Real Time News Classifier

Saini Datta, Sulagna Patra, Pranav Rane

Contents:

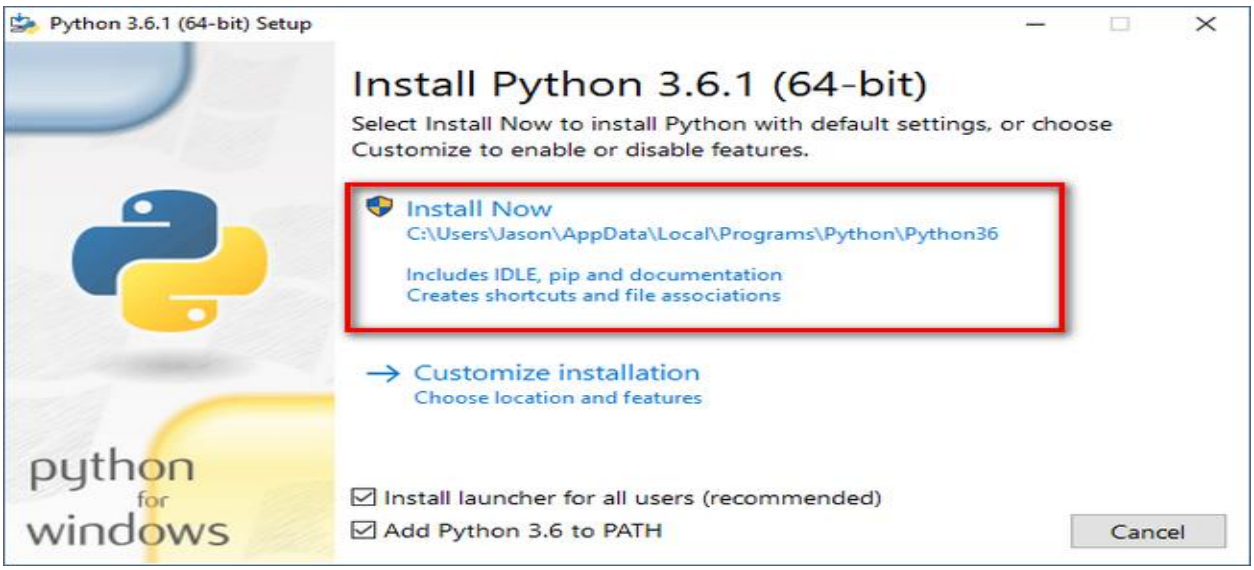
Python Setup - Windows

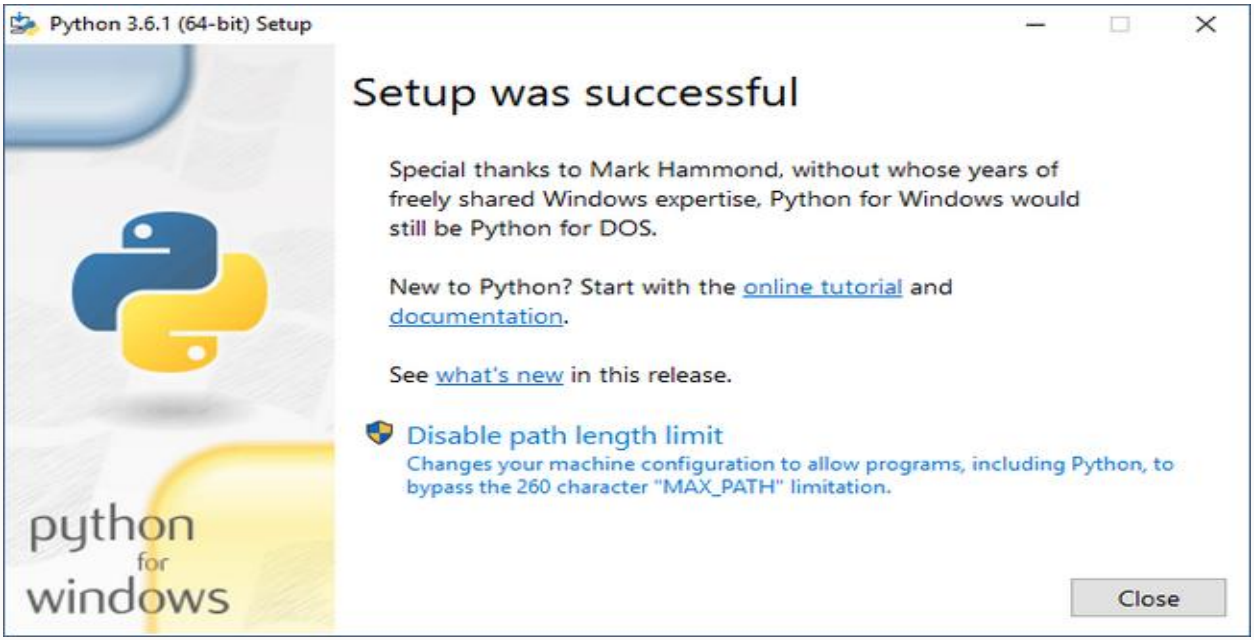
Installation:

Open https://www.python.org/downloads/windows/

Then Download Windows x86-64 executable installer

On the first screen, enable the “Add Python 3.6 to PATH” option and then click “Install Now.”





Usage:

To to Start, search for CMD and open it

Find the location where .py files are located. For eg: `First_Step.py` is at location “C:/Code”

Then set the directory like `cd C:/Code`

Run the code like this:

```
python First_Step.py
```

Python Setup - Ubuntu

Installation:

Open *Terminal*

Check if Python3 is installed by `python3 --version`

If Python is installed then the version number will be displayed.

Eg: Python 3.5.2

If it is not installed then, type the following instructions in *Terminal*

```
sudo apt-get install python3
```

Once installation is complete, validate the process by running `python3 --version`

Usage:

To run a Python Code, Open *Terminal*

Copy the folder address of the Python File.

Example: `First_Step.py` is at location “/home/pranavrane/PycharmProjects/KPT_Project_0”

The location can be identified by checking the Properties of the file.

So to enter the directory via

```
cd /home/pranavrane/PycharmProjects/KPT_Project_0
```

Run the code like this:

```
python3 First_Step.py
```

Python Setup - MacOS

Unfortunately, none of the team members for this project own a Mac based device. So we can't provide exact details.

