

Name:-Akshata Nagesh Kundar

Roll no:540

Subject:- Business Intelligence and Big Data

MSc-CS Part1

Practical 2-Implementation Map-Reduce Program for Word Count Problem.

Step1



Step 2

Home x cloudera-quickstart-vm-5... x

Applications Places System

Sun Mar 19, 9:04 PM clou

Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox

Cloudera Live : Welcom... x


quickstart.cloudera/#/

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

cloudera LIVE Navigation

Welcome to Your Cloudera QuickStart VM!


Your Cluster	
Node	Address
Manager Node	192.168.91.129
Worker Node 1	192.168.91.129



Get Started

The tutorial below guides you through some analytic use cases, using the most popular open source tools included with CDH (including Cloudera Impala, Cloudera Search, and Hue).

[Start Tutorial!](#)



Analyze Your Data

house pointer inside or press Ctrl+G.

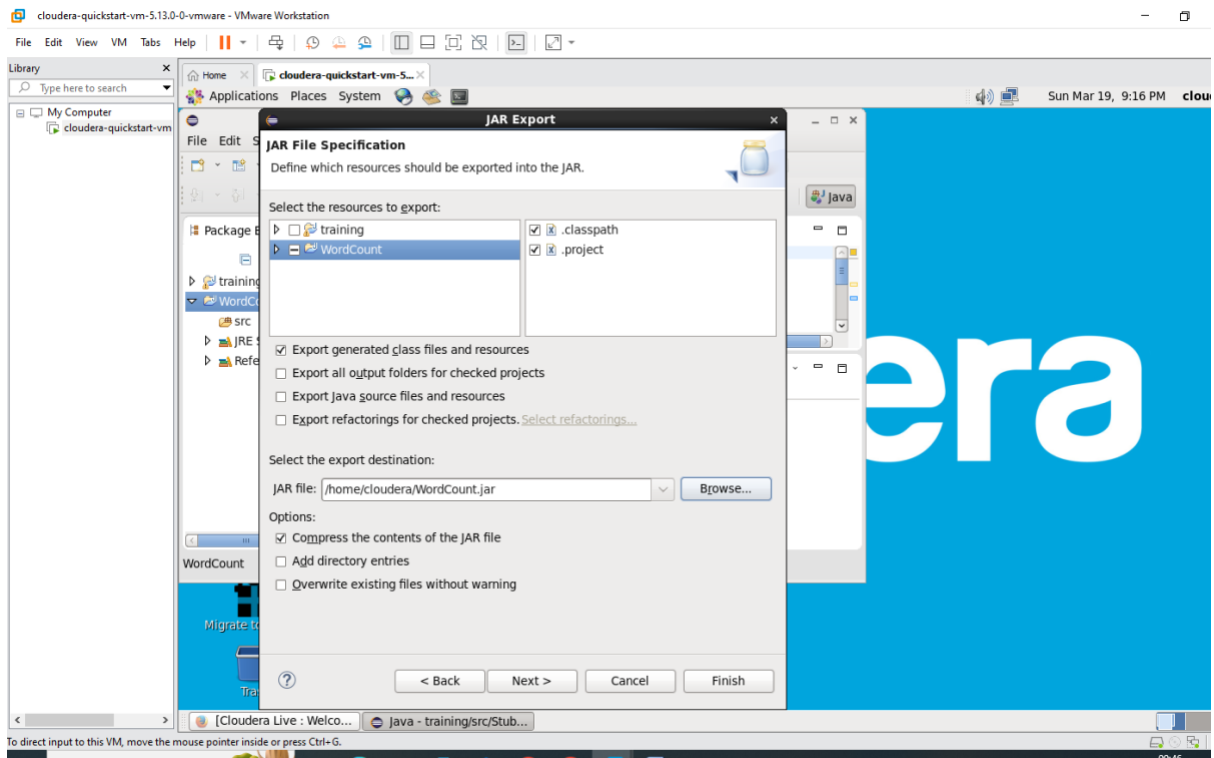
ch

29°C Smoke 09:34 20-03-2023

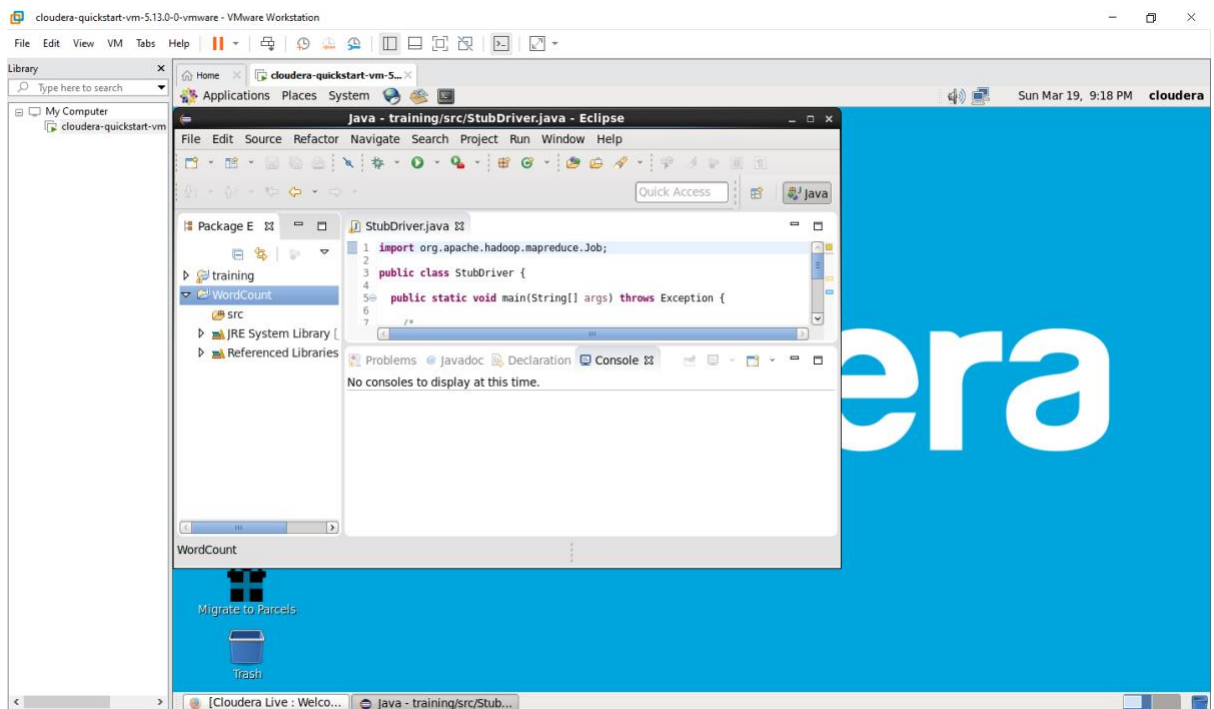
Step 3



Step 4



Step 5



Step 6

```
