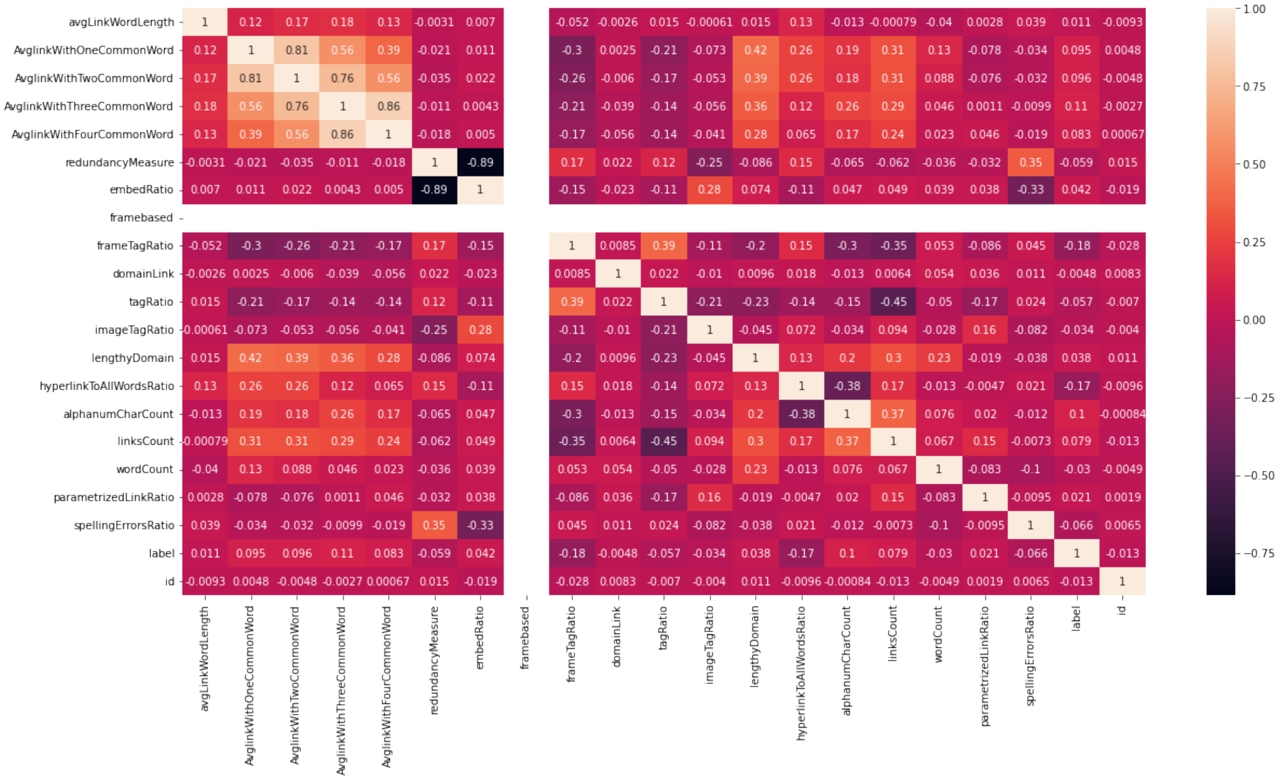
Project Definition

The task is to classify a website as either “relevant” or “irrelevant”, based on attributes such as alchemy category and its score, meta-information of the web pages and a one-line description of the content of each webpage.

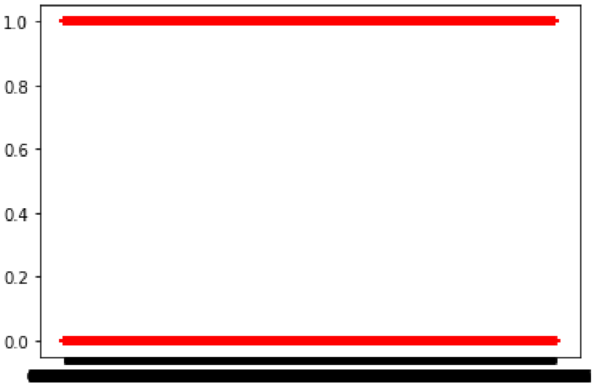
EDA

Correlation matrix for the given features :