The instructions for installation can be taken from here:

http://docs.python-guide.org/en/latest/starting/install3/osx/

The run instructions can be found here:

https://en.wikibooks.org/wiki/Python_Programming/Creating_Python_Programs

Pycharm: Installation and Setup

Download and install Pycharm from:

https://www.jetbrains.com/pycharm/download/

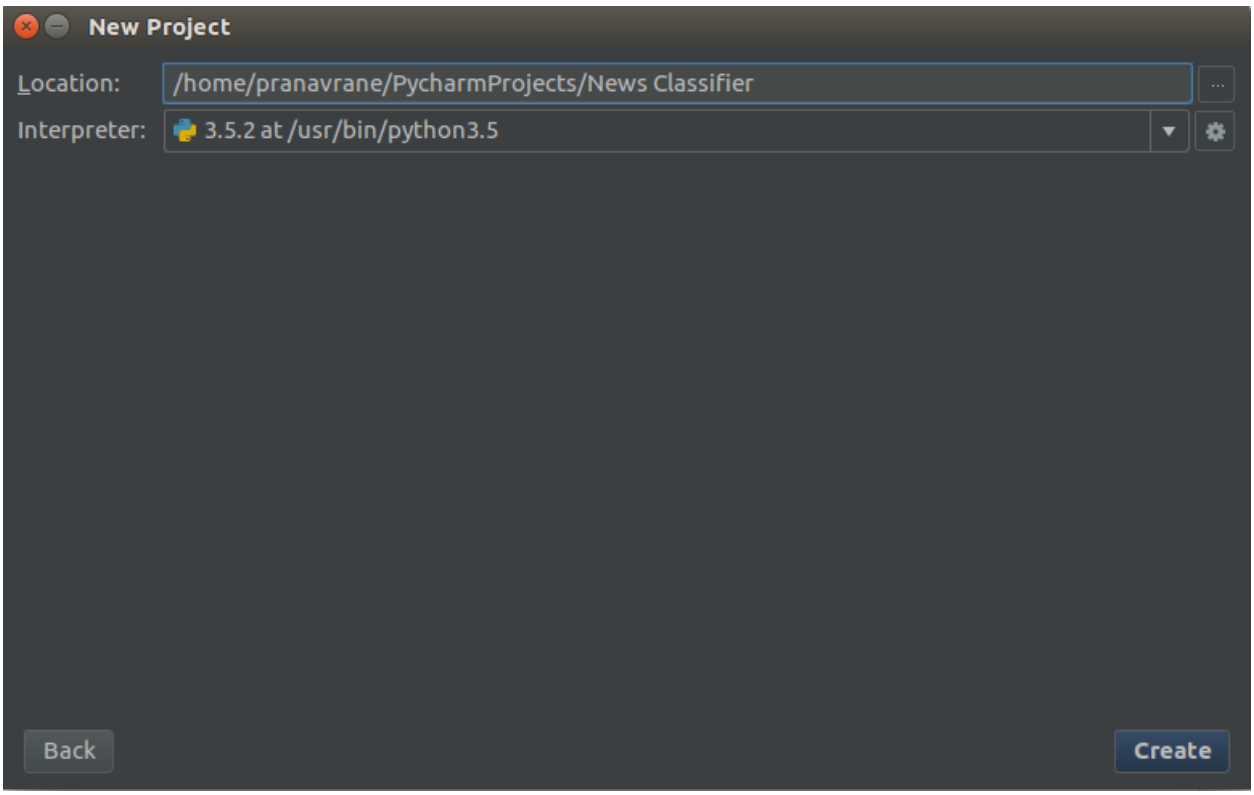
The instructions for installation as per every OS are mentioned here:

https://www.jetbrains.com/help/pycharm/installing-and-launching.html

We use Pycharm as installing libraries becomes very easy. This process stays the same across all Operating Systems.

Start Pycharm, Create New Project

Give Name as 'News Classifier', and ensure Python 3.x is chosen in Python Interpreter and Select 'Create'

