Lab-5

Anmisha Reddy Ramidi Sindhura Uppu

Person#:50208673 50206730

Activity 2

Bigrams:

The word – co-occurrence program for n=2 was implemented Scala environment of Spark.

Input: The input Latin files must be placed in the folder “Latindata”

Lemmatization: The new\_lemmatizer.csv file has been used for lemmatization.

The output: the output for the program is as follows:

((laxis , Sarmata,),<luc. 1.430>)

((noster , quae),<luc. 1.345>)

((spoliare , ignemque),<luc. 1.379>)

((Libyes , leo),<luc. 1.206>)

((o , mille),<luc. 1.299>)

((quoque, , fortes),<luc. 1.447>)

The co-occurring words and their positions.

Trigrams:

The programs Trigrams.scala contains the code for the n=3 co-occurring words.

The input: Latindata(dir) must have all the input latinfiles

The lemmatization: new\_lemmatizer.csv

The output as follows:

(( Inde , alius , per),<uerg. aen. 6.411>)

(( insignis , Possit , uiuido),<uerg. aen. 11.386>)

(( Quem , fugis? , Extremum),<uerg. aen. 6.466>)

(( et , sacer , Ocius),<uerg. aen. 8.278>)

(( carbasus, , et , umbrosa),<uerg. aen. 8.34>)

The words that co-occur and their positions.