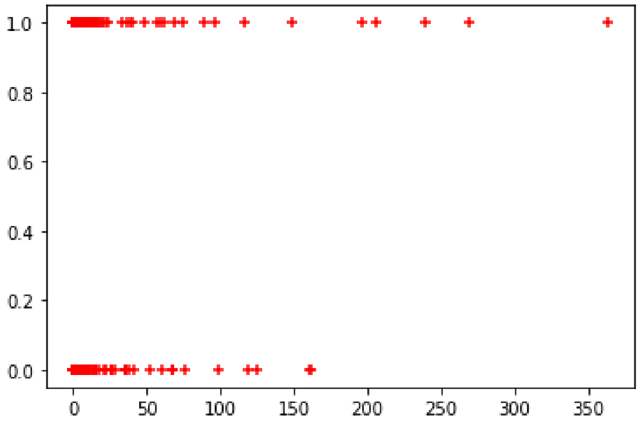


Scatter plots for all the features :

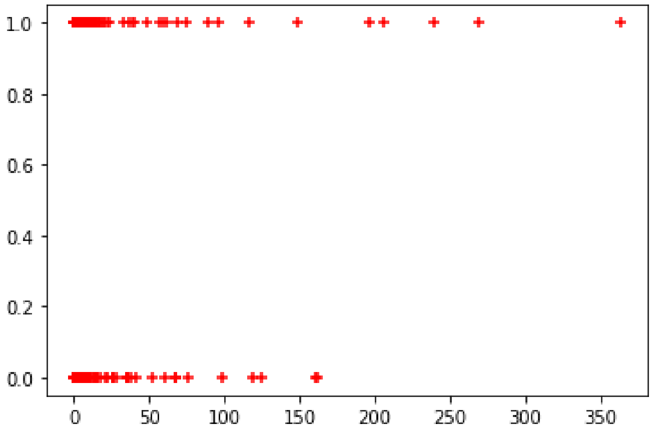
Alchemy_category_score :



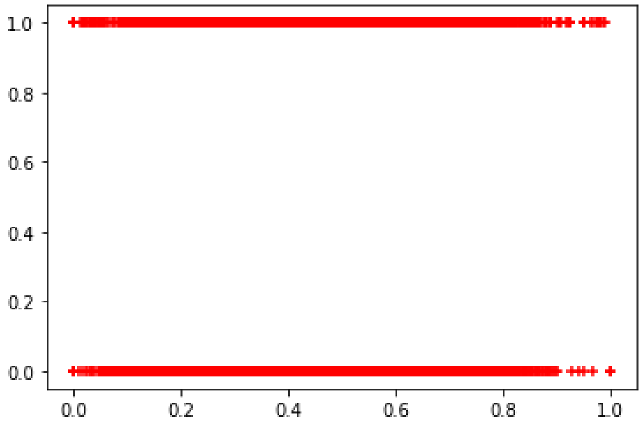
AvgLinkWordLength :



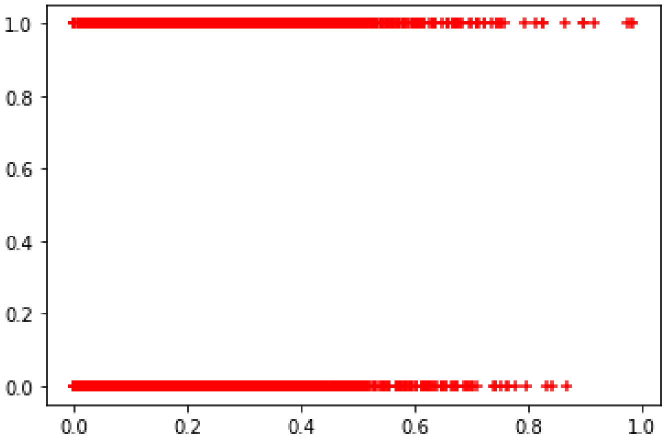
AvglinkWithOneCommonWord :