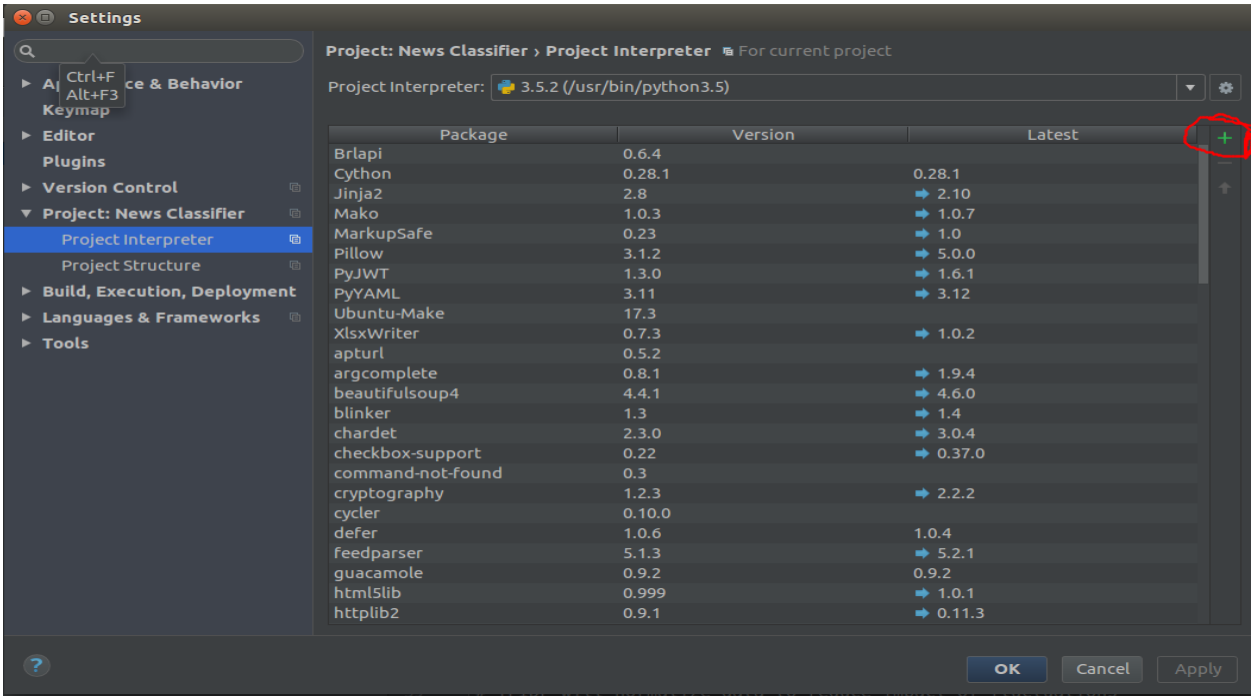


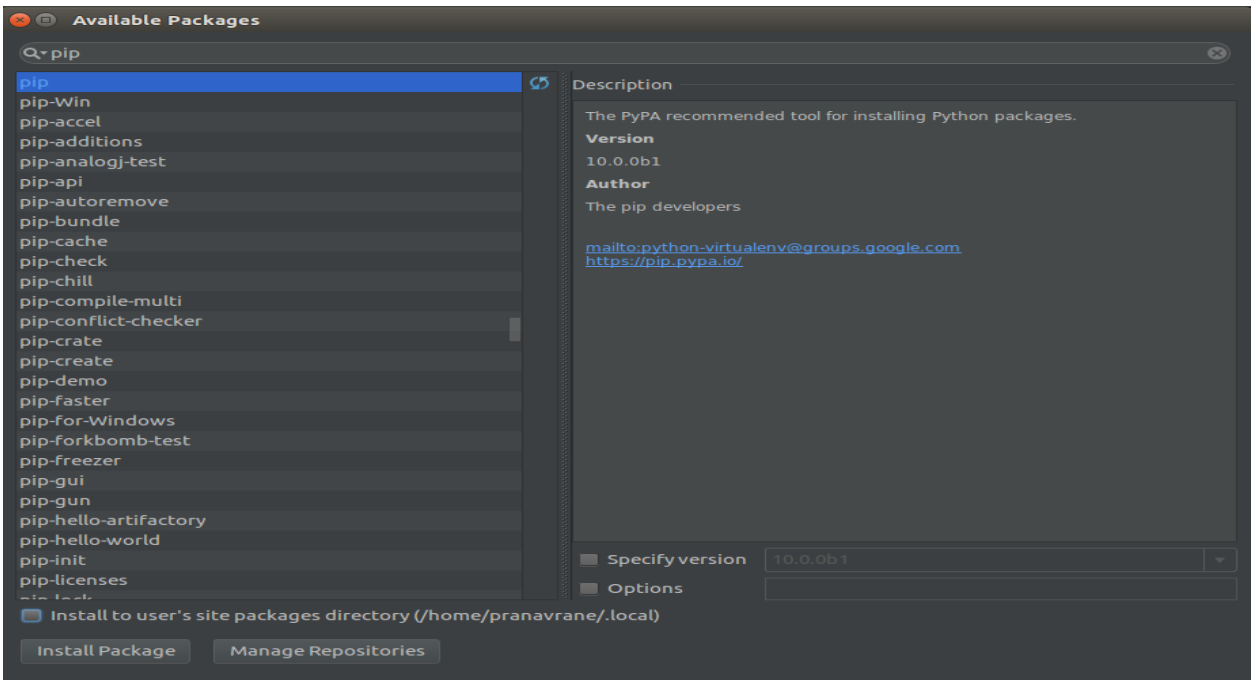
The data file will have to be moved in the 'Location' specified above(Refer Section: Code Execution - Load Dataset)

Pycharm: Install Libraries

Then from Select File>Settings>Select Project: ‘Project Name’> Project interpreter

Click on the Green Plus Symbol in top right and search for the following: Pip





Once the file is located, then choose ‘Install’

Then do the same for the following modules:

- nltk
- pandas

This is the easiest way to install libraries that stays the same across all operating systems.

Incase additional instructions are required:

https://www.jetbrains.com/help/pycharm/installing-uninstalling-and-upgrading-packages.html

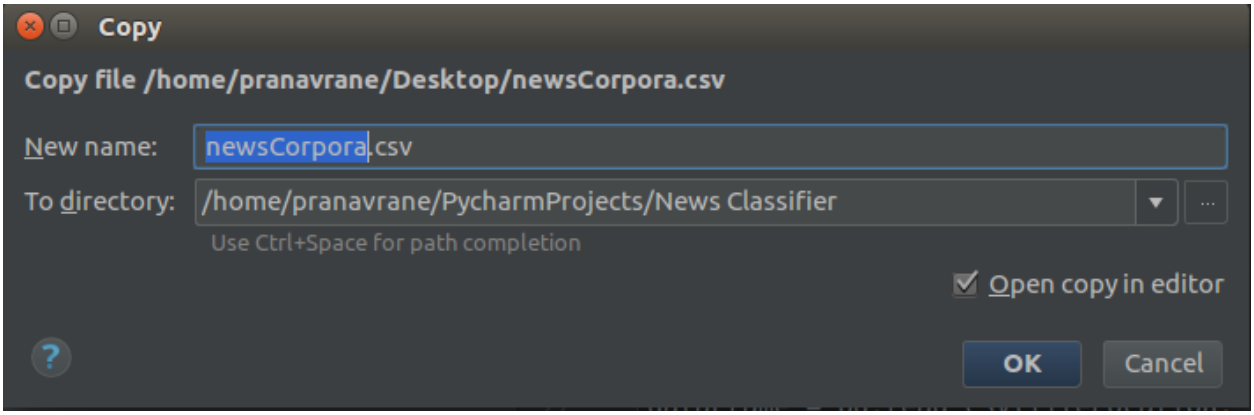
Code Execution - Load Dataset

Get the dataset from:

https://archive.ics.uci.edu/ml/machine-learning-databases/00359/

Unzip the file and move the newsCorpora.csv to the project directory.

The location is specified (‘*To Directory’ in the image below*) when the project is first created in Pycharm. Refer ***‘Python installation and Setup’***



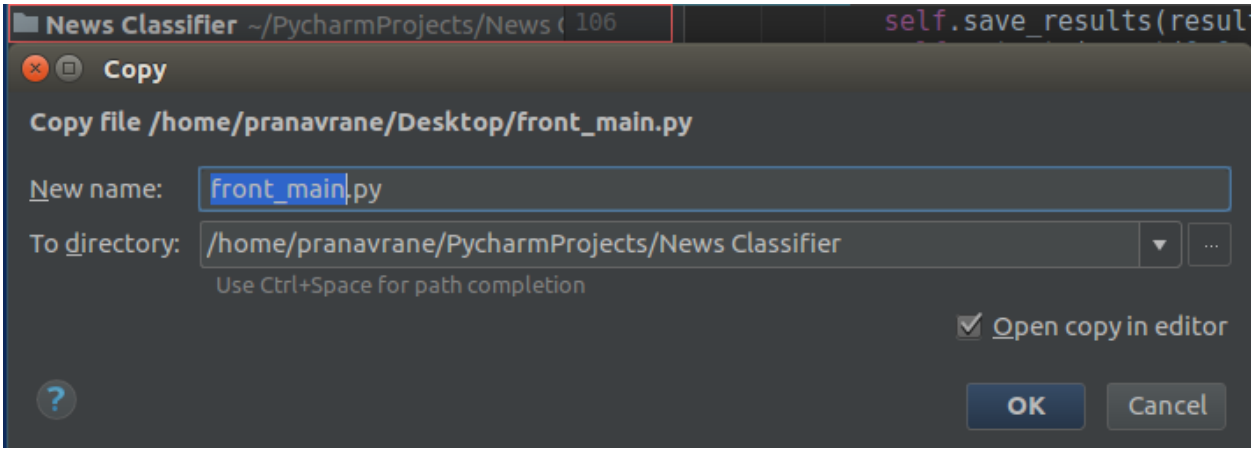
Code Execution - Run Classifier

Table 1. File Names and Use Case

File Name	Type of File	Necessary for code execution?
back_train_get_accuracy_multinaivebayes.py	Code	Yes
back_classify_multinaivebayes.py	Code	Yes
back_import_clean.py	Code	Yes
back_main.py	Code	No. Used for testing
front_train_multinaivebayes.py	Code	Yes
front_classify_ multinaivebayes.py	Code	Yes
front_main.py	Code	Yes
back_train_get_accuracy_knn.py	Code	Yes
back_classify_knn.py	Code	Yes
front_train_knn.py	Code	Yes
front_classify_knn.py	Code	Yes
newsCorpora.csv	Dataset	Yes
classifier_details_knn.txt	Information File	Yes, otherwise Classifier won't run independently. Front_train knn would have to run first otherwise.
sample_input.txt	Information File	No, contains list of headlines for testing.
classifier_details_nb.txt	Information File	Yes, otherwise Classifier won't run independently.