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx - hdfs supergroup          0 2017-10-23 10:29 /benchmarks
drwxr-xr-x - hbase supergroup          0 2023-03-19 20:39 /hbase
drwxr-xr-x - solr solr                  0 2017-10-23 10:32 /solr
drwxrwxrwt - hdfs supergroup          0 2023-03-17 00:31 /tmp
drwxr-xr-x - hdfs supergroup          0 2017-10-23 10:31 /user
drwxr-xr-x - hdfs supergroup          0 2017-10-23 10:31 /var
[cloudera@quickstart Desktop]$ sudo -u hdfs hadoop fs -mkdir /inputdirectory
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 7 items
drwxrwxrwx - hdfs supergroup          0 2017-10-23 10:29 /benchmarks
drwxr-xr-x - hbase supergroup          0 2023-03-19 20:39 /hbase
drwxr-xr-x - hdfs supergroup          0 2023-03-19 21:24 /inputdirectory
drwxr-xr-x - solr solr                  0 2017-10-23 10:32 /solr
drwxrwxrwt - hdfs supergroup          0 2023-03-17 00:31 /tmp
drwxr-xr-x - hdfs supergroup          0 2017-10-23 10:31 /user
drwxr-xr-x - hdfs supergroup          0 2017-10-23 10:31 /var
```