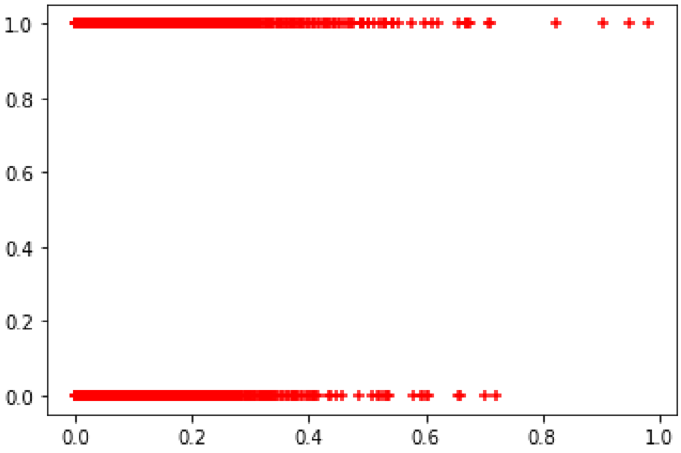
AvglinkWithTwoCommonWord :



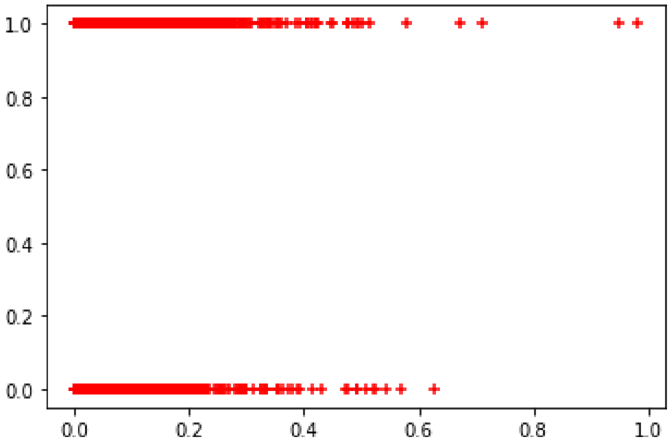
AvglinkWithThreeCommonWord ：



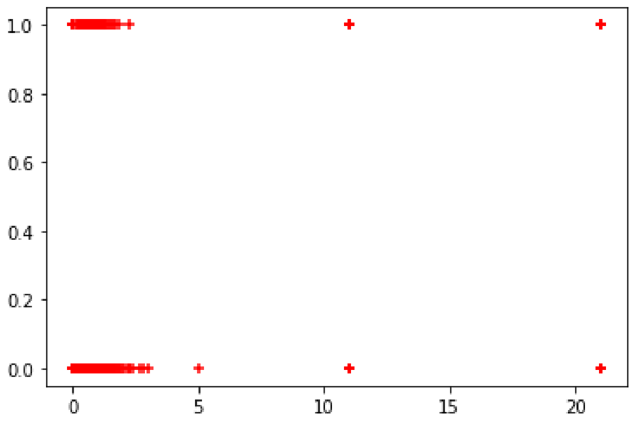
AvglinkWithFourCommonWord ：



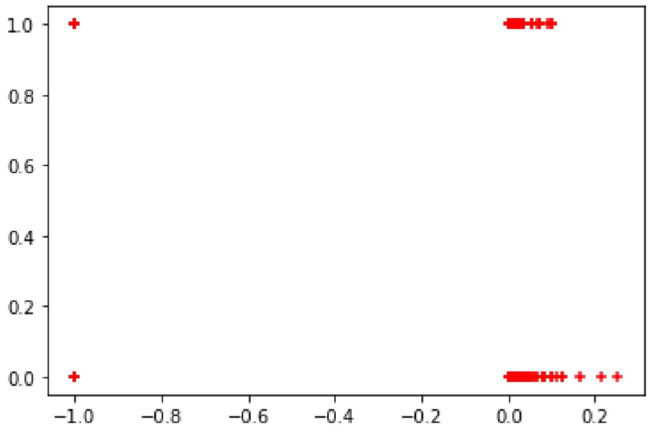
RedundancyMeasure ：



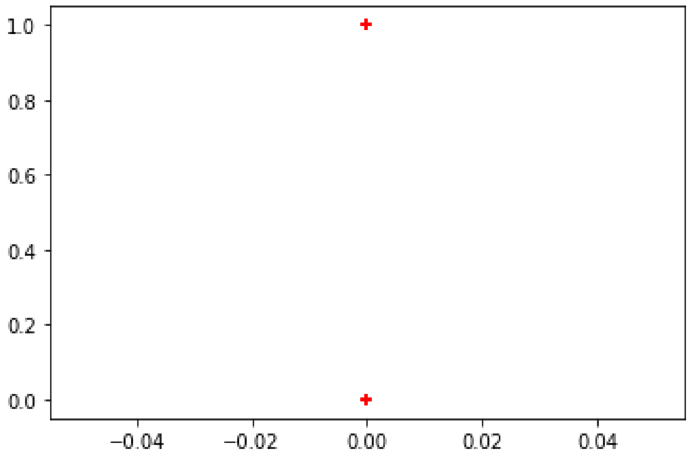
EmbedRatio :