		Front train naïve bayes would have to run first otherwise.

Drag and Move all the files in the project folder.



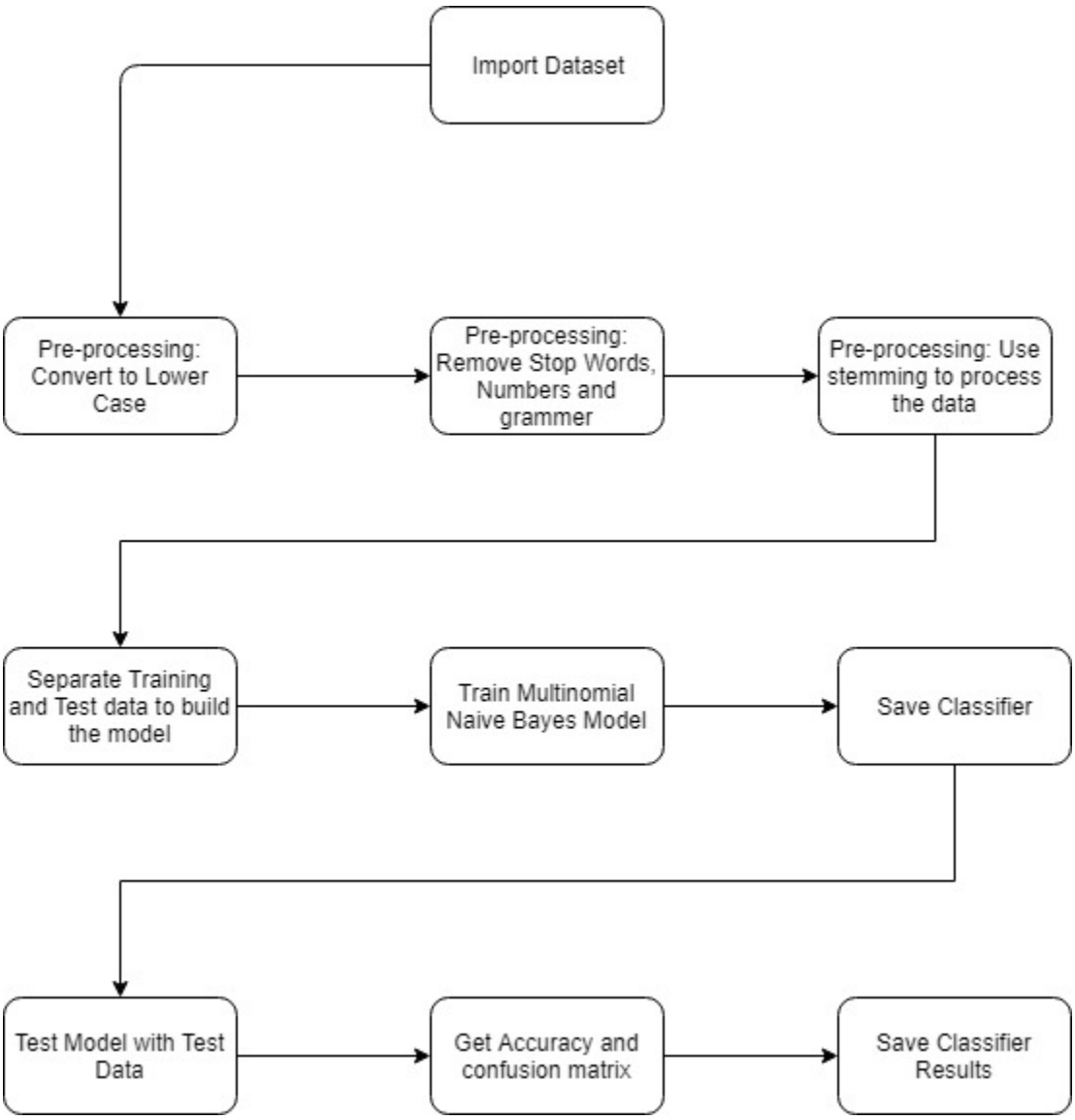
Right Click on Code named **front_main.py**, Select 'Run front_main.py'

Note:

We have tried to test the code on various systems. We have noticed a lot of irregularities while installing packages and Pycharm. Incase the initialization fails, please ask for a LIVE DEMO.

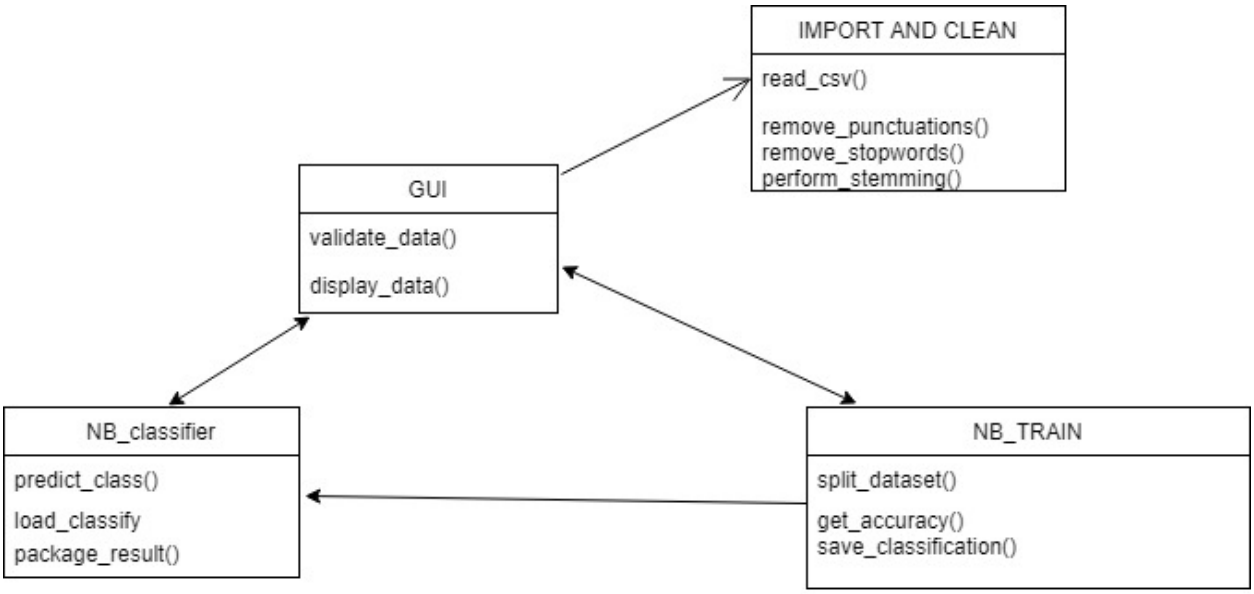
FLOWCHART

The flowchart demonstrates a general representation of the flow of data during classification.