Step 7

```
[cloudera@quickstart Desktop]$ cat > /home/cloudera/ProcessFile.txt
helloworld
good morning
have a good day^C
[cloudera@quickstart Desktop]$ cat /home/cloudera/ProcessFile.txt
helloworld
good morning
have a good day[cloudera@quickstart Desktop]$ sudo -u hdfs hadoop fs -chmod -R fs -chmod -R
-chmod: Not enough arguments: expected 2 but got 0
Usage: hadoop fs [generic options] -chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...
[cloudera@quickstart Desktop]$ ^C
[cloudera@quickstart Desktop]$ sudo -u hdfs hadoop fs -chmod -R 777 /inputdirectory
[cloudera@quickstart Desktop]$ sudo -u hdfs hadoop fs -put /home/cloudera/ProcessFile.txt/inputdirectory
put: '.': No such file or directory
[cloudera@quickstart Desktop]$ sudo -u hdfs hadoop fs -put /home/cloudera/ProcessFile.txt /inputdirectory
[cloudera@quickstart Desktop]$ hdfs dfs -ls /inputdirectory
Found 1 items
-rw-r--r-- 1 hdfs supergroup          39 2023-03-19 21:35 /inputdirectory/ProcessFile.txt
```

Step 8

```

Map input records=1
Map output records=8
Map output bytes=60
Map output materialized bytes=72
Input split bytes=127
Combine input records=8
Combine output records=7
Reduce input groups=7
Reduce shuffle bytes=72
Reduce input records=7
Reduce output records=7
Spilled Records=14
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=791
CPU time spent (ms)=3270
Physical memory (bytes)
snapshot=459796480
Virtual memory (bytes)
snapshot=3146891264
Total committed heap usage
(bytes)=389021696
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=28
File Output Format Counters
Bytes Written=38
[cloudera@quickstart ~]$ hdfs dfs -ls /out1
Found 2 items
-rw-r--r--  1 cloudera supergroup          0
2023-01-05 23:02 /out1/_SUCCESS
-rw-r--r--  1 cloudera supergroup        38
2023-01-05 23:02 /out1/part-r-00000
[cloudera@quickstart ~]$ hdfs dfs -cat
/out1/part -r-00000
cat: `/out1/part': No such file or directory
cat: `-r-00000': No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -cat
/out1/part-r-00000
Hii 2
How 1
am 1
are 1
fine 1
i 1
u 1
[cloudera@quickstart ~]$

```

Practical-3:- Write a Pig Script For Solving counting Problems.

Steps :

```
cat> /home/cloudera/input.csv
```

```
cat /home/cloudera/input.csv
```

```
pig -x local
```

```
lines = load '/home/cloudera/input.csv' as (line:chararray);
```

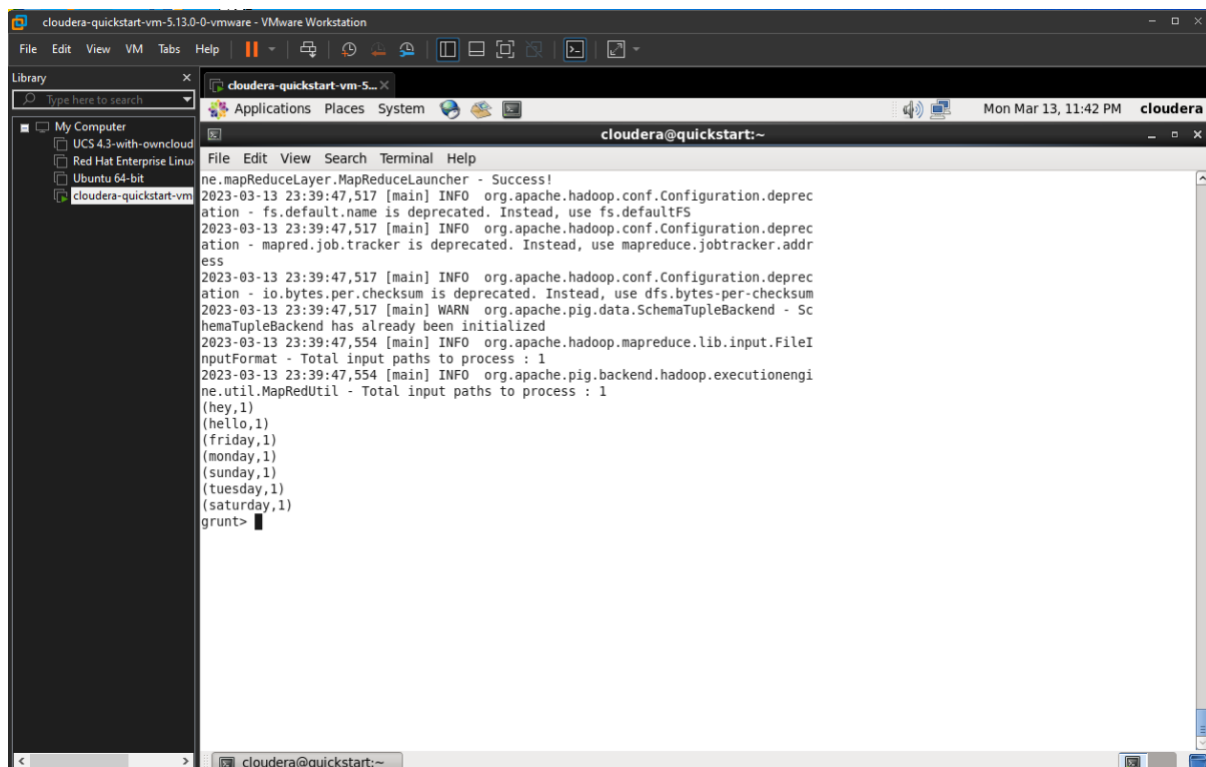
```
words = foreach lines GENERATE FLATTEN(TOKENIZE(line)) as word;
```

```
grouped = GROUP words by word;
```

```
wordcount = foreach grouped GENERATE group, COUNT(words);
```

```
dump wordcount;
```

