Approach

In the Dataset, the important columns are urlid, boilerplate, label. In the Column Boilerplate each row is in JSON Format and consists of the Title of the page and the Body of the page, so as it is an important Column to classify the pages. So, I converted that Column into a DataFrame. The DataFrame consists of 4 Columns namely, URLId, Title, Body and Label from the Train Data. The Column urlid refers to the id of the URL of each row so I used it in the dataframe, so that each title and body could be mapped to that particular url referred by the ID.

As we know that Label is a Dependent variable, I assigned it to a variable 'y'. Then I used word tokenizer to split the sentence into words for both train data and test data. I then reduced those words after tokenizing into their stem words using PorterStemmer. As I got the stemmed words, I converted those text into Feature Vectors using TfidfVectorizer. I converted those Feature Vectors into Sparse Matrix and called Logistic Regression to classify the Page as "Ephemeral" or "Evergreen".

Conclusion

1. Model Accuracy is about 92%.
2. Model is neither overfitted nor underfitted.
3. The Precision for Evergreen (1) is 0.94 and for Ephemeral (0) is 0.89
4. The Recall for Evergreen (1) is 0.90 and for Ephemeral (0) is 0.93
5. The ROC and AUC curve is 97.3%

References

- https://github.com/NirantK/nlp-python-deep-learning
- LetsUpgrade AIML Course