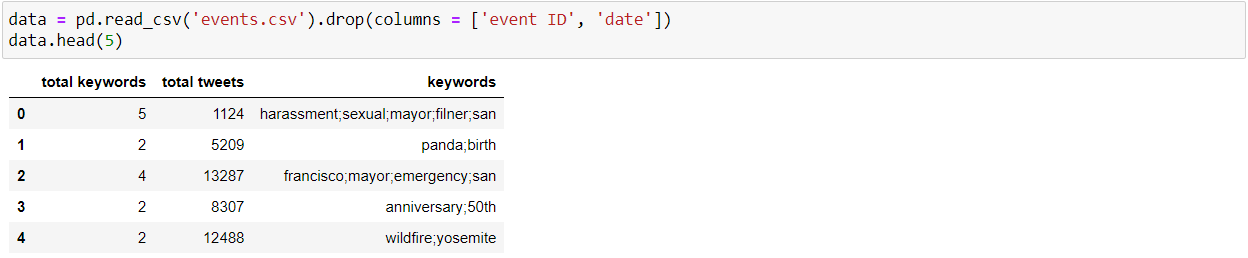
Shayma Report

# Code Progress

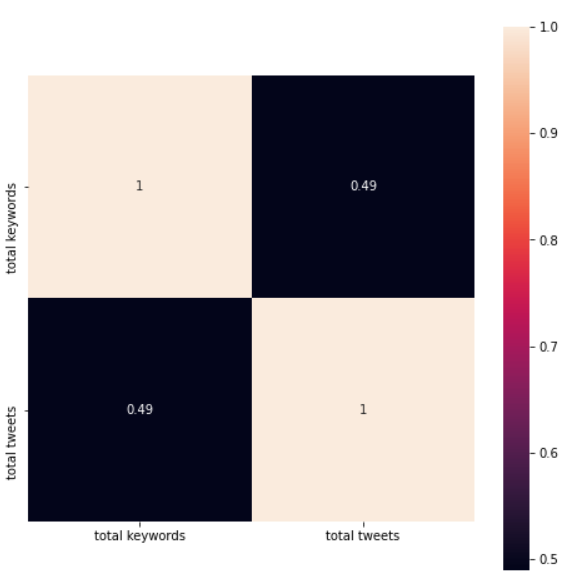
The code for this project is tasked with performing two different and crucial processes for the research. These processes are as follows:

## Data Pre-processing

This section performs an in-depth analysis on the dataset. The in-depth analysis includes exploring the dataset’s metadata to study about the properties of the dataset and the features in it. Then comes the data cleaning part where all the duplicate and null instances are removed from the dataset. Care needs to be taken here since carelessly removing instances from the data can result in the loss of the quality of prediction and the integrity of the dataset as well.



The figure above shows the importing of the dataset into the code. It can be seen that the features ‘event\_id’ and ‘date’ have been removed since they do not contribute to the objective of this research. To understand how each feature affects the dataset, a correlation matrix is plotted.



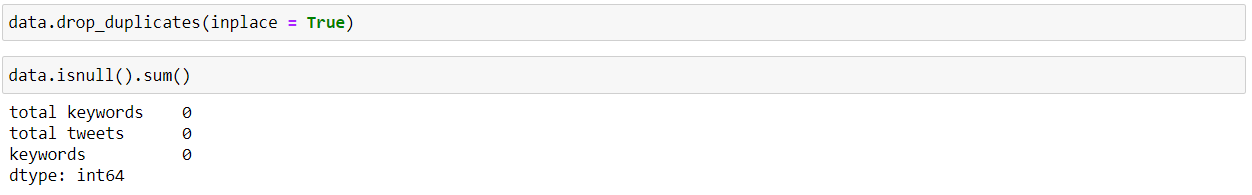
The correlation matrix in the figure above is designed as a heatmap. To features in the dataset currently have good correlation values and hence are important to the dataset. A correlation matrix is a table that shows the dependence of each input feature in the dataset with other variables. A correlation matrix is used to measure the strength of the relation between the features in the dataset. The greater the value of the correlation between the features, the better the quality of the dataset and better the results from the Machine Learning Algorithm.

The issue is that correlation matrices do not work with categorical data, that is why there are only 2 features dealt with in the correlation matrix instead of three. Since this research is using a Natural Language Processing approach for now, there is no need of encoding the categorical data.

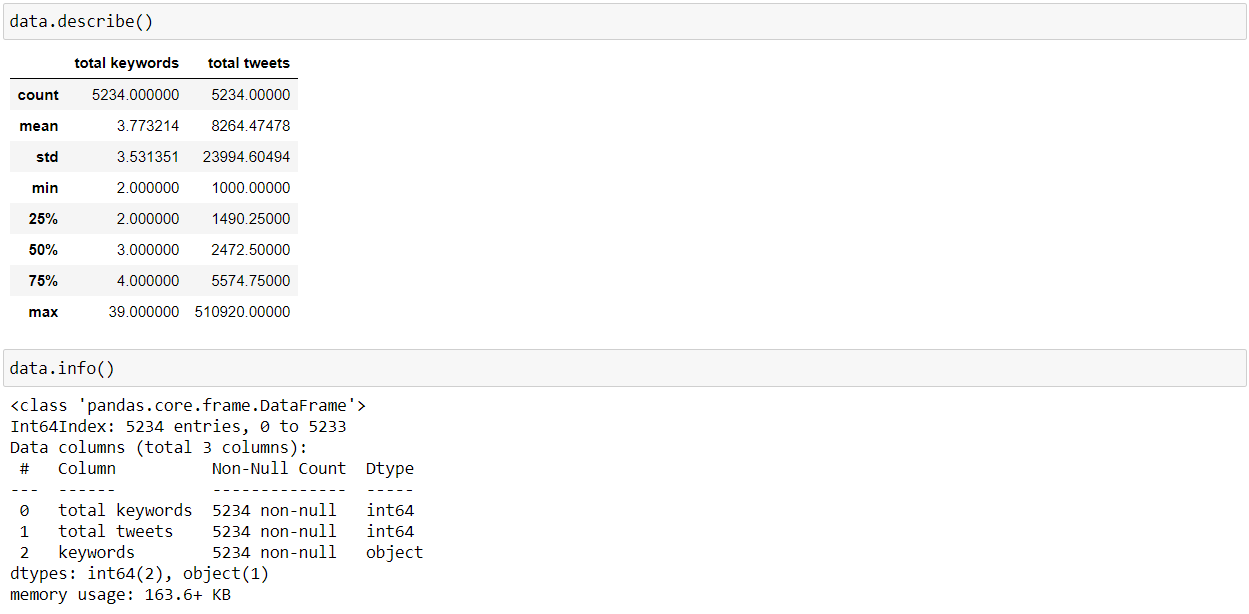
The use of “Total Keywords” feature in the correlation matrix is to just convey the quality of the dataset used in this research by showing the strength of the correlation between the numerical features in the dataset.

