Text Summarization using Firefly Algorithm

Text summarization gives us an overview of large text which help in understanding large text just by seeing its summary. In the past all summaries were given manually which is very time consuming and requires lots of human efforts. In the past many algorithms such as machine learning and deep learning algorithms were introduced but all those algorithms were lack of optimization so its summarization is not accurate.

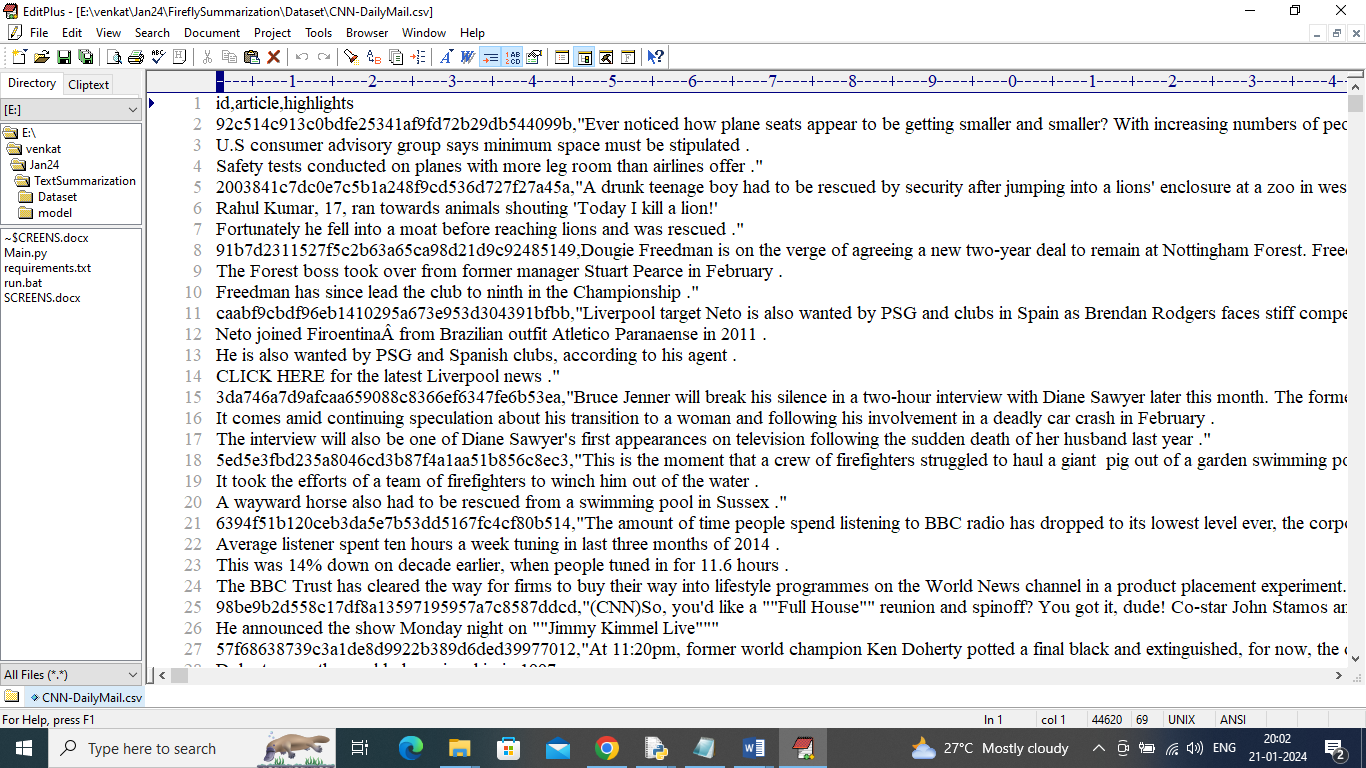
To overcome from above issue we are employing Firefly optimization algorithm on training dataset to optimize training features and this optimize features will give better summary.

In propose work training dataset contains ‘Article and summaries’ and this article will be converted to numeric vector which will replace each word in the vector with its average frequency and this vector features will be weighted and the sentences which has high weight will be consider as better sentences for summaries. This weight will be optimized with Firefly algorithm to select only those weight which has high weight and this high weight sentences will be summarize.

Firefly algorithm (FA) works on the principle of directing the less shiny than the light intensity emitted by fireflies in nature towards the bright. The algorithm can adaptively select the best subset of features and improve classification accuracy.

