Problem1

```
hz2558@big-data-analysis:~$ ./hadoop/bin/hdfs dfs -ls /hbase
Found 1 items
drwxr-xr-x   - hz2558 supergroup          0 2018-09-17 03:10 /hbase/.tmp
hz2558@big-data-analysis:~$
```

```
hz2558@big-data-analysis:~$ spark-shell
2018-09-17 03:36:21 WARN  NativeCodeLoader:62 - Unable to load native-hadoop library for your platform... using bui
ltin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://big-data-analysis.c.shaped-terrain-216119.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1537155396424).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.3.1
      /_/

Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_181)
Type in expressions to have them evaluated.
Type :help for more information.

scala> :quit
hz2558@big-data-analysis:~$
```

```
hz2558@big-data-analysis:~$ hive

Logging initialized using configuration in jar:file:/home/hz2558/hive/lib/hive-common-1.2.2.jar!/hive-log4j.propert
ies
hive> exit;
hz2558@big-data-analysis:~$
```

Problem2

Using NYPD_Motor_Vehicle Dataset

a-c

gsutil ls

gsutil ls gs://happeng_storage/

gsutil cp gs://happeng_storage/text1.txt .

./bin/hdfs dfs -put ~/text1.txt /user/hz2558/ NYPD_Motor_Vehicle.csv

./bin/hdfs dfs -ls /user/hz2558

./bin/hdfs dfs -tail /user/hz2558/ NYPD_Motor_Vehicle.csv

```
hz2558@big-data-analysis:~/hadoop$ ./bin/hdfs dfs -put ./NYPD_Motor_Vehicle_Collisions.csv /user/hz2558/NYPD_Motor_
Vehicle_Collisions.csv
hz2558@big-data-analysis:~/hadoop$ ./bin/hdfs dfs -ls /user/hz2558
Found 1 items
-rw-r--r--   1 hz2558 supergroup  298261523 2018-09-17 19:34 /user/hz2558/NYPD_Motor_Vehicle_Collisions.csv
hz2558@big-data-analysis:~/hadoop$ ./bin/hdfs dfs -tail /user/hz2558/NYPD_Motor_Vehicle_Collisions.csv
,,,2896726,PASSENGER VEHICLE,UNKNOWN,,,
07/01/2012,9:50,MANHATTAN,10019,40.7667789,-73.9967618,"(40.7667789, -73.9967618)",WEST SIDE HIGHWAY
,WEST 50 STREET                  ,,0,0,0,0,0,0,0,0,Driver Inattention/Distraction,Unspecified,,,,37635,VAN,SPORT UT
ILITY / STATION WAGON,,,
07/01/2012,9:50,QUEENS,11001,40.7362448,-73.7029656,"(40.7362448, -73.7029656)",265 STREET                      ,85
 AVENUE                     ,,0,0,0,0,0,0,0,0,View Obstructed/Limited,Unspecified,,,,219395,PASSENGER VEHICLE,SPO
RT UTILITY / STATION WAGON,,,
07/01/2012,9:57,MANHATTAN,10065,40.7652424,-73.9578679,"(40.7652424, -73.9578679)",1 AVENUE
,EAST 68 STREET                  ,,0,0,0,0,0,0,0,0,Other Vehicular,Other Vehicular,,,,44907,PASSENGER VEHICLE,TAXI,
,,
07/01/2012,9:59,BRONX,10452,40.835397,-73.920305,"(40.835397, -73.920305)",EAST 167 STREET                  ,GERARD
AVENUE                   ,,0,0,0,0,0,0,0,0,Glare,Unspecified,,,,85154,PASSENGER VEHICLE,SPORT UTILITY / STATION WAG
ON,,,
```

d

gsutil ls

gsutil ls gs://happeng_storage/

gsutil cp gs://happeng_storage/text1.txt .

