dollars?"      2
```

Finally, we can print the *train.txt* file with the following command:

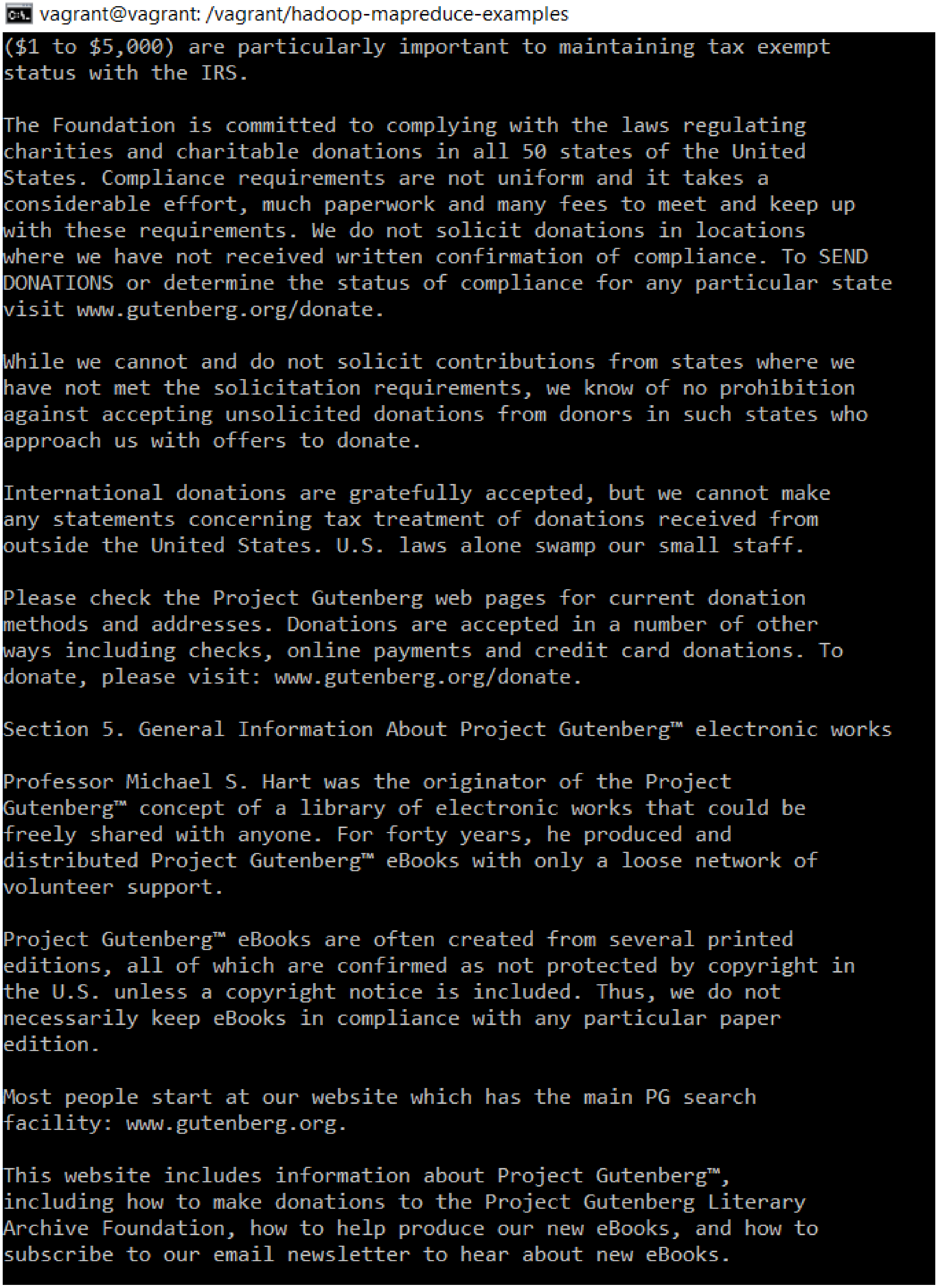
```
docker exec namenode hdfs dfs -text /user/hdfs/input/train.txt
```

A random peak at the data:



```
kind to attract the men. And, besides, you're getting on. How old are
you now?"

"Thirty-three," Katherine told her.