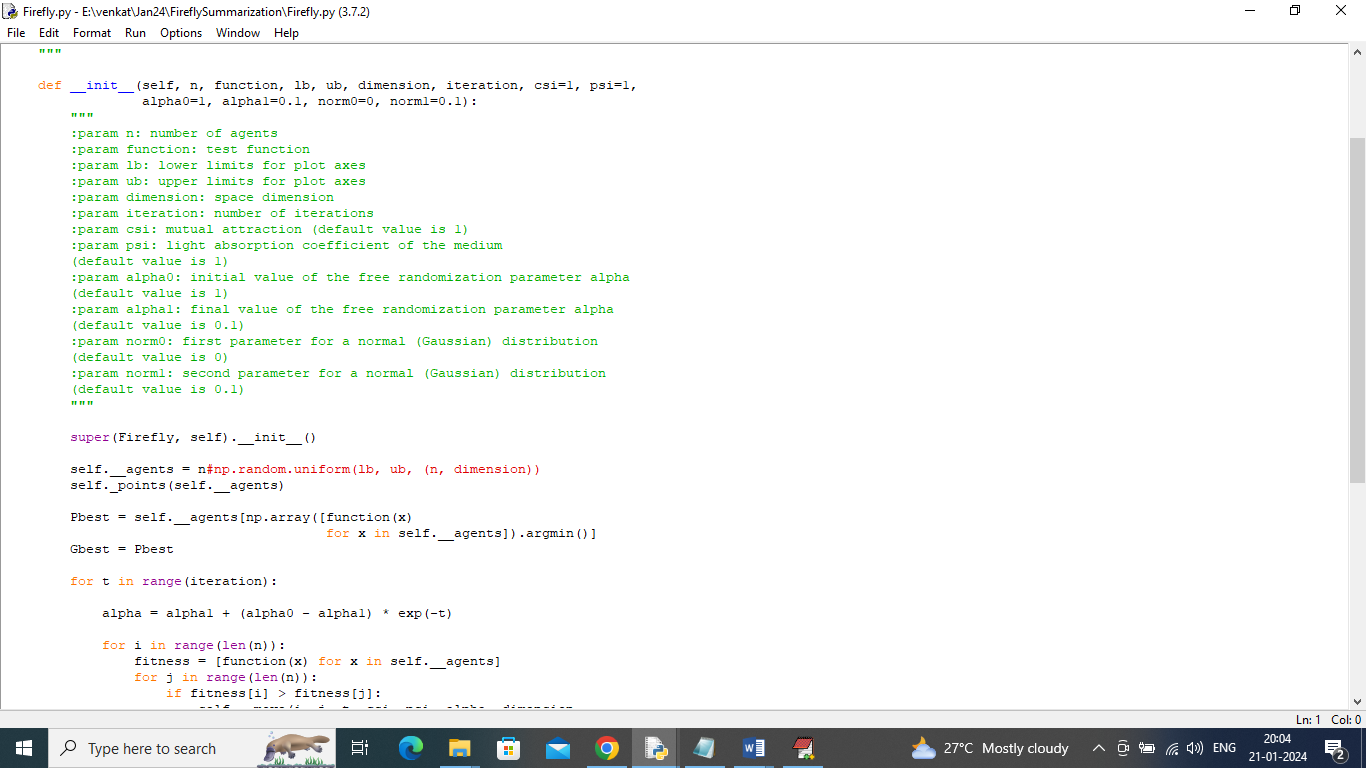
So the above technique of firefly will be adapted to select those sentences or features which his high weight.

To train Firefly for optimization we have used ‘CNN Daily Mail’ dataset and below screen displaying dataset values

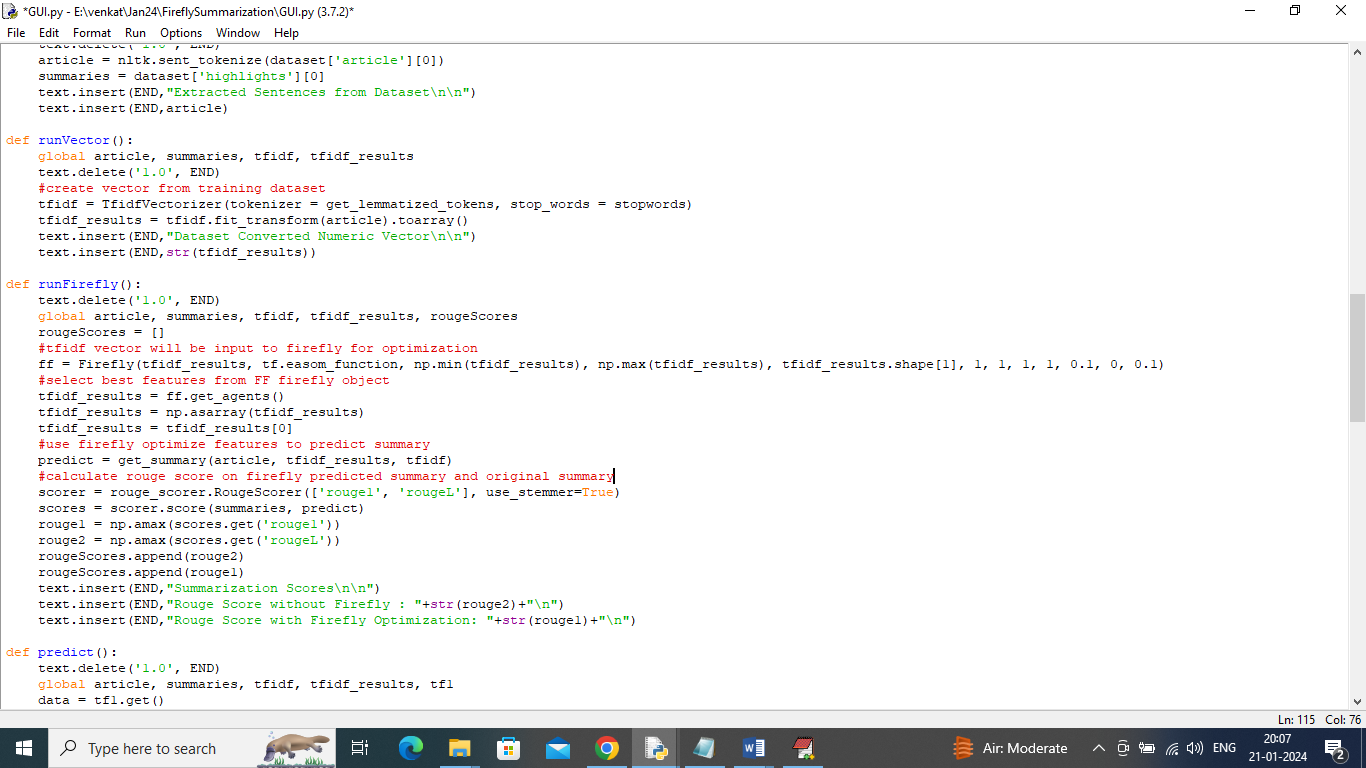


In above dataset first row represents dataset column names and remaining rows represents dataset values. Second column contains articles and 3rd highlight column contains summary and by using articles we will train Firefly algorithm for summarization by selecting features with high weight.