./bin/hdfs dfs -put ~/text1.txt /user/hz2558/text1.txt

./bin/hdfs dfs -ls /user/hz2558

./hadoop/bin/hadoop jar ./hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.1.jar wordcount /user/hz2558/text1.txt /user/hz2558/output

./bin/hdfs dfs -ls /user/hz2558/output
./bin/hdfs dfs -cat /user/hz2558/output/part-r-00000

Text1:

```
hz2558@big-data-analysis:~/hadoop$ ./bin/hdfs dfs -ls /user/hz2558/output
Found 2 items
-rw-r--r--   1 hz2558 supergroup          0 2018-09-17 20:19 /user/hz2558/output/_SUCCESS
-rw-r--r--   1 hz2558 supergroup       1477 2018-09-17 20:19 /user/hz2558/output/part-r-00000
hz2558@big-data-analysis:~/hadoop$ ./bin/hdfs dfs -cat /user/hz2558/output/part-r-00000
Abilities       1
Balls   1
Betrayed        1
Boy     1
Cottage 1
Dare    1
Disposal        1
Entire  1
Had     1
He      1
Its     1
Known   1
Lose    1
Match   1
Meant   1
Mrs     1
No      1
Old     1
Preserved       1
Received        1
Remain  1
She     1
So      1
Travelling      1
True    1
Two     1
Winter  1
Yet     1
adapted 1
an      1
an.     1
any     1
appetite        1
as      2
at      1
at.     2
be      2
body.   1
boy     1
branched.       1
but     1
by.     1
can     1
change  1
charm   1
charmed 1
child   1
comfort 1
conviction      1
cordially       1
dashwood        1
decay   1
delay   1
demands.        1
```

```
design  1
desire  1
discretion      2
diverted        2
do      1
doors   1
downs   1
drew.   1
edward. 1
effect  1
end     1
estimating      1
event   1
exposed 1
extremely       1
eyes    1
fact    1
fanny   1
fat     2
fat.    1
few.    1
folly   1
forfeited       1
formerly.       1
forming 1
friendship      1
front   1
get     1
gone    1
had     2
hardly  1
has     1
he      1
her     1
him.    1
impossible      1
improve 1
in.     1
income  1
incommode       1
increasing      1
indeed  1
inquietude      1
instrument      1
invited.        1
is      2
it      2
its.    1
just    1
keeps   1
kindness        1
large   1
learning        1
least   1
led     1
letter  1
lively  1
looking 1
man     2
may     1
```

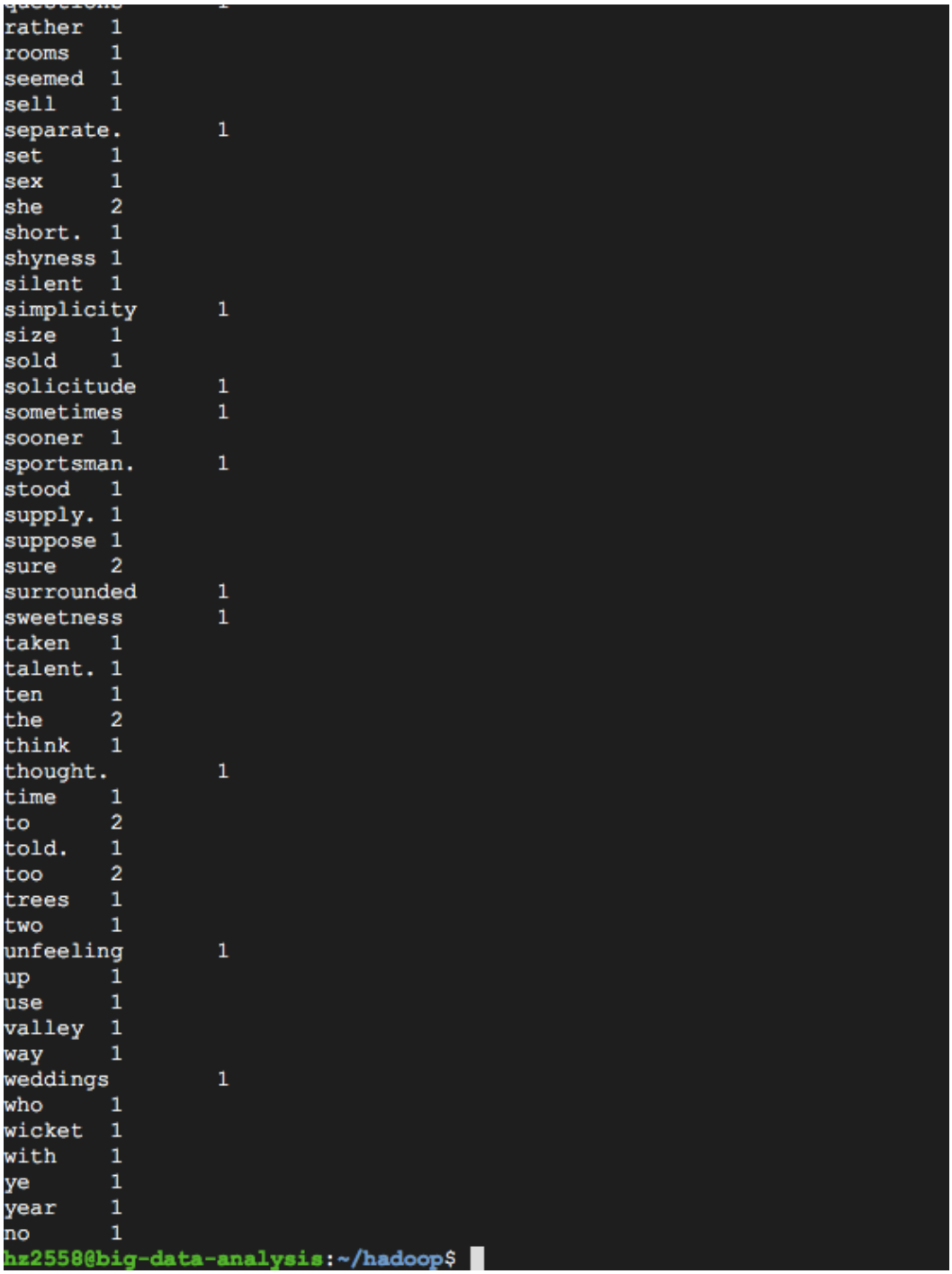
```
means    1
money    1
most     1
mr       2
my       2
name     1
needed   1
noisier  1
now      1
now.     1
of       1
off      3
on       1
on.      1
one      1
oppose   1
or.      1
picture  1
point    1
prepared         1
principles       1
produced         1
promise  1
put      1
quick    1
quick.   1
raptures.        1
resembled.       1
returned         1
round    1
sentiments       1
share    1
she      2
shew.    1
shy      1
situation        1
so       2
solicitude       1
spot     1
state    1
strictly         1
style    1
ten      1
tended   1
the.     1
there    2
those    1
three.   1
to       5
to.      1
two      1
understood       1
unsatiable       1
vexed    1
views.   1
voice    1
way      3
we       1
weeks    1
```