"Well," remarked Miss Viner doubtfully, "that's not so very bad. You've
lost your first freshness, of course."

"I'm afraid so," said Katherine, much entertained.

"But you're a very nice girl," said Miss Viner kindly. "And I'm sure
there's many a man might do worse than take you for a wife instead of
one of these flibbertigibbets running about nowadays showing more of
their legs than the Creator ever intended them to. Good-bye, my dear,
and I hope you'll enjoy yourself, but things are seldom what they seem
in this life."

Heartened by these prophecies, Katherine took her departure. Half the
village came to see her off at the station, including the little maid
of all work, Alice, who brought a stiff wired nosegay and cried openly.

"There ain't a many like her," sobbed Alice when the train had finally
departed. "I'm sure when Charlie went back on me with that girl from
the Dairy, nobody could have been kinder than Miss Grey was, and though
particular about the brasses and the dust, she was always one to notice
when you'd give a thing an extra rub. Cut myself in little pieces for
her, I would, any day. A real lady, that's what I call her."

Such was Katherine's departure from St. Mary Mead.




                    8. Lady Tamplin Writes a Letter


"Well," said Lady Tamplin, "well."

She laid down the continental _Daily Mail_ and stared out across the
blue waters of the Mediterranean. A branch of golden mimosa, hanging
just above her head, made an effective frame for a very charming
picture. A golden-haired, blue-eyed lady in a very becoming negligee.
That the golden hair owed something to art, as did the pink-and-white
complexion, was undeniable, but the blue of the eyes was Nature's gift,
and at forty-four Lady Tamplin could still rank as a beauty.

Charming as she looked, Lady Tamplin was, for once, not thinking of
herself. That is to say, she was not thinking of her appearance. She
was intent on graver matters.
```

… and in the end of the file:



vagrant@vagrant: /vagrant/hadoop-mapreduce-examples

```
($1 to $5,000) are particularly important to maintaining tax exempt
status with the IRS.

The Foundation is committed to complying with the laws regulating
charities and charitable donations in all 50 states of the United
States. Compliance requirements are not uniform and it takes a
considerable effort, much paperwork and many fees to meet and keep up
with these requirements. We do not solicit donations in locations
where we have not received written confirmation of compliance. To SEND
DONATIONS or determine the status of compliance for any particular state
visit www.gutenberg.org/donate.

While we cannot and do not solicit contributions from states where we
have not met the solicitation requirements, we know of no prohibition
against accepting unsolicited donations from donors in such states who
approach us with offers to donate.

International donations are gratefully accepted, but we cannot make
any statements concerning tax treatment of donations received from
outside the United States. U.S. laws alone swamp our small staff.

Please check the Project Gutenberg web pages for current donation
methods and addresses. Donations are accepted in a number of other
ways including checks, online payments and credit card donations. To
donate, please visit: www.gutenberg.org/donate.

Section 5. General Information About Project Gutenberg™ electronic works

Professor Michael S. Hart was the originator of the Project
Gutenberg™ concept of a library of electronic works that could be
freely shared with anyone. For forty years, he produced and
distributed Project Gutenberg™ eBooks with only a loose network of
volunteer support.

Project Gutenberg™ eBooks are often created from several printed
editions, all of which are confirmed as not protected by copyright in
the U.S. unless a copyright notice is included. Thus, we do not
necessarily keep eBooks in compliance with any particular paper
edition.

Most people start at our website which has the main PG search
facility: www.gutenberg.org.

This website includes information about Project Gutenberg™,
including how to make donations to the Project Gutenberg Literary
Archive Foundation, how to help produce our new eBooks, and how to
subscribe to our email newsletter to hear about new eBooks.
```

PART II

In this part, we implemented two java classes to implement the application. You can see them in the corresponding .java files, which include the code and some explanatory comments. The results are in the ‘output_partb.txt’.

Note that I have changed the name of the ‘wordcount’ to ‘danceability’, and in order to work I changed the name of the class, the package, the .java file and the name of the class in the pom.xml file. So for your convenience, I will be attaching .xml file too.

Figure 1: Types of dancing[1]

[1] Adapted from https://uk.harlequinfloors.com/en/news/popular-types-of-dance-list-of-top-dance-genres.