In below screen showing code for firefly algorithm



In above Firefly.py file algorithm will continuously iterate all features to select sentences with high weight or best weight. In below screen showing code to optimize training features using Firefly



In above screen read red or green colour comments to know about Firefly coding.

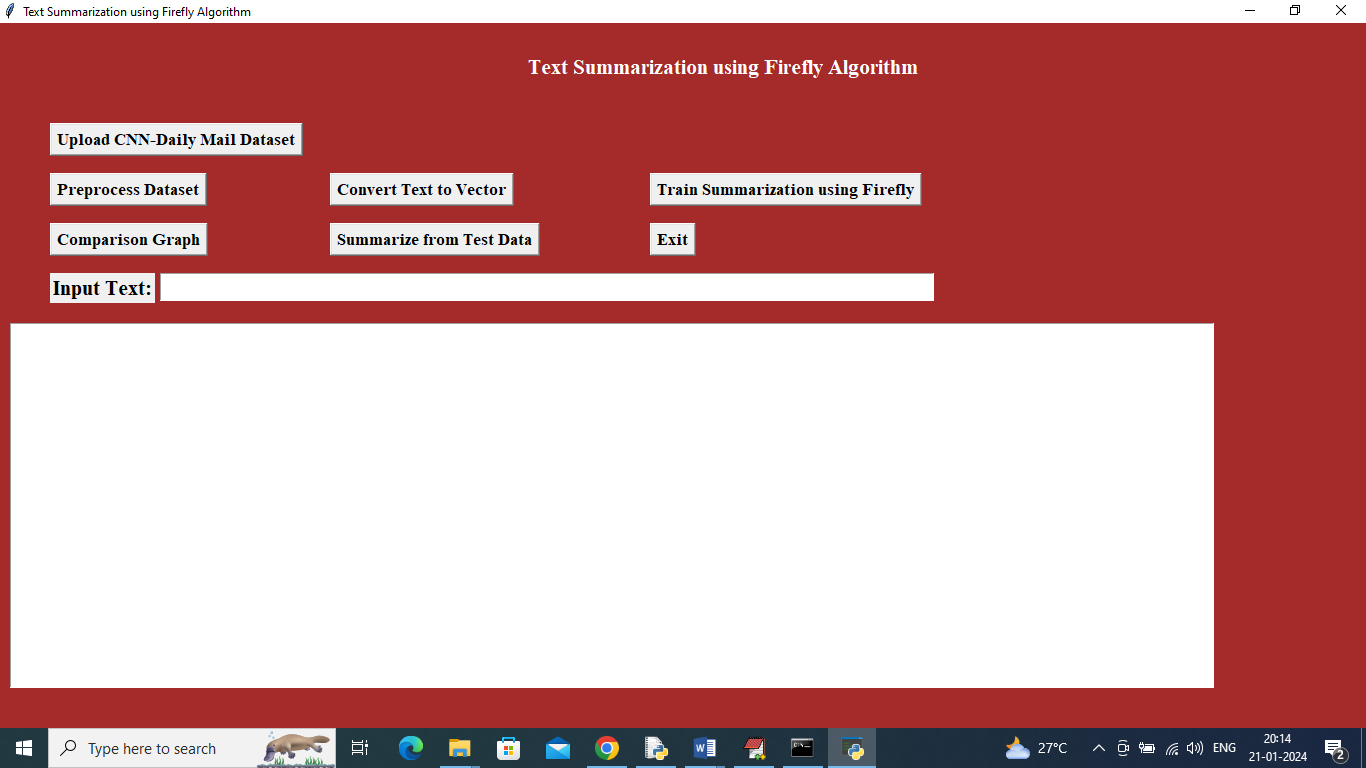
ROUGE score, or Recall-Oriented Understudy for Gisting Evaluation, is a metric used to evaluate the quality of machine-generated summaries and document translation models. It compares the output of an NLP model to one or more human reference texts. The higher the Rouge score the better is the summary and its algorithm.

