Readme

4 a) The folder n2 contains co-occurrence of words as pairs and the related input and output file and the time taken to execute them.

The input files:

The input files are randomly chosen.

The input file sizes are taken as multiples of 5 (i.e) 5, 10,……45.

The output files are present there.

The times are noted in n2.txt

Hadoop Commands:

Hdfs dfs –put ~/Data4\_1/ ~/input

hadoop com.sun.tools.javac.Main Stripes\_lemma.java

jar cf Stripes\_lemma.jar Stripes\_lemma \*.class

hadoop jar Stripes\_lemma.jar Stripes\_lemma ~/input/Data ~/output4\_1

hdfs dfs -get ~/output4\_1

Output format:

! = (ipse , 1)= (seruasset , 1)= (ad , 1)= (uerum , 1)= (caritatis , 1)= (adamauit.,1)

' = (iustus , 1)= (si , 1)= (saluus , 1)= (uix , 1)

'II = (negabat. , 1)= (Dauid , 1)= (sed , 1)= (flagitium, , 1)

'aspicit = (non , 1)= (delet , 1)= (et , 1)= (enim , 1)

'cruenta, = (si , 2)= (est: , 1)= (tetra, , 1)

'diues' = (est. , 1)= (huiusmodi , 1)= (inquit , 1)= ((et , 1)

'ecce = (sum, , 1)= (optulit , 1)= (dicens: , 1)= (Christus , 1)= (ego , 1)= (haeresis: , 1)= (detestanda , 1)= (negauit , 1)

The Jupyter notebook contains the plots of the execution.

4 b)

The folder cooc\_n3 contains the implementation of the Activity 4(b)

The input files are multiples of 5 and are selected randomly.

The times are stored in logfilestoren4

Hadoop Commands:

Hdfs dfs –put ~/Data/ ~/input

hadoop com.sun.tools.javac.Main Stripes\_try.java

jar cf Stripes\_try.jar Stripes\_try \*.class

hadoop jar Stripes\_try.jar Stripes\_try ~/input/Data ~/output

hdfs dfs -get ~/output

Output format:

(will,small=3)(joined,declared=2)("869",video=14)(help,missing=14)

(reason,says=5)(may,Paton=3)(Why,one=14)(says,use=7)(Walsh,Partridge=8)

("How,like=10)(pound,like=9)("How,affect=10)(just,"How=5)

(progress,season=11)(keep,needs=13)(take,lid=9)(custody,take=7)

The jupyter notebook contains the plots related to the time.