words: to, off, way,

Text2:

```
Add     1
Added   1
Ask     1
Chapter 1
Course  1
Dear    1
End     1
Everything      1
Excellent       1
Extensive       1
Full    1
Greatly 1
He      1
Less    1
Man     1
Or      1
Projection      1
Rent    1
Son     1
Tastes  1
Tolerably       1
Week    1
Wicket  1
Years   1
abode   1
affixed 1
age     1
agreeable       1
all     1
and     1
as      1
ask.    1
assurance       1
at      2
boy     2
breakfast.      1
but     3
chief   1
child   1
compact 1
concluded       1
connection      1
cordial 1
county  1
decisively.     1
dependent       1
devonshire      1
did     1
direct  1
discovery       2
discretion      1
distrusts       1
do      2
door.   1
earnestly       1
elegance        1
enable  1
enquire 1
extremely       1
extremity       1
```

```
feel    1
few     1
fine    1
formal  1
formed  1
forth   1
frequently      1
friendship      1
gentleman.      1
giving  1
great   1
he      2
her     1
how     1
if      1
immediate       1
in      3
in.     2
insensible      1
it      1
its     1
john.   1
led.    1
literature      1
loud.   1
manner  1
matter  1
matters 1
may.    1
me      1
mention 1
middleton       1
more    1
motionless      1
mr      1
my.     1
neglected       1
no      2
noisy   1
not.    1
now     1
of      1
offered 1
often   1
oh      1
on.     1
oppose  1
or      1
others  1
our     1
own     1
parlors 1
passed  1
pianoforte      1
place   1
prepare 2
projecting.     1
promise.        1
put.    1
questions       1
```



words: in, but, no

Problem3

start-hbase.sh
hbase shell
create ‘LocationID’,’ Borough,Zone,service_zone’
exit

gsutil cp gs://happeng_storage/test.csv .
./bin/hdfs dfs -put ./test.csv /user/hz2558/taxi.csv