To implement this project we have designed following modules

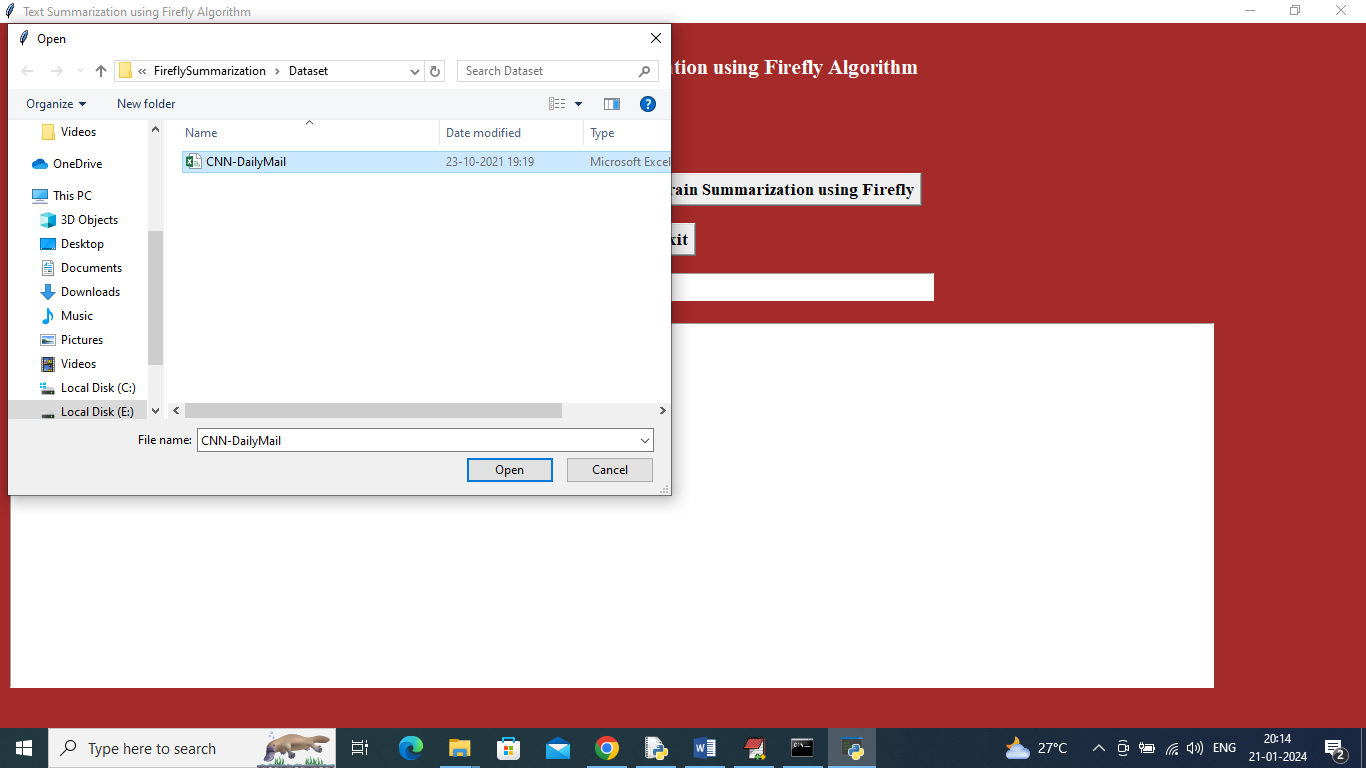
1. Upload CNN-Daily Mail Dataset: using this module we will upload dataset to application and then read and display all articles from dataset
2. Pre-process Dataset: using this module extracted articles will be cleaned by removing special symbols and stop words. Stop words are those words which are not important like ‘and the are what’ and many more are consider as stop words
3. Convert Text to Vector: process articles will be input to this module to convert all text data into numeric vector and this vector will replace each words with its average frequency
4. Train Summarization using Firefly: Vector will be input to this module to optimize all features as per occurrence and its weight values and this high weight sentences will be chosen by firefly for summary
5. Comparison Graph: using this module we will plot Rouge score comparison between Non-Firefly and Firefly summary.
6. Summarize from Test Data: in this module we will input test data and then Firefly will summarize from given test data.

SCREEN SHOTS

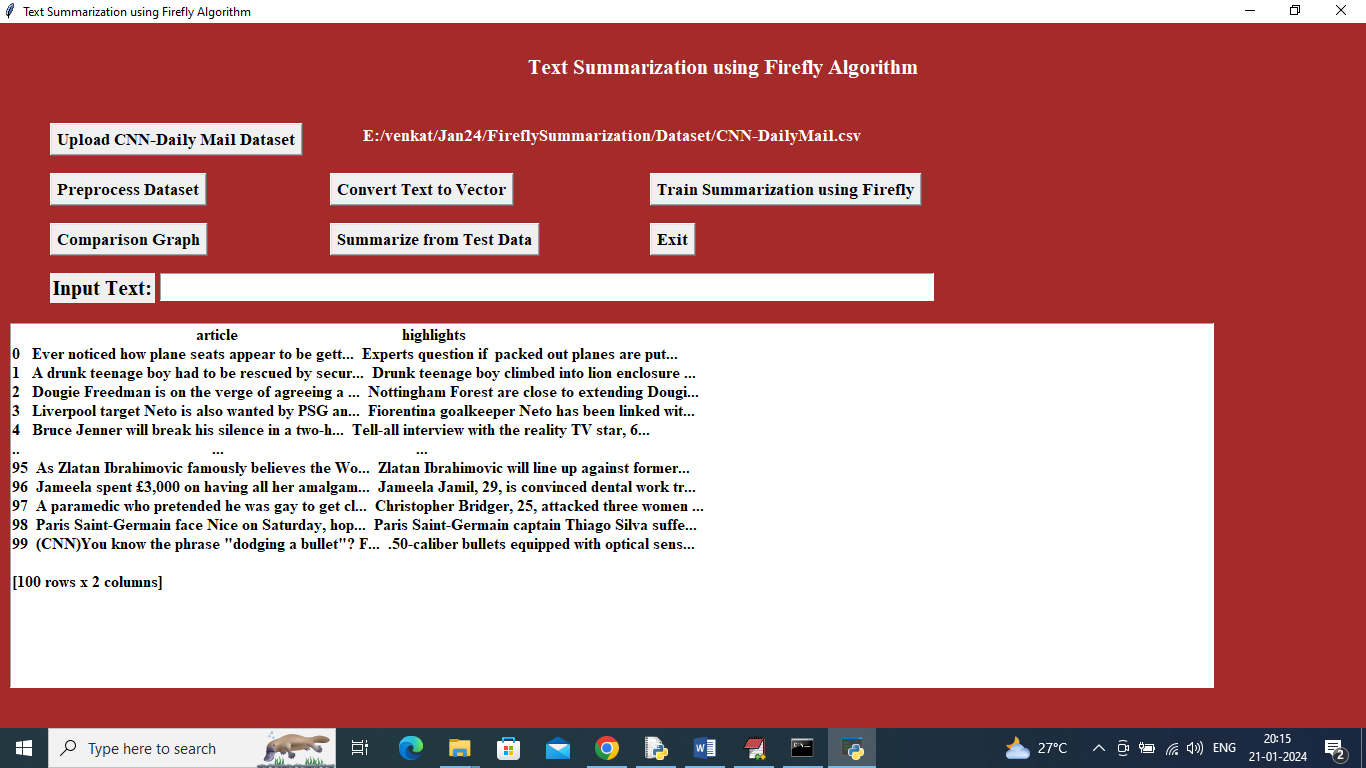
To run project double click on run.bat file to get below screen



