This package contains the python code and relevant files for authorship profiling model.

The directory contents of this distribution are as follows:

|  |  |
| --- | --- |
| **Files** | **Description** |
| Author Profiling.ipynb | Python implementation of final classifier in Jupyter Notebook File Format. |
| pred\_labels.csv | Label prediction on the test documents is stored. |

We’ve used Python 3.7.6 and conda 4.8.2. This python code has been implemented and tested in Jupyter Notebook.

===========================================================

**Input Files and their data format**

|  |  |
| --- | --- |
| **Files** | **Description** |
| 3600 XML files | These XML files have author's tweets which are to be extracted and gender classification is to be done using this. |
| train\_labels.csv | This file contains all the training document id and gender of the author of that document. |
| test\_labels.csv | This file contains the test document id and gender of the author of that document. This is not to be used in model development but is to be used for label comparison after model is run. |

===========================================================

**Output files and their data format**

|  |  |
| --- | --- |
| **Files** | **Description** |
| pred\_labels.csv | Label prediction on the test documents is stored. This contains document id and predicted gender labels. |

===========================================================

**Running Instruction**

* Firstly, the set of libraries used in the first cell, the required input files such as XML documents (tweet details), train\_labels.csv and test\_labels.csv must be present in the user’s system.
* Once, they’re available, the user has to set the corresponding path or place all the input files in the same directory where the Author Profiling.ipynb file is present to be able to read the input data into the dataframe.
* Lastly, the user can run all the cells and wait until the output file is generated. It’ll be placed in the same path as the Author Profiling.ipynb file.