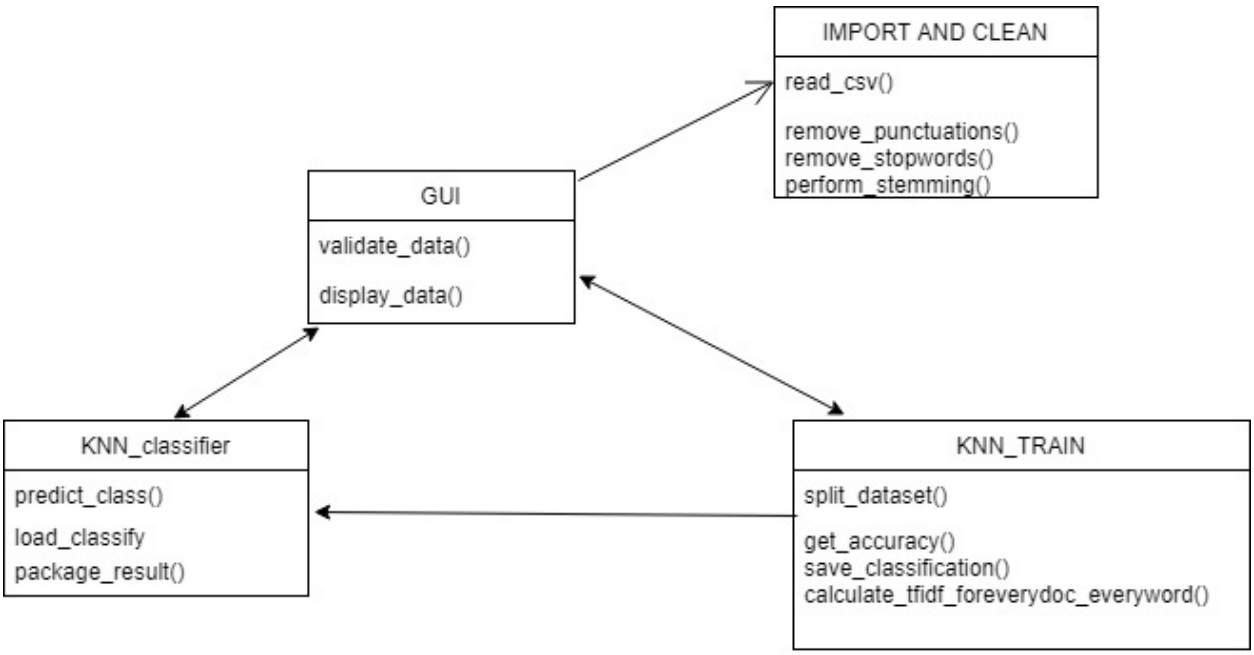


UML Diagram

CLASS DIAGRAM FOR MULTINOMIAL NAÏVE BAYES

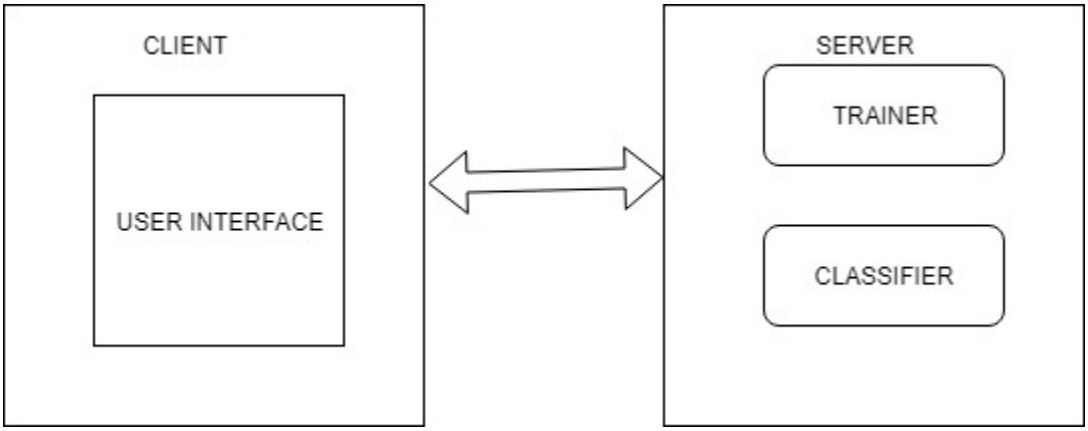


CLASS DIAGRAM FOR K NEAREST NEIGHBOUR



SYSTEM ARCHITECTURE

Client-Server architecture:



The News Classification application is developed using standard client-server architecture. The client side utilizes PyCharm. The server side includes trainer and classifier that listen to user requests and responds to client.

CODE DETAILS AND USER INTERFACE

The following are the different modules of our project -

A. Backend Code:

a. The Backend Code is Written in Python

b. There are three major objectives for this code:

i. Preprocessing:

1. Remove Words with Numbers
2. Remove Non English Words
3. Remove Punctuations, convert to lowercase
4. Remove Stop Word
5. Perform Stemming on every word

ii. Training

1. Split Dataset for Training and Testing
2. Build a Multinomial Naive Bayes Classifier and train it
3. Build a K Nearest Neighbour Classifier and train it
4. Save the classifier

iii. Testing
    1. Get Accuracy Model
    2. Get Confusion Matrix
iv. Prediction
    1. Load Saved Classifier (Training Every Time is expensive)
    2. Predict for Single/Multiple Headlines
    3. Save Predictions

c. The Backend Codes are in the format "back_*.py"

B. Data

a. The data is sourced from https://archive.ics.uci.edu/ml/machine-learning-databases/00359/
b. It contains Headlines, Categories, Source, Address etc.
c. We will retain and use Headlines and their respective Categories.
d. The data will undergo significant preprocessing as described in Section A, point b-1.
e. There are 4 news categories: Entertainment, Science and Technology, Business and Health.

