Activity 3:

* New\_Lemmetization.csv file is used to identify the lemmas for each word in the .tess files and is its path is passed to the jar file as an argument through command line.
* Only lucan.bellum\_civile.part.1 and vergil.aeneid are used to calculate the lemma word count.
* Brief Code Logic:
  + A global lemma map is created in main function and used throughout the execution.
  + A Inmap combiner is used for better processing of Mapper output.
  + A Location string is constructed in Inmap Combiner and Reducer by appending the new location identified.
  + Associative Array flushing logic is also added to avoid hitting the heap size of Hadoop Java execution.
* LemmetizationWordCount.jar is used to identify the location of the words from the .tess file and get its count.
* Output format: normalized\_word/lemma\_of\_word {<location1>,<location2>…} {count:wordOccurences}

Commands:

* hdfs dfs -put Latin2Files LemmaWordCountInput
* hadoop jar LemmetizationWordCount.jar LemmaWordCountInput LemmaWordCountOutput
* hdfs dfs -get LemmaWordCountOutput