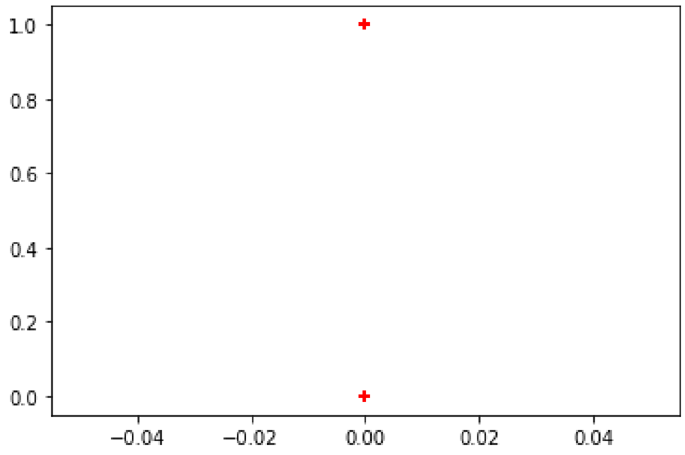
Framebased :



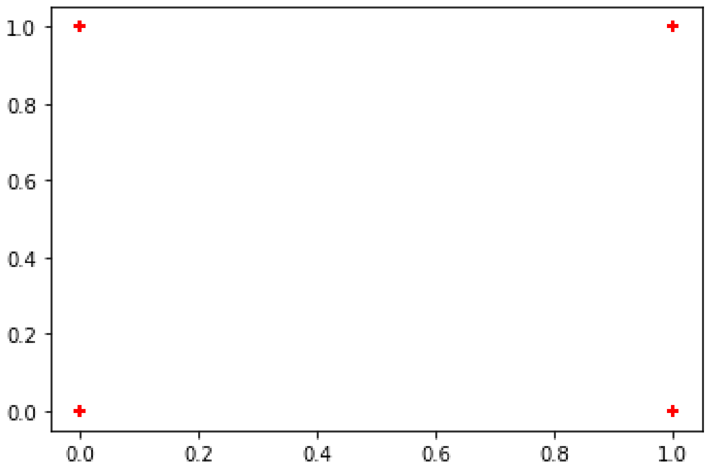
FrameTagRatio :



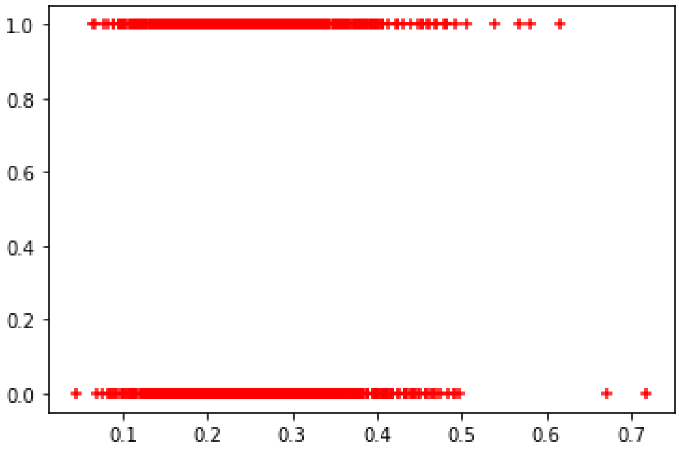
DomainLink :



