The code file submitted for this assignment is written using python language. The code is executed using the packages mentioned in the importing libraries section. If the cell throws a module error, please uncomment the pip cell which will install the packages required to run this code.

The directory contents of file:

./

group26\_ass2\_imp1.zip

* pred\_labels.csv
* group26\_ass2\_impl.ipynb
* group\_26\_README.docx
* group26\_assignment2\_cover\_sheet.pdf

group26\_ass2\_report.pdf

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Building Instructions:

The code file uses a python version (Python 3.7.7) this should work for the versions of 3 and above.

To run the file, place the code ipynb file and input files for the data in a same file.

The input files include:

* train\_labels.csv
* test.csv
* test\_labels.csv
* data.zip

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Running Instructions:

To run the code use jupyter notebook to open the ipynb file and click on restart and run all cells in the notebook by including the input files.

If needed, you can modify the location of input files given in the code file.

Please refer to markdown cells in the code file to get understanding about the code. If you need more explanation on the algorithms and steps performed, you can refer to the report where we the explanations are given in-depth.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Input Data format :

There are four input files. Three are CSV files and ne is a zip file. Each file on a zip file has approximately 100 tweets.

Train Data : Contains train ID and gender

Test Data : Contains test ID (gender to be predicted)

data.zip

* \_MACOSX
* data
* 1a6ce29582265a11b005501fbb92e9f8.xml
  + <author lang="en">
  + <documents>
  + <document><![CDATA[@nzherald Raurimu is not in Northland]]></document>
  + <document><![CDATA[This is all by design - but what's the end game?]]></document>
  + <document><![CDATA[@OliverCarle I wasnt aiming for soothing 😉]]></document>

The code file will read all the four files accordingly to the file extensions. Note there is no need to unzip the data file the code we wrote will do it for you.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Output Files:

The output files generated after running the code can be found in the same directory as the code which is named as **pred\_labels.csv**.

* pred\_labels.csv contains test ID and predicted gender (‘male’/’female’)

|  |  |
| --- | --- |
|  |  |
|  |  |

The End. Now please run the code file.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*