output:



```
cloudera-quickstart-vm-5.13.0-0-vmware - VMware Workstation
File Edit View VM Tabs Help
Library
Type here to search
My Computer
UCS 4.3-with-owncloud
Red Hat Enterprise Linux
Ubuntu 64-bit
cloudera-quickstart-vm
cloudera-quickstart-vm-5...
Applications Places System
cloudera@quickstart:~
File Edit View Search Terminal Help
ne.mapReduceLayer.MapReduceLauncher - Success!
2023-03-13 23:39:47,517 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2023-03-13 23:39:47,517 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2023-03-13 23:39:47,517 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-03-13 23:39:47,517 [main] WARN org.apache.pig.data.SchemaTupleBackend - Sc
emaTupleBackend has already been initialized
2023-03-13 23:39:47,554 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2023-03-13 23:39:47,554 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(hey,1)
(hello,1)
(friday,1)
(monday,1)
(sunday,1)
(tuesday,1)
(saturday,1)
grunt>
```

Practical-4:-Install HBase and use the HBase Data model Store and retrieve Data

Steps :

```
//Start HBase
```

```
hbase shell
```

```
//HBase Commands
```

```
status
```

```
version,
```

table_help

whoami

//Data Definition Language

create 'employee', 'Name', 'ID', 'Designation', 'Salary', 'Department'

//Verify created table

list

//Disable single table

disable 'employee'

scan 'employee'

//or

is_disable 'employee'

//Disable multiple tables

disable_all 'e.*'

// Enabling table

enable 'employee'

//Or

is_enabled 'employee'

//create new table

create 'student', 'name', 'age', 'course'

```
put 'student', 'sharath', 'name:fullname', 'sharathkumar'
```

```
put 'student', 'sharath', 'age:presentage', '24'
```

```
put 'student', 'sharath', 'course:pursuing', 'Hadoop'
```

```
put 'student', 'shashank', 'name:fullname', 'shashank R'
```

```
put 'student', 'shashank', 'age:presentage', '23'
```

```
put 'student', 'shashank', 'course:pursuing', 'Java'
```

```
//Get Information
```

```
get 'student', 'shashank'
```

```
get 'student', 'sharath'
```

```
get 'student', 'sharath', 'course'
```

```
get 'student', 'shashank', 'course'
```

```
get 'student', 'sharath', 'name'
```

```
//Scan
```

```
scan 'student'
```

```
//Count
```

```
Count 'student'
```

```
//Alter
```

```
alter 'student', NAME=>'name', VERSIONS=>5
```

```
put 'student', 'shashank', 'name:fullname', 'shashank Rao'
```

```
scan 'student'
```

```
//Delete
```

```
delete 'student', 'shashank', 'name:fullname'
```


Practical-5:-Install Hive and use Hive Create Store Structured databases.