In above screen click on ‘Upload CNN-Daily Mail Dataset’ button to upload dataset and then will get below output



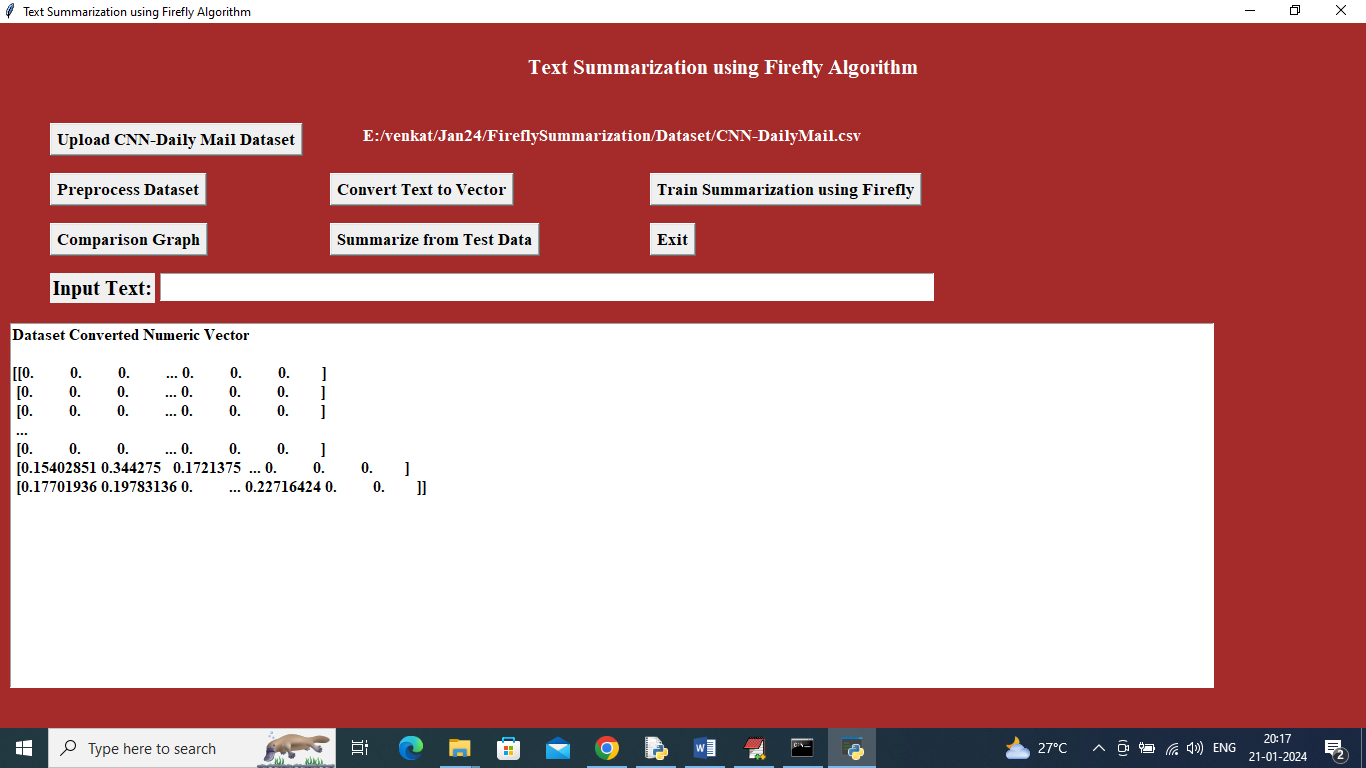
In above screen selecting and uploading ‘CNN Daily Mail’ dataset and then click on ‘Open’ button to load dataset and then will get below output