./bin/hdfs dfs -ls /user/hz2558
hbase shell
scan ‘name’

hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.separator="," -Dimporttsv.columns="HBASE_ROW_KEY,Borough,Zone,service_zone" LocationID hdfs://localhost:1234/user/hz2558/taxi.csv
hbase shell
scan ‘LocationID’,{LIMIT => 10}

```
ROW                           COLUMN+CELL
 "LocationID"                 column=Borough:, timestamp=1537290750935, value="Borough"
 "LocationID"                 column=Zone:, timestamp=1537290750935, value="Zone"
 "LocationID"                 column=service_zone:, timestamp=1537290750935, value="service_zone"
 1                            column=Borough:, timestamp=1537290750935, value="EWR"
 1                            column=Zone:, timestamp=1537290750935, value="Newark Airport"
 1                            column=service_zone:, timestamp=1537290750935, value="EWR"
 10                           column=Borough:, timestamp=1537290750935, value="Queens"
 10                           column=Zone:, timestamp=1537290750935, value="Baisley Park"
 10                           column=service_zone:, timestamp=1537290750935, value="Boro Zone"
 100                          column=Borough:, timestamp=1537290750935, value="Manhattan"
 100                          column=Zone:, timestamp=1537290750935, value="Garment District"
 100                          column=service_zone:, timestamp=1537290750935, value="Yellow Zone"
 101                          column=Borough:, timestamp=1537290750935, value="Queens"
 101                          column=Zone:, timestamp=1537290750935, value="Glen Oaks"
 101                          column=service_zone:, timestamp=1537290750935, value="Boro Zone"
 102                          column=Borough:, timestamp=1537290750935, value="Queens"
 102                          column=Zone:, timestamp=1537290750935, value="Glendale"
 102                          column=service_zone:, timestamp=1537290750935, value="Boro Zone"
 103                          column=Borough:, timestamp=1537290750935, value="Manhattan"
 103                          column=Zone:, timestamp=1537290750935, value="Governor's Island/Ellis Island/Liberty
                              Island"
 103                          column=service_zone:, timestamp=1537290750935, value="Yellow Zone"
 104                          column=Borough:, timestamp=1537290750935, value="Manhattan"
 104                          column=Zone:, timestamp=1537290750935, value="Governor's Island/Ellis Island/Liberty
                              Island"
 104                          column=service_zone:, timestamp=1537290750935, value="Yellow Zone"
 105                          column=Borough:, timestamp=1537290750935, value="Manhattan"
 105                          column=Zone:, timestamp=1537290750935, value="Governor's Island/Ellis Island/Liberty
                              Island"
 105                          column=service_zone:, timestamp=1537290750935, value="Yellow Zone"
 106                          column=Borough:, timestamp=1537290750935, value="Brooklyn"
 106                          column=Zone:, timestamp=1537290750935, value="Gowanus"
 106                          column=service_zone:, timestamp=1537290750935, value="Boro Zone"
10 row(s) in 0.8010 seconds
```

scan 'LocationID',{LIMIT => 10,COLUMNS => ['Borough']}

```
hbase(main):007:0> scan 'LocationID',{LIMIT => 10,COLUMNS => ['Borough']}
ROW                           COLUMN+CELL
 "LocationID"                 column=Borough:, timestamp=1537290750935, value="Borough"
 1                            column=Borough:, timestamp=1537290750935, value="EWR"
 10                           column=Borough:, timestamp=1537290750935, value="Queens"
 100                          column=Borough:, timestamp=1537290750935, value="Manhattan"
 101                          column=Borough:, timestamp=1537290750935, value="Queens"
 102                          column=Borough:, timestamp=1537290750935, value="Queens"
 103                          column=Borough:, timestamp=1537290750935, value="Manhattan"
 104                          column=Borough:, timestamp=1537290750935, value="Manhattan"
 105                          column=Borough:, timestamp=1537290750935, value="Manhattan"
 106                          column=Borough:, timestamp=1537290750935, value="Brooklyn"
10 row(s) in 0.0310 seconds
```

Problem4
hive

create table taxi (LocationID int, Borough string, Zone string, service_zone string) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