Steps :

```
cat > /home/cloudera/employee.txt
```

```
1~Sachin~Pune~Product Engineering~100000~Big Data
```

```
2~Gaurav~Bangalore~Sales~90000~CRM
```

```
3~Manish~Chennai~Recruiter~125000~HR
```

```
4~Bhushan~Hyderabad~Developer~50000~BFSI
```

```
cat /home/cloudera/employee.txt
```

```
sudo -u hdfs hadoop fs -put /home/cloudera/employee.txt /inputdirectroy
```

```
hdfs dfs -ls /
```

```
hdfs dfs -ls /inputdirectory
```

```
hadoop fs -cat /inputdirectory/employee.txt
```

```
hive
```

```
show databases;
```

```
create database organization;
```

```
show databases;
```

```
use organization;
```

```
show tables;
```

```
hive> create table employee(
```

```
> id int,
```

```
> name string,
```

```
> city string,
```

```
> department string,
```

```
> salary int,
```

```
> domain string)
```

```
> row format delimited
```

```
> fields terminated by '~';
```

```
show tables;
```

```
select * from employee;
```

```
show tables;
```

```
load data inpath '/inputdirectory/employee.txt' overwrite into table employee;
```

```
show tables;
```

```
select * from employee;
```

Practical-6:-Write a program to construct different types of k-shingles for a given document.

```
install.packages("tm")
```

```
require("tm")
```

```
install.packages("devtools")
```

```
readinteger <-function()
```

```
{
```

```
  n<-readline(prompt="Enter value of k-1:")
```

```
  k<- as.integer(n)
```

```
  u1<- readLines("C:/MSC Notes/file.txt")
```

```
  Shingle <-0
```

```
  i<-0
```

```
  while(i<nchar(u1)-k+1){
```

```
    Shingle[i] <- substr(u1,start=i,stop=i+k)
```

```
    print(Shingle[i])
```

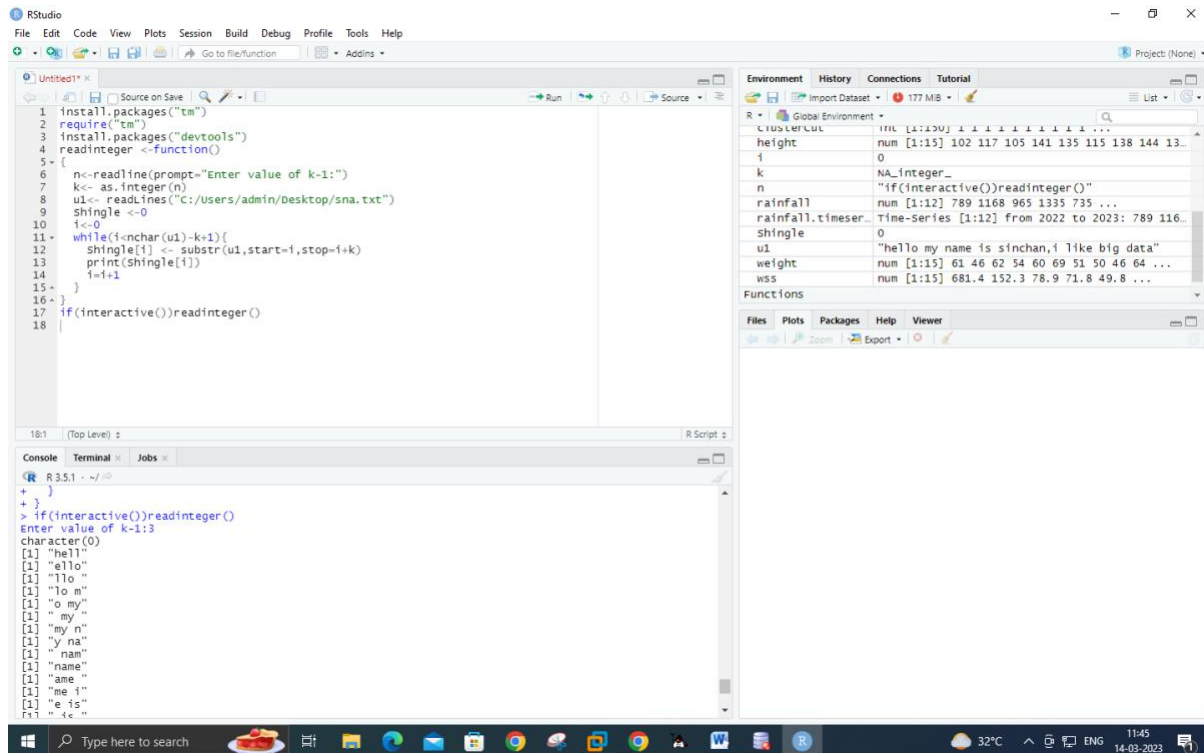
```
    i=i+1
```

```
  }
```

```
}
```

```
if(interactive())readinteger()
```

output:



The screenshot shows the RStudio interface. The script editor on the left contains the following code:

```
1 install.packages("tm")
2 require("tm")
3 install.packages("devtools")
4 readinteger <-function()
5 {
6   n<-readline(prompt="Enter value of k-1:")
7   k<- as.integer(n)
8   u1<- readlines("C:/Users/admin/Desktop/sna.txt")
9   shingle <-0
10  f<-0
11  while(f<nchar(u1)-k+1){
12    shingle[f] <- substr(u1,start=f,stop=f+k)
13    print(shingle[f])
14    f=f+1
15  }
16 }
17 if(interactive())readinteger()
18
```

The console on the bottom left shows the execution of the script, with the prompt 'Enter value of k-1:3' and the output of the shingle function:

```
R 3.5.1 ~\> if(interactive())readinteger()
Enter value of k-1:3
character(0)
[1] "hell"
[1] "ello"
[1] "llo "
[1] "lo m"
[1] "o my"
[1] "my "
[1] "my n"
[1] "y na"
[1] "nam"
[1] "name"
[1] "ame "
[1] "me i"
[1] "e is"
[1] "ic"
```

The environment pane on the right shows the objects in the global environment:

Object	Class	Attributes
clusterLUC	tmL	[1:130] 1 1 1 1 1 1 1 1 1 1 ...
height	num	[1:15] 102 117 105 141 135 115 138 144 13...
f	0	
k	NA_integer_	
n	"if(interactive())readinteger()"	
rainfall	num	[1:12] 789 1168 965 1335 735 ...
rainfall.timeser	Time-Series	[1:12] from 2022 to 2023: 789 116...
shingle	0	
u1	"hello my name is sinchan,i like big data"	
weight	num	[1:15] 61 46 62 54 60 69 51 50 46 64 ...
wss	num	[1:15] 681.4 152.3 78.9 71.8 49.8 ...

Practical-7:- Write a program for measuring similarity among documents and detecting passages which have been reused.

```
install.packages("tm")
```

```
require("tm")
```

```
install.packages("devtools")
```

```
my.corpus <- Corpus(DirSource("C:/MSC Notes/r-corpus"))
```

```
my.corpus<- tm_map(my.corpus, removeWords ,stopwords("english"))
```

```
my.tdm<- TermDocumentMatrix(my.corpus)
```

```
my.dtm<- DocumentTermMatrix(my.corpus,control=list(weighting= weightTfIdf
,stopwords=TRUE))
```

```
my.df<- as.data.frame(inspect(my.tdm))
```