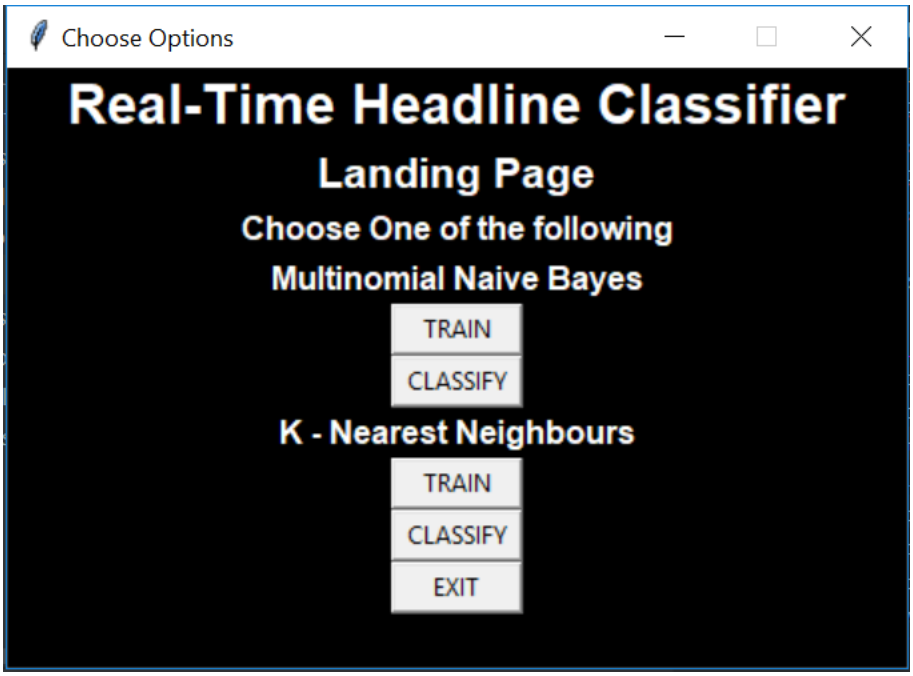
C. Business Logic

a. The tool is designed for Editors and Managers who work at a news publishing agency.
b. Significant news headlines are generated everyday in the 'wires'; It is the job of Managers to categorize the news headlines and assign a journalist to them.
c. Then a journalist will then follow up the story with sources, add context and publish it in a newspaper or website.
d. The tool is designed such that editors don't have to manually decide the news type; the news story will be classified and sent to the respective department. Eg: Business department.
e. We have not applied any business rules to data i.e. all news stories are considered as is without any change.
f. The changes are only algorithmic and technical not business oriented.

D. User Interface

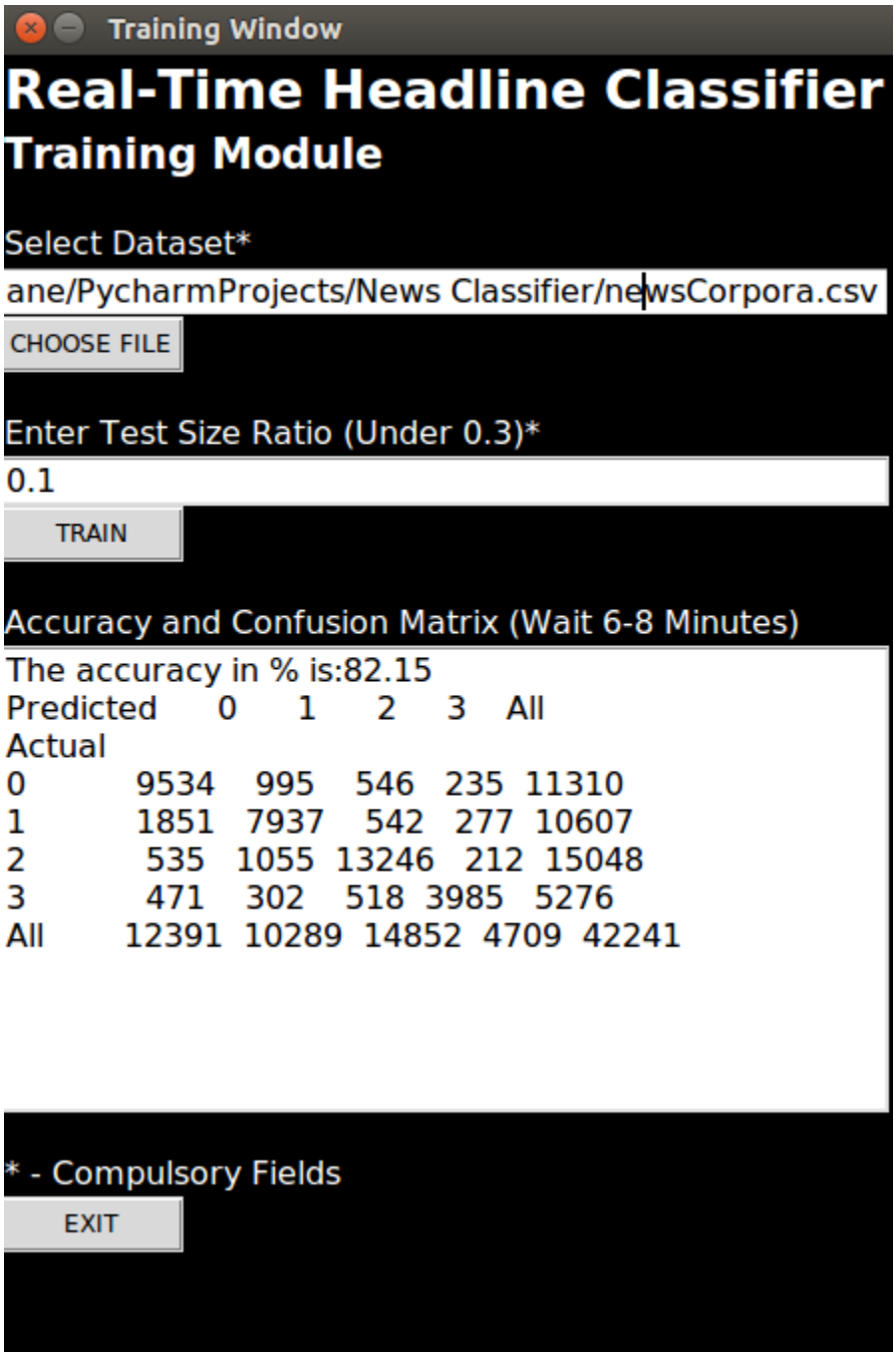
a. The User Interface is made using Tkinter Library
b. There are 2 major modes and a landing page

c. The following is the landing Page:



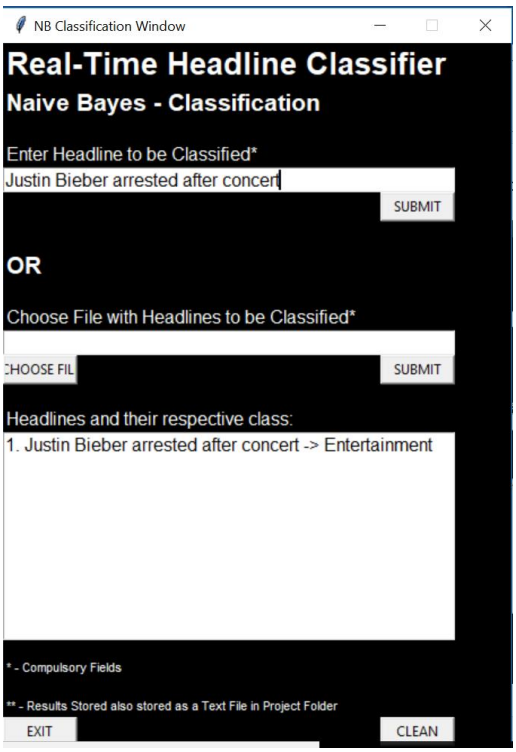
d. We can go to the Train and Classification section from the landing page