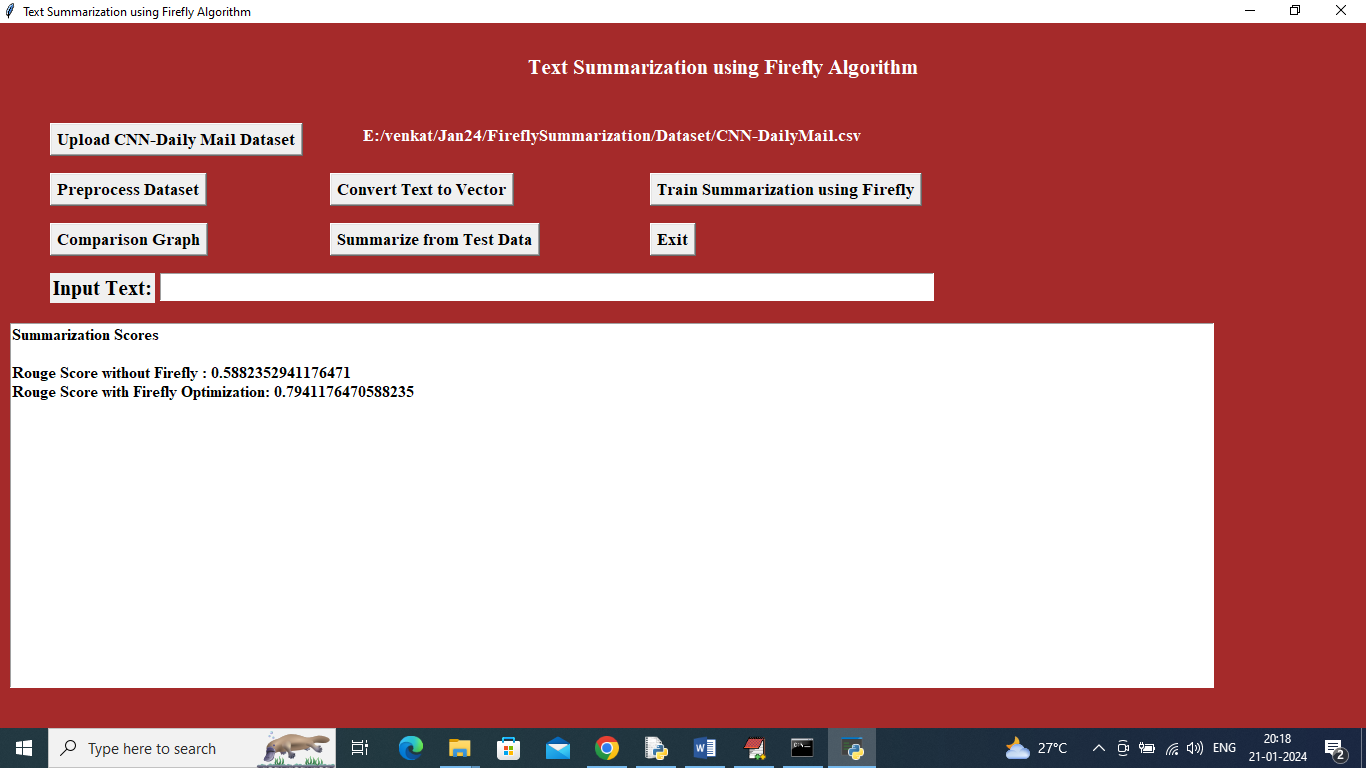
In above screen dataset loaded and now click on ‘Preprocess Dataset’ button to convert all dataset values in to sentences and then will get below output



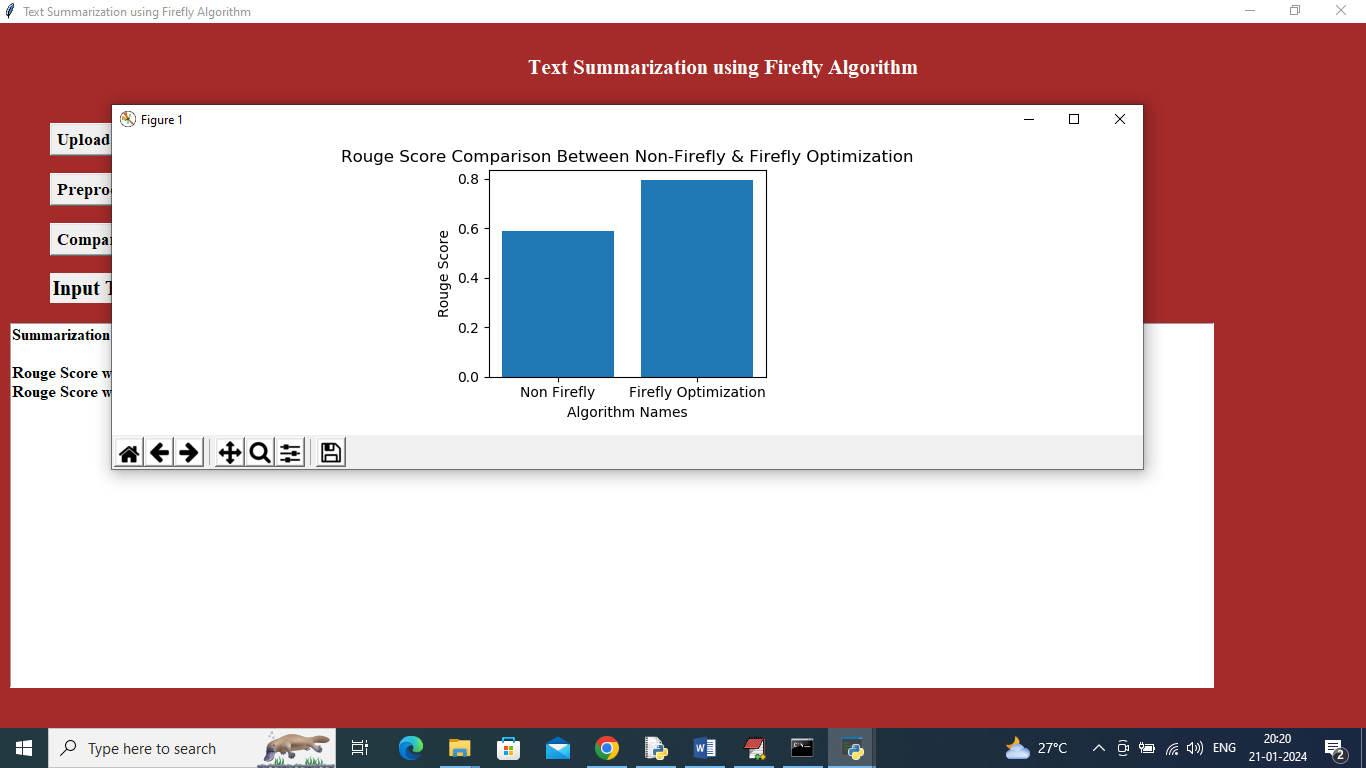
In above screen dataset converted to sentences and displaying few sentences from them and now click on ‘Convert Text Vector’ button to convert above sentences into vector and then will get below output



