CS777 – Assignment 4

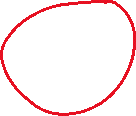
Gaurav Tungare

------------------------------------------------------------------------------------------------------------------------------------------

**Task 1**

Graphical user interface, text, application

Description automatically generated



**Task 2**

**--Task2 -> printing out the five words with the largest regression coefficients-------**

**Task2 : [('eurest', 15141)]**

**Task2 : [('fubilan', 16338)]**

**Task2 : [('mysims', 19642)]**

**Task2 : [('otml', 19869)]**

**Task2 : [('ansto', 17285)]**

A screenshot of a computer

Description automatically generated

**Task 3**

**----Starting Task 3----**

**21/10/03 23:17:02 INFO org.apache.hadoop.mapred.FileInputFormat: Total input files to process : 1**

**---Task 3 ->----F1 Score----**

**F1 : 0.0**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Graphical user interface, application, table

Description automatically generated

Graphical user interface, application, table

Description automatically generated

Graphical user interface, application, table

Description automatically generated

**Question -** Write paragraph describing why you think it is that your model was fooled. Were the bad documents about Australia? The legal system?

**Answer –**

**There is total imbalance of the data set where most of the records are Wikipedia on which the model was trained so there not enough samples in the training set which would have given us good test result so class imbalance is the major cause of skewed F1 score on the test data set**

**Log File**

**Log file attached**

****

**Task 2**

**Task 3**