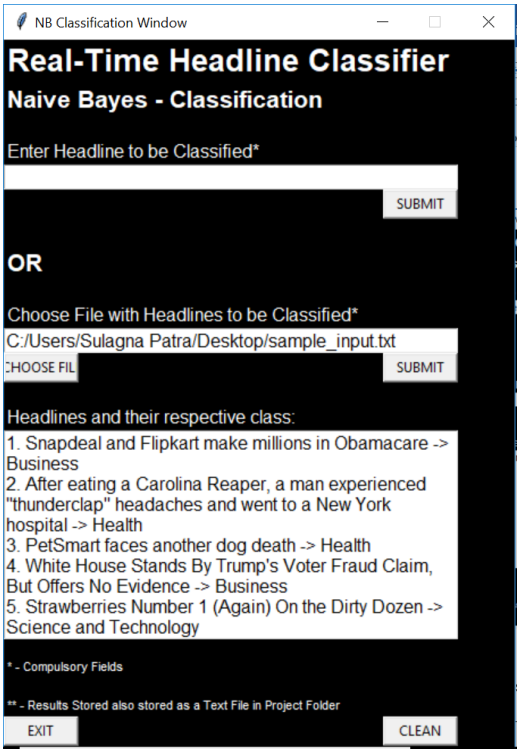
e. This is the training section. **Note**: Training can take upto 10 minutes:



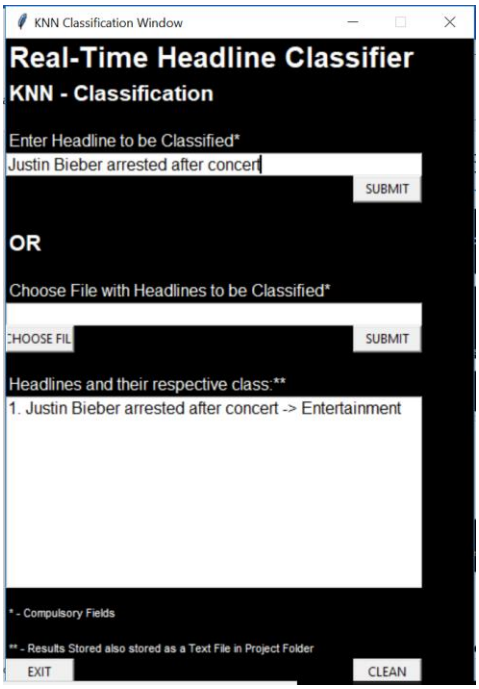
f. We can classify individual headlines and files containing headlines in Classification Module. The file containing headline can be imported manually