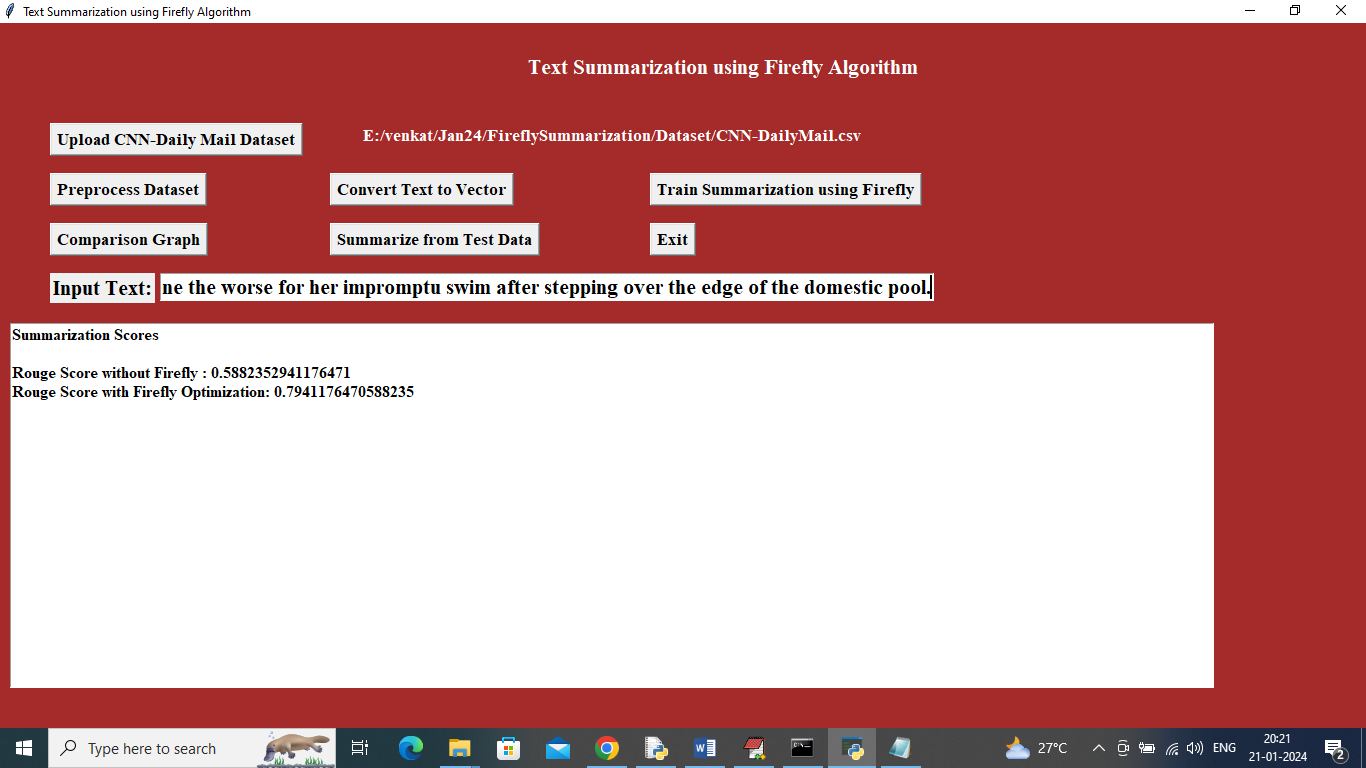
In above screen sentences converted to vector and now click on ‘Train Summarization using Firefly’ button to optimize vector with Firefly to summarize data with high weight values



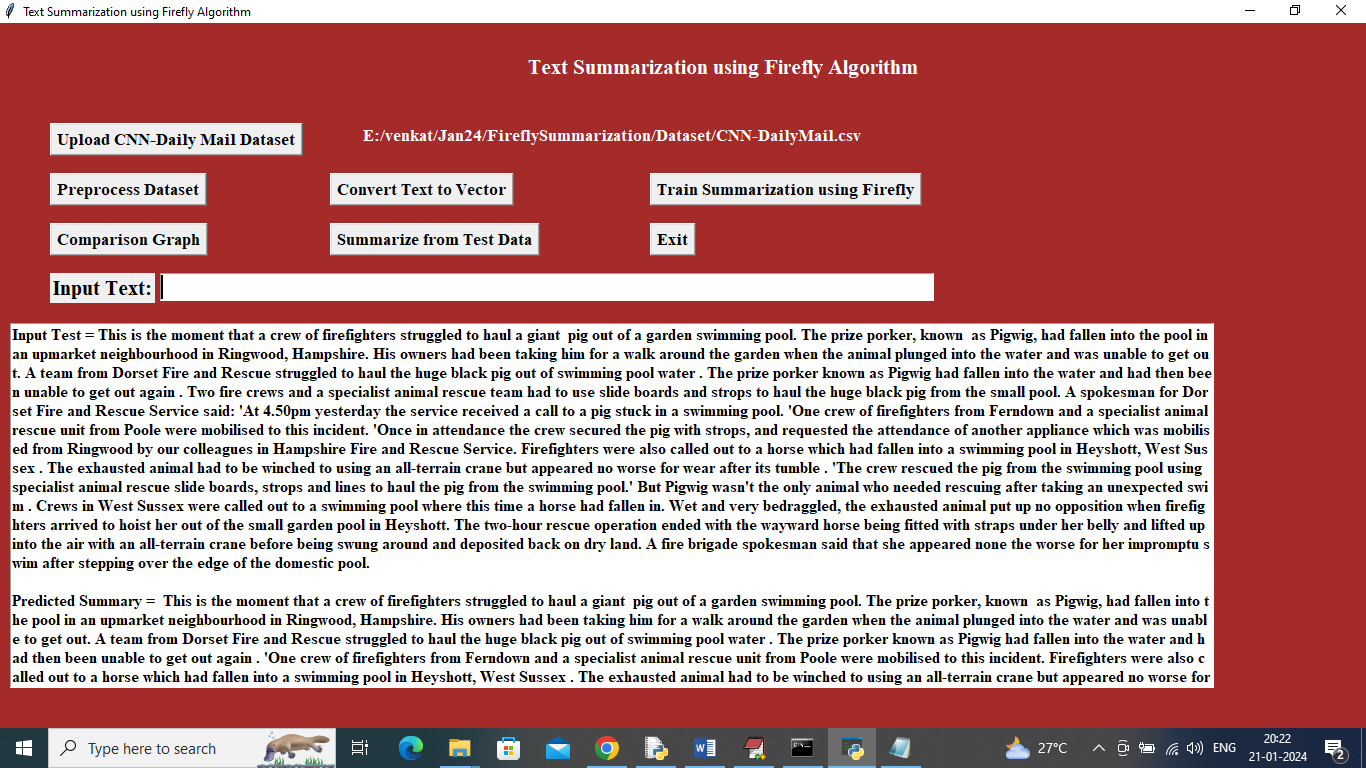
In above screen Firefly training completed where Non-firefly generated summary got 58% rouge score or accuracy and Firefly optimized generated summary got 79% rouge score or accuracy. Now click on ‘Comparison Graph’ button to get below graph



