**SUBMISSION**

**Done By:**

Abhinav Kumar (Person No: 50247660)

Ravi Malik (Person No: 50247302)

**The Data**

We combined all the articles from the four categories into one file named train4.csv and assigned labels to

each article i.e category name. We used Pyspark to do the classification.

**The Environment:**

For classification of data,we used:

* Hadoop VM Xubuntu 14.04 LT
* Pyspark
* Python scripts for collecting the articles and converting them into csv

Two pyspark scripts included :

* logistic\_clasifier.py
* randomforest.py

Python notebook for collecting the articles:

* Articles\_collect.ipynb

Python notebook for converting the articles into a csv:

* Text-to-csv.ipynb

**RUNNING THE CODE :**

Go into the directory where the script/code is present and run the following command :

**spark-submit logistic\_clasifier.py**

**Screenshots**

**The final output of the code :**

D:\My docs\MS in CS\DIC\labs\lab 3\logistic.PNG

**Intermediate output**