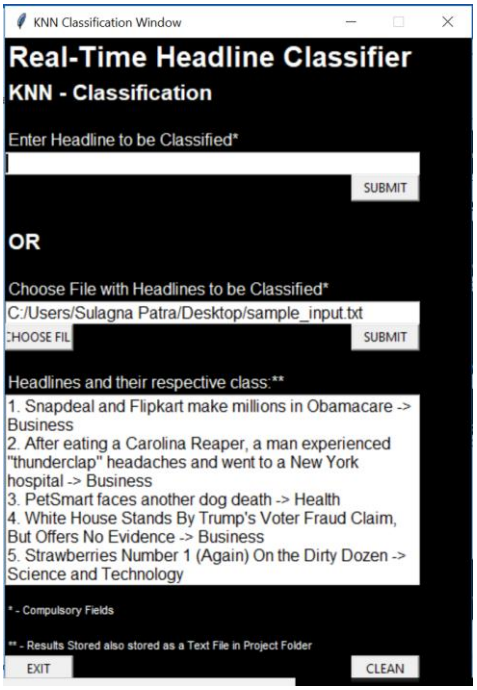
NAÏVE BAYES CLASSIFICATION



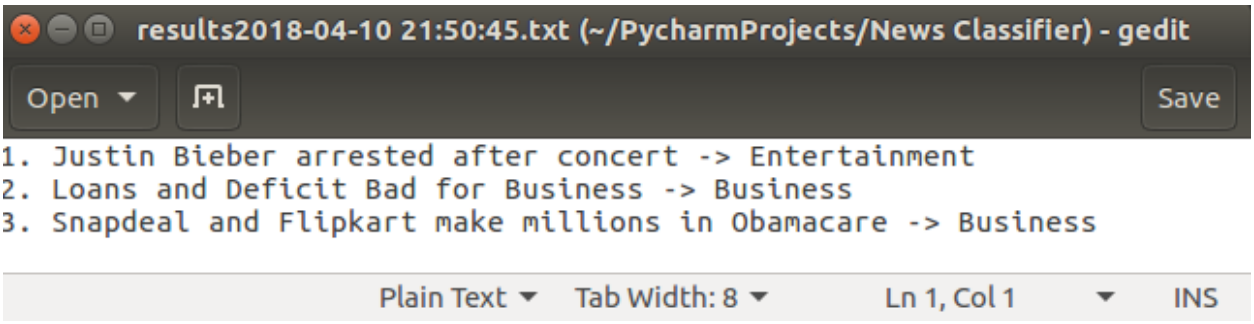


KNN CLASSIFICATION

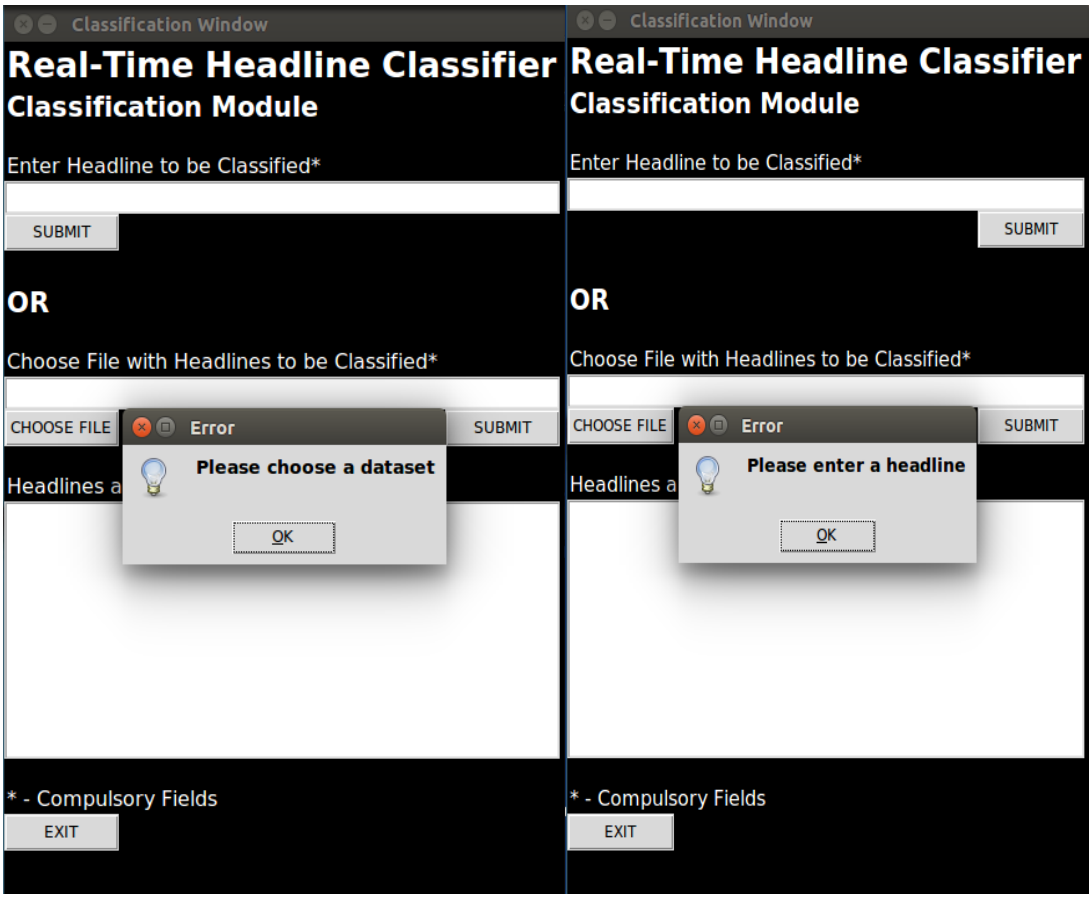




g. The results are exported to a text file:



h. Faulty user inputs are met with error prompts



TEST RESULTS

i. With Multinomial Naive Bayes Classifier we have 82% accuracy. The total rows in dataset are over 400,000. We have 10% of them as test dataset.

j. The confusion matrix for Multinomial Naive Bayes Classifier:

Predicted	0	1	2	3	All
Actual					
0	9534	995	546	235	11310
1	1851	7937	542	277	10607
2	535	1055	13246	212	15048
3	471	302	518	3985	5276
All	12391	10289	14852	4709	42241

k. With K Nearest Neighbor Classifier have 78.26% accuracy. The total rows in dataset are over 400,000. We have 10% of them as test dataset.

l. The confusion matrix for Multinomial Naive Bayes Classifier:

Predicted	0	1	2	3	All
Actual					
0	9264	930	814	302	11310
1	2040	7404	928	235	10607
2	941	956	12974	177	15048
3	665	462	733	3416	5276
All	12910	9752	15449	4130	42241

Result Classification for Multinomial Naïve Bayes			
Input Testcases	**Actual**	**Predicted**	**Pass/Fail**
bskdnsijamklasmklsksa	Not Applicable	Not Applicable	Pass
After eating a Carolina Reaper, a man experienced "thunderclap" headaches and went to a New York hospital	Health	Health	Pass
PetSmart faces another dog death	Health	Health	Pass
White House Stands By Trump's Voter Fraud Claim, But Offers No Evidence	Business	Business	Pass
Snapdeal and Flipkart make millions in Obamacare	Business	Business	Pass
Strawberries Number 1 (Again) On the Dirty Dozen	Health	Health	Pass
All the Times Ryan Reynolds And Blake Lively Roasted Each Other	Entertainment	Entertainment	Pass
Intersectionality-focused series continues with Jahmal B. Golden	Entertainment	Entertainment	Pass
The Tavern at Gibbs offers laidback fare and downtown proximity	Health	Entertainment	Fail
Google staff protest firm's involvement with Pentagon drones programme	Science and Technology	Science and Technology	Pass
JFK Jr.'s wedding is the subject of a TLC special boasting never-before-seen footage of the private ceremony	Entertainment	Entertainment	Pass
Students solve math, science and technology problems at STEM Challenge	Health	Science and Technology	Fail
Hubble Telescope Discovers a Light-Bending 'Einstein Ring' in Space	Science and Technology	Science and Technology	Pass
Brain effects of 'hottest pepper in the world' put man in hospital	Health	Health	Pass
Fake pot likely tainted with rat poison kills 3, sickens 100	Health	Health	Pass

Instagram rolls out Focus portrait mode for videos and photos	Entertainment	Entertainment	Pass
Russia-linked account pushed fake Hillary Clinton sex video	Entertainment	Entertainment	Pass
Apple ordered to pay VirnetX $502.6M in patent infringement row	Science and Technology	Science and Technology	Pass
Yulia Skripal, poisoned daughter of ex-spy, out of hospital	Business	Business	Pass
LeBron James Says He'll Ditch Social Media Again for 2018 NBA Playoffs	Entertainment	Entertainment	Pass
326731781380181	Not Applicable	Not Applicable	Pass

Result Classification for K Nearest Neighbour			
Input Testcases	**Actual**	**Predicted**	**Pass/Fail**
bskdnsijamklasmklsksa	Not Applicable	Not Applicable	Pass
After eating a Carolina Reaper, a man experienced "thunderclap" headaches and went to a New York hospital	Business	Health	Fail
PetSmart faces another dog death	Health	Health	Pass
White House Stands By Trump's Voter Fraud Claim, But Offers No Evidence	Business	Business	Pass
Snapdeal and Flipkart make millions in Obamacare	Business	Business	Pass
Strawberries Number 1 (Again) On the Dirty Dozen	Science and Technology	Health	Fail
All the Times Ryan Reynolds And Blake Lively Roasted Each Other	Entertainment	Entertainment	Pass
Intersectionality-focused series continues with Jahmal B. Golden	Entertainment	Entertainment	Pass
The Tavern at Gibbs offers laidback fare and downtown proximity	Entertainment	Entertainment	Pass
Google staff protest firm's involvement with Pentagon drones programme	Science and Technology	Science and Technology	Pass
JFK Jr.'s wedding is the subject of a TLC special boasting never-before-seen footage of the private ceremony	Entertainment	Entertainment	Pass
Students solve math, science and technology problems at STEM Challenge	Science and Technology	Science and Technology	Pass
Hubble Telescope Discovers a Light-Bending 'Einstein Ring' in Space	Science and Technology	Science and Technology	Pass
Brain effects of 'hottest pepper in the world' put man in hospital	Entertainment	Health	Fail
Fake pot likely tainted with rat poison kills 3, sickens 100	Health	Health	Pass

Instagram rolls out Focus portrait mode for videos and photos	Entertainment	Entertainment	Pass
Russia-linked account pushed fake Hillary Clinton sex video	Business	Entertainment	Fail
Apple ordered to pay VirnetX $502.6M in patent infringement row	Science and Technology	Science and Technology	Pass
Yulia Skripal, poisoned daughter of ex-spy, out of hospital	Health	Business	Fail
LeBron James Says He'll Ditch Social Media Again for 2018 NBA Playoffs	Entertainment	Entertainment	Pass
326731781380181	Not Applicable	Not Applicable	Pass

KEY ISSUES

Predicting class for unknown words was significant challenge. Initially a lot of records were mis-classified as Entertainment. That was default predicted class for all the times when we tried to predict the new words. We overcame that issue by validating every word for its presence in the dictionary. If brand new words appear then ‘Not Applicable’ class is predicted. But this approach is expensive and not very smart. We can use deep learning to identify the meaning of words and classifying new words accordingly.