**Pre Processing**

1) Loading data from json to csv

Run ‘loadfiles.py path’, Give input location to the folder where the original data is present.

2) Removing all non-alphanumeric characters

Run ‘remove\_chars.csv filename’

3) Extract hashtags from raw data and merge with the output of above step for different dataset

Run ‘hashtags.py raw\_filename merge\_filename’

**Classification**

1) ‘nb\_classifier.py’ has the implementation of tf-ids, Inverse cluster frequencies and Naïve Bayes classifier.

It has different modes:

* Run ‘python nb\_classifier.py filepath 2 1000’ for Inverse Cluster Frequency Model, change 1000 to any number to increase number of tweets to process
* Run ‘python nb\_classifier.py filepath 1 1000’ for unique words model, change 1000 to any number to increase number of tweets to process
* Search for below line and Change False to True to enable cross validation

*def main(path, n =1000, cv=False)*

‘python nb\_classifier.py filepath 1 1000’ to run cross validation.