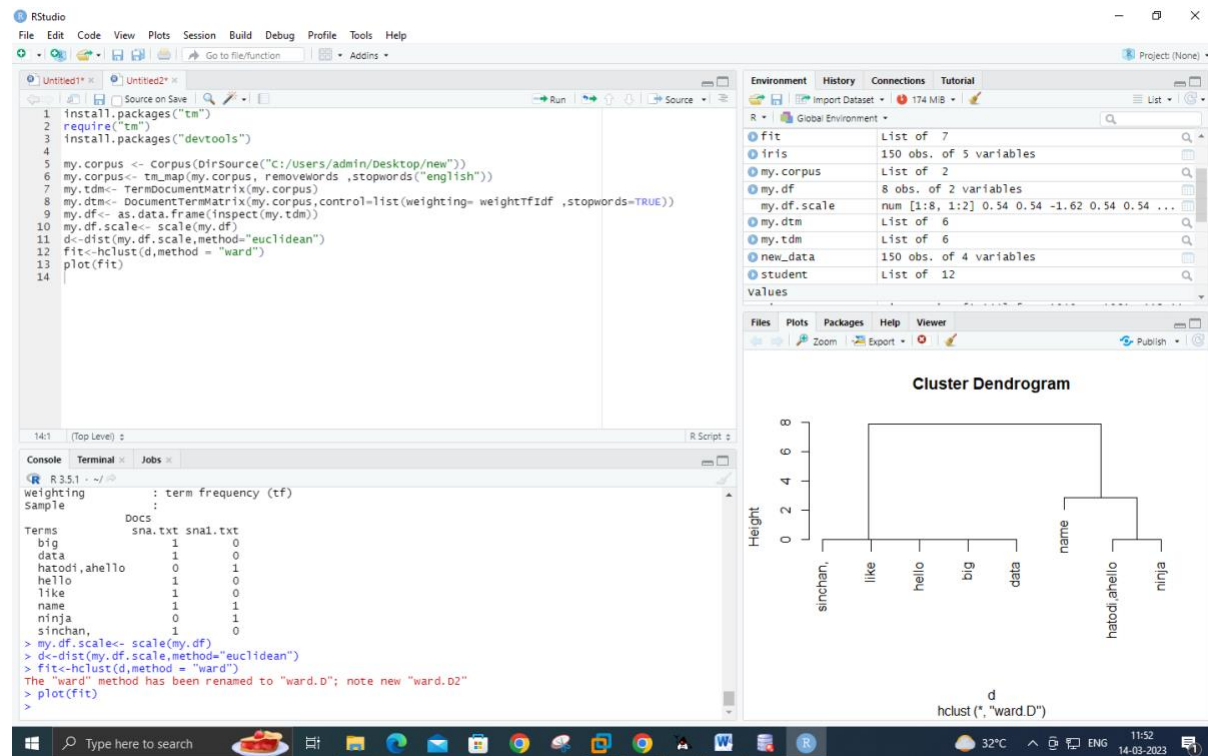
```
my.df.scale<- scale(my.df)
```

```
d<-dist(my.df.scale,method="euclidean")
```

```
fit<-hclust(d,method = "ward")
```

```
plot(fit)
```

output:



Practical-8:-Write a program to compute the n-moment for a given stream where n is given.

```
import java.io.*;
```

```
import java.util.*;
```

```
public class n_moment
```

```
{
```

```
    public static void main(String args[]) {
```

```
        int n=15;
```

```
        String stream[]= {"a","b","c","b","d","a","c","d","a","b","d","c","a","a","b"};
```

```
        int zero_moment=0,first_moment=0,second_moment=0,count=1,flag=0;
```

```
        ArrayList<Integer> arrlist=new ArrayList();
```

```
        System.out.println("Arraylist elements are::");
```

```

for (int i=0;i<15;i++)
{
    System.out.println(stream[i]+" ");
}

Arrays.sort(stream);
for(int i=1;i<n;i++)
{
    if(stream[i]==stream[i-1])
    {
        count++;

    }
    else
    {
        //System.out.println("Hello"+i);
        arrlist.add(count);
        count=1;
    }
}

arrlist.add(count);

zero_moment=arrlist.size();

System.out.println("\n\n\nValue of Zeroth moment for given
stream::"+zero_moment);

for(int i=0;i<arrlist.size();i++)
{

```

```
        first_moment+=arrlist.get(i);
    }

    System.out.println("\n\nValue of First moment for given
stream::"+first_moment);

    for (int i=0;i<arrlist.size();i++)
    {
        int j=arrlist.get(i);

        second_moment+=(j*j);
    }

    System.out.println("\n\nValue of Second moment for given
stream::"+second_moment);

    }

}
```

Output:

```
C:\WINDOWS\system32\cmd.exe
ArrayList elements are::
a
b
c
c
b
d
a
c
d
a
b
d
c
a
a
b

Value of Zeroth moment for given stream::4
Value of First moment for given stream::15
Value of Second moment for given stream::59
C:\Users\admin\Desktop>Pause
Press any key to continue . . .
```