In above screen x-axis represents algorithm names and y-axis represents Rouge Score and in both algorithm Firefly optimized summary got high score and now enter some text in the text filed and then click on ‘Summarize from Test Data’ button to predict summary



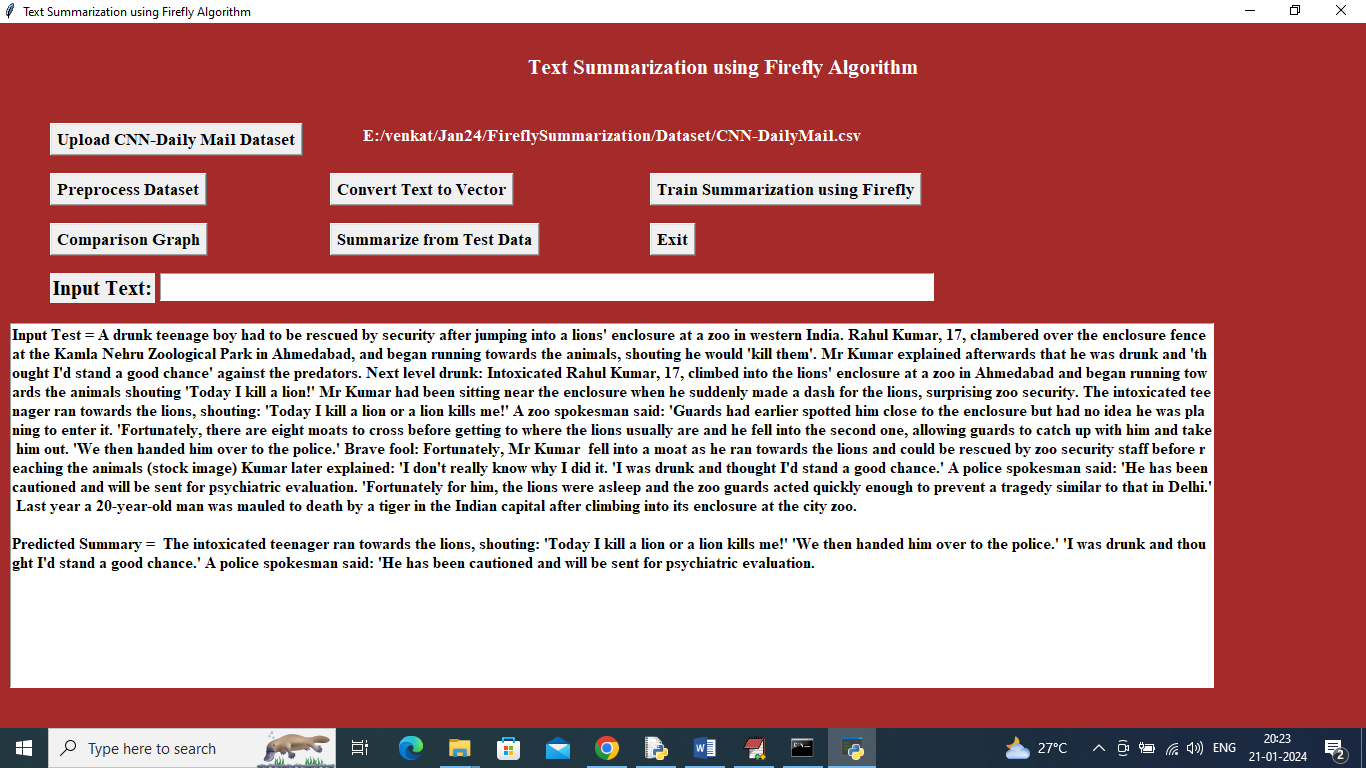
In above screen in text field I entered some Text data and then press ‘Summarize from Test Data’ button to get below output



In above screen first paragraph is the input sentences and second paragraph is the ‘Predicted summary’ and similarly you can enter some paragraphs and get summary or you can copy text from ‘testData’ folder. Below is another example



In above screen can see summary from another input



In above screen can see summary from another input