LOAD DATA INPATH 'hdfs://localhost:1234/user/hz2558/taxi.csv' INTO TABLE taxi;

```
hive> select * from taxi limit 5;
OK
NULL    "Borough"       "Zone"  "service_zone"
1       "EWR"   "Newark Airport"        "EWR"
2       "Queens"        "Jamaica Bay"   "Boro Zone"
3       "Bronx" "Allerton/Pelham Gardens"       "Boro Zone"
4       "Manhattan"     "Alphabet City" "Yellow Zone"
Time taken: 0.119 seconds, Fetched: 5 row(s)
```

```
hive> select Borough as place, Zone as service from taxi where LocationID < 10;
OK
"EWR"   "Newark Airport"
"Queens"        "Jamaica Bay"
"Bronx" "Allerton/Pelham Gardens"
"Manhattan"     "Alphabet City"
"Staten Island" "Arden Heights"
"Staten Island" "Arrochar/Fort Wadsworth"
"Queens"        "Astoria"
"Queens"        "Astoria Park"
"Queens"        "Auburndale"
Time taken: 0.122 seconds, Fetched: 9 row(s)
```

```
hive> select Borough from taxi where LocationID > 250;
OK
"Staten Island"
"Queens"
"Queens"
"Bronx"
"Brooklyn"
"Brooklyn"
"Brooklyn"
"Queens"
"Bronx"
"Queens"
"Manhattan"
"Manhattan"
"Manhattan"
"Unknown"
"Unknown"
Time taken: 0.111 seconds, Fetched: 15 row(s)
```

```
hive> select * from taxi where LocationID < 10;
OK
1       "EWR"   "Newark Airport"        "EWR"
2       "Queens"        "Jamaica Bay"   "Boro Zone"
3       "Bronx" "Allerton/Pelham Gardens"       "Boro Zone"
4       "Manhattan"     "Alphabet City" "Yellow Zone"
5       "Staten Island" "Arden Heights" "Boro Zone"
6       "Staten Island" "Arrochar/Fort Wadsworth"       "Boro Zone"
7       "Queens"        "Astoria"       "Boro Zone"
8       "Queens"        "Astoria Park"  "Boro Zone"
9       "Queens"        "Auburndale"    "Boro Zone"
Time taken: 0.569 seconds, Fetched: 9 row(s)
```

```
hive> select Zone as zone,service_zone as service from taxi where LocationID < 5;
OK
"Newark Airport"        "EWR"
"Jamaica Bay"   "Boro Zone"
"Allerton/Pelham Gardens"       "Boro Zone"
"Alphabet City" "Yellow Zone"
Time taken: 0.095 seconds, Fetched: 4 row(s)
```

Problem5:

Python Code:

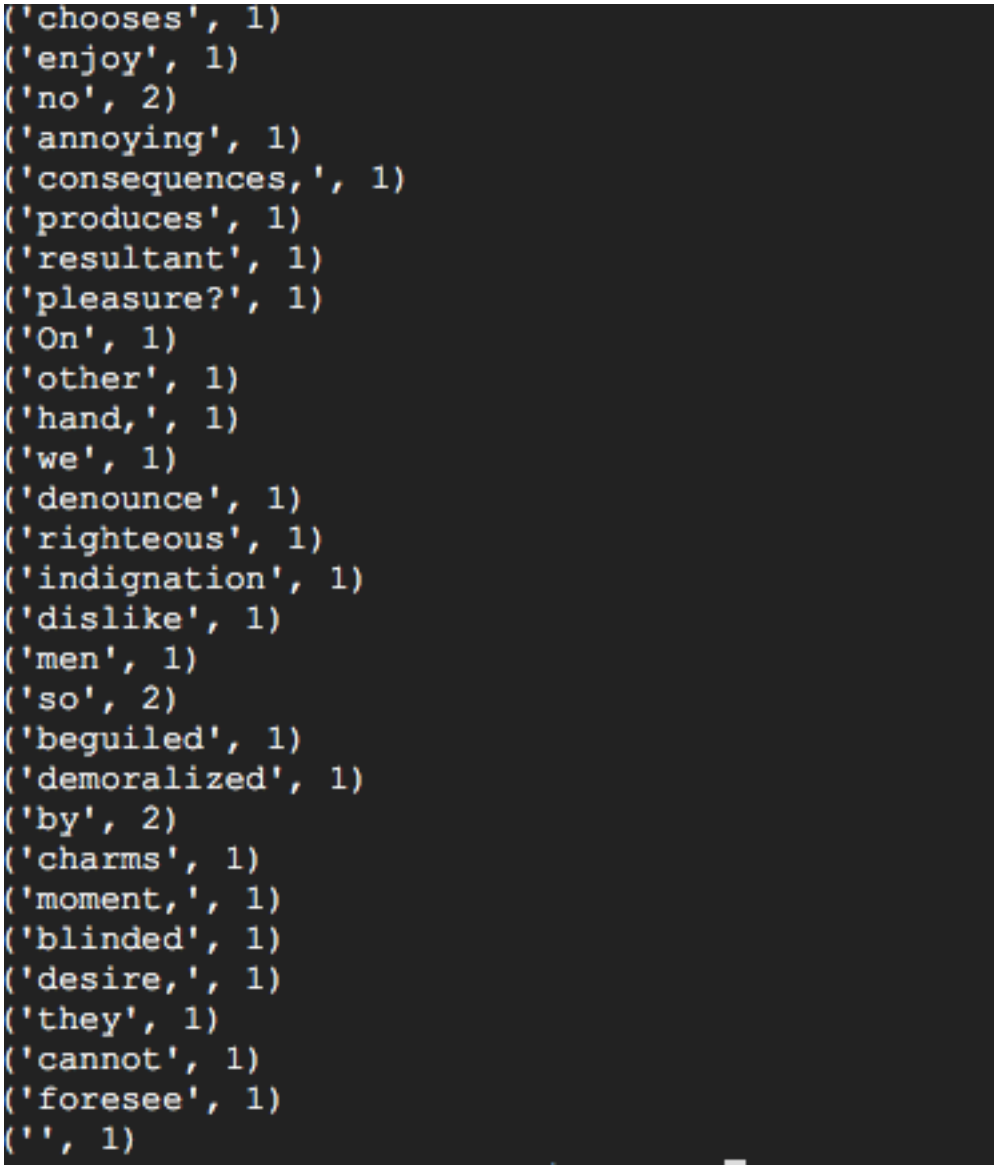
```
from pyspark import SparkConf, SparkContext
import random
conf = (SparkConf().setMaster("local").setAppName("My app").set("spark.executor.memory", "1g"))
sc = SparkContext(conf = conf)

text_file = sc.textFile("hdfs://localhost:1234/user/hz2558/text.txt")
counts = text_file.flatMap(lambda line: line.split(" ")).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
counts.saveAsTextFile("hdfs://localhost:1234/user/hz2558/out.txt")
```

```
hz2558@big-data-analysis:~/hadoop$ ./bin/hdfs dfs -cat out.txt/part-00000
('But', 2)
('I', 2)
('must', 1)
('explain', 1)
('to', 6)
('you', 2)
('how', 2)
('all', 1)
('this', 1)
('mistaken', 1)
('idea', 1)
('of', 9)
('denouncing', 1)
('pleasure', 5)
('and', 6)
('praising', 1)
('pain', 4)
('was', 1)
('born', 1)
('will', 1)
('give', 1)
('a', 5)
('complete', 1)
('account', 1)
('the', 8)
('system,', 1)
('expound', 1)
('actual', 1)
('teachings', 1)
('great', 2)
('explorer', 1)
('truth,', 1)
('master-builder', 1)
```

```
('human', 1)
('happiness.', 1)
('No', 1)
('one', 2)
('rejects,', 1)
('dislikes,', 1)
('or', 4)
('avoids', 2)
('itself,', 2)
('because', 4)
('it', 2)
('is', 3)
('pleasure,', 1)
('but', 2)
('those', 1)
('who', 6)
('do', 1)
('not', 1)
('know', 1)
('pursue', 1)
('rationally', 1)
('encounter', 1)
('consequences', 1)
('that', 4)
('are', 2)
('extremely', 1)
('painful.', 1)
('Nor', 1)
('again', 1)
('there', 1)
('anyone', 1)
('loves', 1)
('pursues', 1)
('desires', 1)
```

```
('obtain', 2)
('pain,', 1)
('occasionally', 1)
('circumstances', 1)
('occur', 1)
('in', 1)
('which', 2)
('toil', 1)
('can', 1)
('procure', 1)
('him', 1)
('some', 2)
('pleasure.', 1)
('To', 1)
('take', 1)
('trivial', 1)
('example,', 1)
('us', 1)
('ever', 1)
('undertakes', 1)
('laborious', 1)
('physical', 1)
('exercise,', 1)
('except', 1)
('advantage', 1)
('from', 1)
('it?', 1)
('has', 2)
('any', 1)
('right', 1)
('find', 1)
('fault', 1)
('with', 2)
('man', 1)
```



words: of, the, to