Practical-9:-Write a program to demonstrate the ALON-Matias-Szegedy Algorithm for second moments.

```
import java.io.*;
```

```
import java.util.*;
```

```
class AMSA
```

```
{
```

```
    public static int findCharCount(String stream,char XE,int random,int n)
```

```
    {
```

```
        int countoccurance=0;
```

```
        for(int i=random;i<n;i++)
```

```
        {
```

```
            if(stream.charAt(i)==XE)
```

```
            {
```

```
                countoccurance++;
```

```
            }
```

```

    }

    return countoccurance;
}

public static int estimateValue(int XV1,int n)
{
    int ExpValue;

    ExpValue=n*(2*XV1-1);

    return ExpValue;
}

public static void main(String args[])
{
    int n=15;

    String stream="abcbdacdabdcaab";

    int random1=3,random2=8,random3=13;

    char XE1,XE2,XE3;

    int XV1,XV2,XV3;

    int ExpValuXE1,ExpValuXE2,ExpValuXE3;

    int apprSecondMomentValue;

    XE1=stream.charAt(random1-1);

    XE2=stream.charAt(random2-1);

    XE3=stream.charAt(random3-1);

    XV1=findCharCount(stream,XE1,random1-1,n);

    XV2=findCharCount(stream,XE2,random2-1,n);

    XV3=findCharCount(stream,XE3,random3-1,n);

    System.out.println(XE1+"="+XV1+" "+XE2+"="+XV2+" "+XE3+"="+XV3);

    ExpValuXE1=estimateValue(XV1,n);

```



```

ExpValuXE2=estimateValue(XV2,n);

ExpValuXE3=estimateValue(XV3,n);

System.out.println("Expected value for"+XE1+" is::"+ExpValuXE1);

System.out.println("Expected value for"+XE2+" is::"+ExpValuXE2);

System.out.println("Expected value for"+XE3+" is::"+ExpValuXE3);

apprSecondMomentValue=(ExpValuXE1+ExpValuXE2+ExpValuXE3)/3;

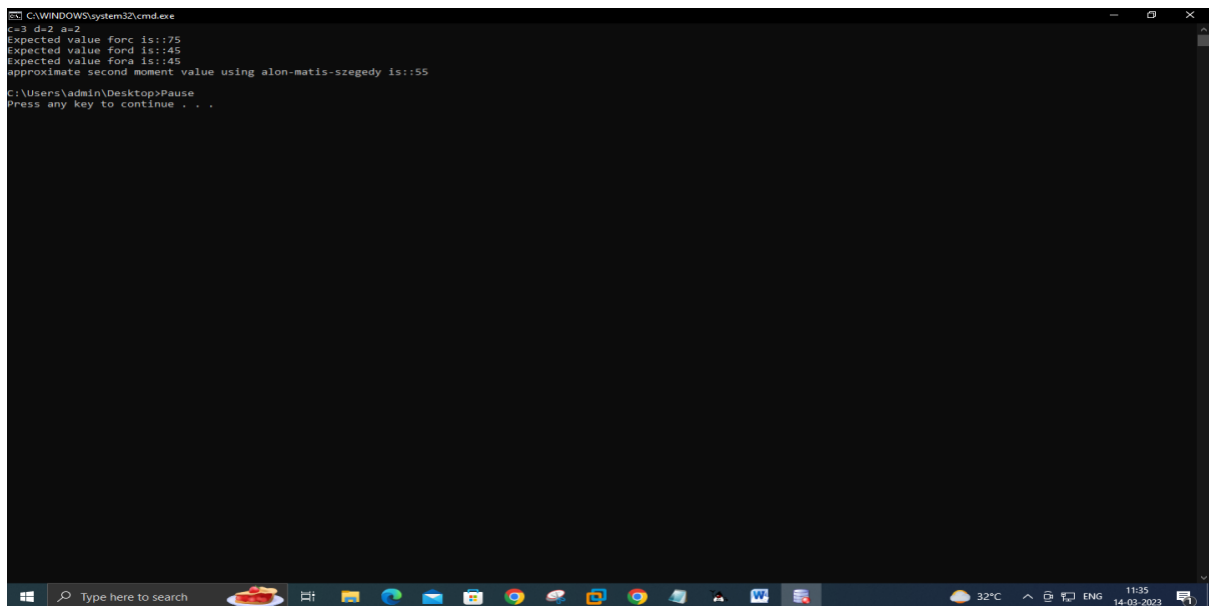
System.out.println("approximate second moment value using alon-matis-szegedy
is::"+apprSecondMomentValue);

}

}

```

Output:



```

C:\WINDOWS\system32\cmd.exe
C:\> java d=2 a=2
Expected value for c is::75
Expected value for d is::45
Expected value for a is::45
approximate second moment value using alon-matis-szegedy is::55
C:\Users\admin\Desktop> Pause
Press any key to continue . . .

```