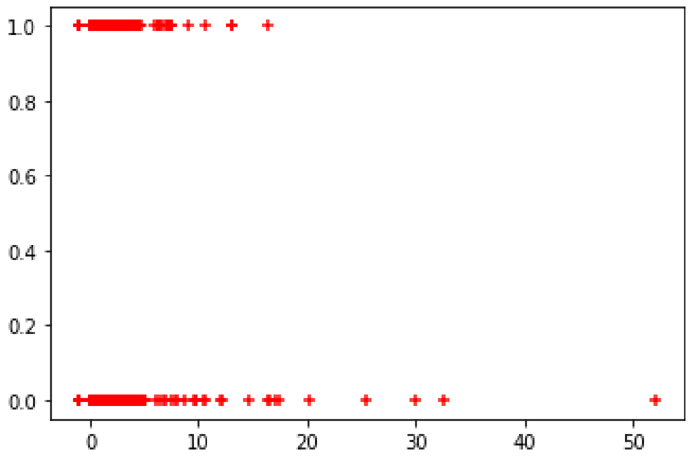
TagRatio :



ImageTagRatio :



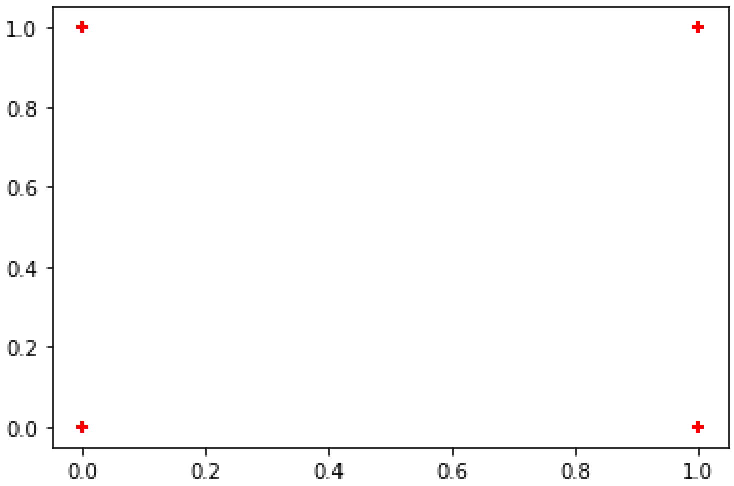
IsNews :



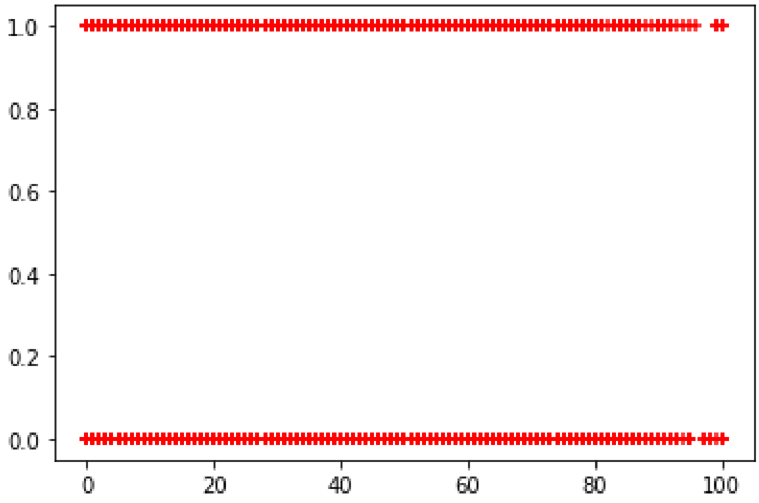
LengthyDomain :

1.0
0.8
0.6
0.4
0.2
0.0
?
1

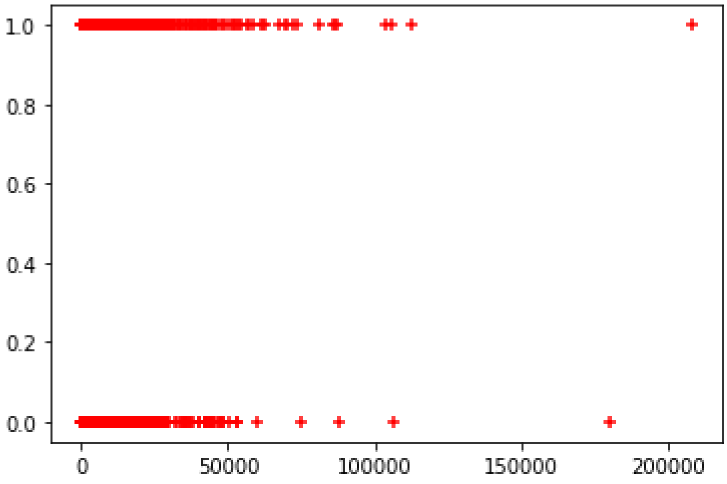
HyperlinkToAllWordsRatio :