The “Date” feature from the dataset is removed since the research doesn’t have any use for it. The approach used in this research is the NLP approach in which keywords related to the project domain will be extracted from the categorical features in the dataset i.e., from “Keywords” feature. These keywords also known as the “bag of words” will then be processed using spell correction and stemming and then used by a Naïve Bayes algorithm to identify the topics. There is no need for the “Date” feature since it doesn’t contribute to any of the above-mentioned processes and also date cannot be used for data visualization since the data is about keywords and their frequencies.

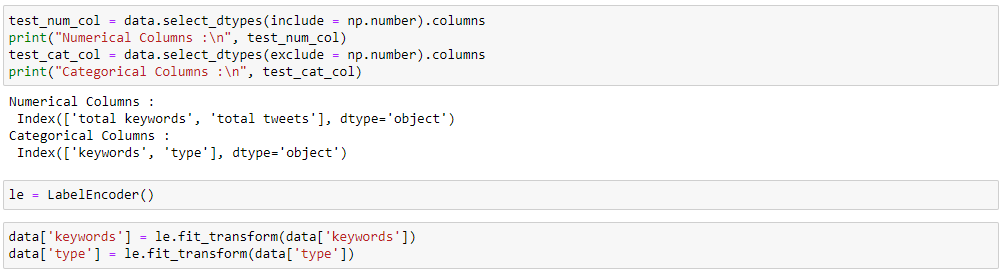
My findings from the correlation matrix convey that the correlation between the features in the dataset is positive and strong. This implies that all the features in the dataset contribute to the dataset and are an important part of the dataset.



The next step is described in the figure above where the dataset is rid of all the duplicate and null instances. There are no more duplicate and null values in the dataset. The next step would be to understand the dataset from a statistical point of view and also look at the metadata of the dataset.



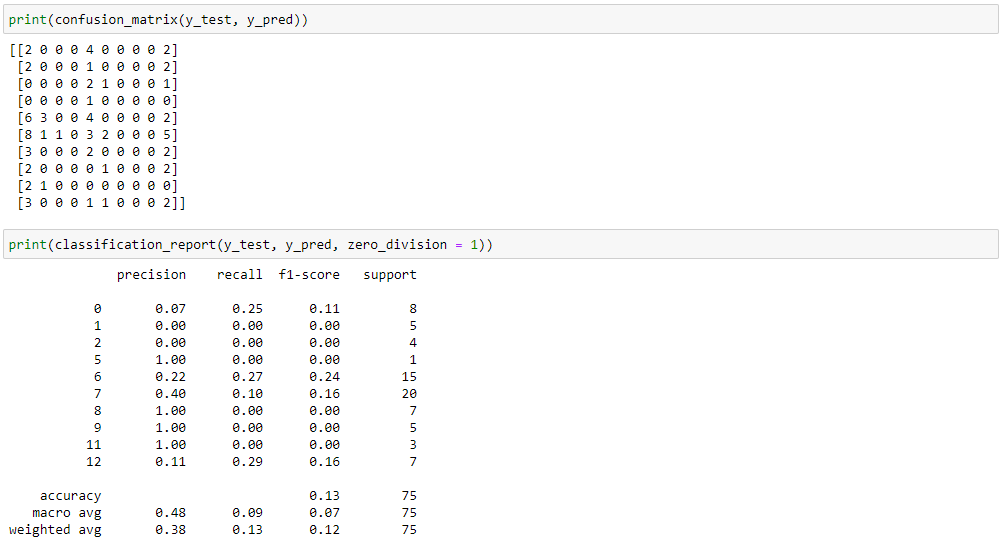
The central tendency measurements for each feature in the dataset along with the datatypes and the number of non-null values in each feature are seen here. They provide good insight into understanding the dataset. The dataset has the “keywords” feature which is categorical in nature. This needs to be encoded for the classification approach employed in this research. The encoding is done using Label Encoding method. The code below is the code responsible for Label encoding the dataset.



This completes the data pre-processing part.

## Machine Learning

The classification paradigm is used for this project. To that extent, the SVM and the MLP algorithms are applied in comparison to see which algorithm performs better and to understand how classification works on this dataset. The figure below shows the results for the SVM algorithm. The accuracy is only 13% which is very subpar. The reason behind this maybe the inherent bias or the number of output classes in the dataset. Or it could be the lack of parameters in the dataset.



Like the SVM, the MLP also performs similarly and provides very subpar results as shown below giving an accuracy of only 9%. This can only mean that the dataset does not possess enough parameters and there is a lot of inherent bias in the dataset.