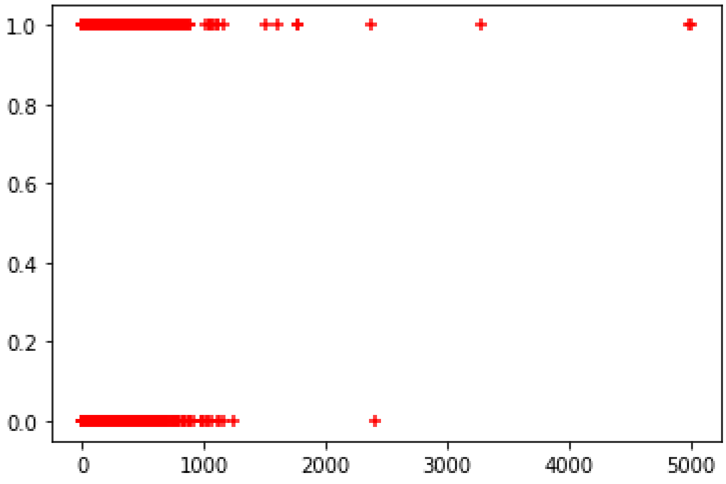
IsFrontPageNews :



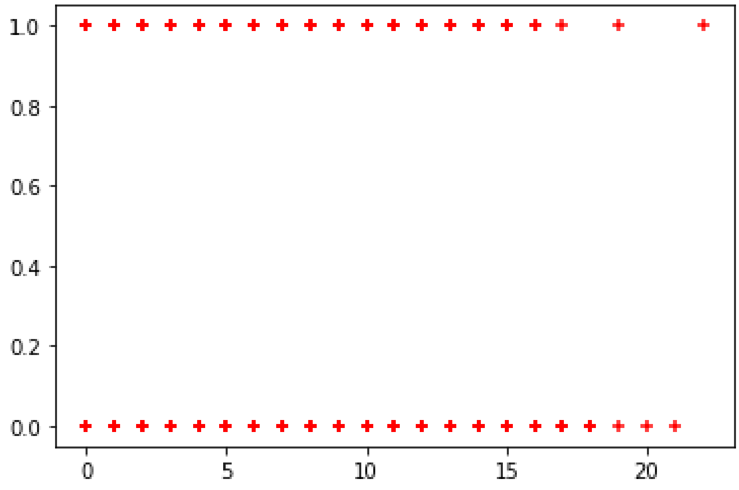
AlphanumCharCount :



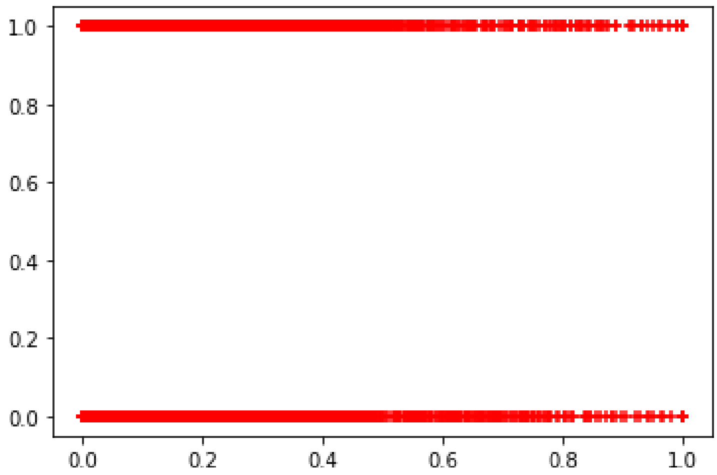
linksCount :



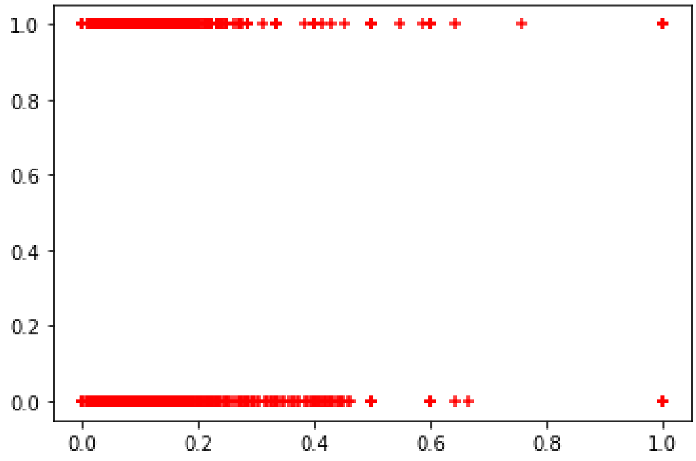
WordCount :



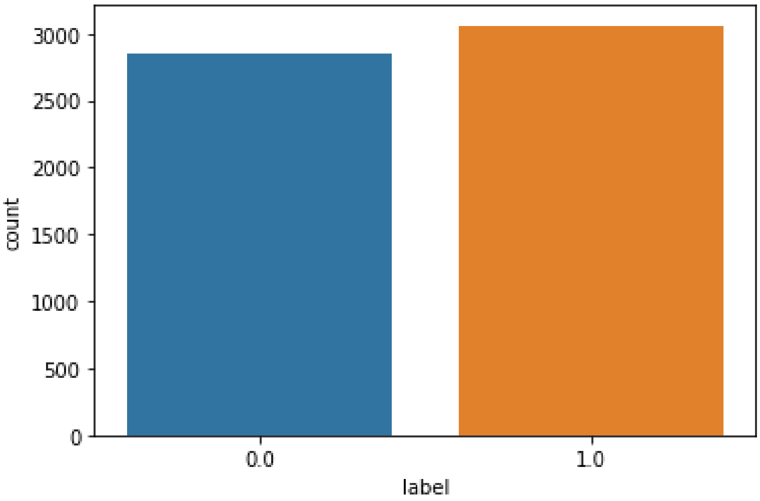
ParametrizedLinkRatio :



SpellingErrorsRatio :



Count plot for label :



Pre processing and feature selection :

Four features had missing values (?) Alchemy_category, alchemycategoryscore, isNews and isFrontPageNews.

Alchemy_category is in str format. The missing values in this column were changed into "Unknown" as it is difficult to classify them into any other field.

Alchemy_category_score is in int format. The missing values were replaced by the mean of the remaining value.

The isNews feature of the dataset is in integer format. The missing values were replaced by 0 as 1 represented the webpage that are news pages. The logic behind this was such that if a page is not a news article, it cannot be a news page

The isFrontPageNews feature was in integer format. The logic involved in solving this is similar to the isNews feature. The missing values were replaced by 0.

There are two columns in the dataset that are in object format. URL and webpage description.

The url column was dropped as every webpage has a different url and that was considered as an important feature.

NLP was performed on the webpage description column.

"isFrontPagenews" contains most of the values as zeros so we ignored the column.

pd.dummies is applied on alchemy_category column and countplots are plotted for each column.

We excluded all the features except the web page description in our model as we got the best score with only the web page column.

If we observe the scatter plotts and correlation matrix we can see that the feature except were not important cant affect the score much.

Experiments conducted and challenges faced :

We have applied NLP on the “web page Description” column and trained the data using 8 models .

We tried applying both count and Tfidf vectorizer but memory occured so we used Tfidf vectorizer only with parameter min_df=3.

Memory errors occured while applying SVM,Bagging models so we used applied PCA on the data and applied the models.

Models Used :

Logistic regression ,XG Boost,Adaboost 0.8368868617729972,Decision tree,Random forest,Bagging classifier, PCA + SVM.

Table of Models and their scores :

Logistic regression	0.88020 (kaggle)
XG Boost	0.8536759702422817
Adaboost	0.8368868617729972
Decision tree	0.7339978642375586
Random forest	0.8268697210998299
Bagging Classifier	0.4718601861858771
PCA + SVM kernel = rbf	0.5174227608803801
PCA + SVM kernel = poly	0.7626194397268968

Individual contributions.

We have done it together.

Conclusions

References

https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning d44498845d5b

https://towardsdatascience.com/text-classification-in-python-dd95d264c802 see m