Eutopia Task - Davide Garbelotto

The task was to classify the use of AI by some companies, given in input the text scraped from their websites.

The basic idea was to check the presence of some AI relevant keywords and their occurences inside the webpages and training a model able to classify the two cases.

The different fields inside the webpages have been merged in a unique field, assuming that the keywords don't depend on the part of the website where they appear.

After building a vocabulary from all the samples, two approaches have been implemented to extract meaningful features from the text:

1. The first method (not reported in the code) was to encode each website through a one-hot-encoding of the words appearing with respect to the vocabulary, i.e. storing information only about which word appears.
2. The second approach, which brought to slightly better results, was to encode also how many times each word appears in each website.

The variable encoding took more than 5 minutes, therefore the variables have been saved and can be uploaded directly in the notebook at the subtitle ‘Data Encoded’.

Some of the techniques which have been implemented are:

- **Feature engineering**: encoding, for example, whether in a sample the word 'ai' was appearing more than a certain number of times
- **Feature selection**: only the features which were more correlated to the target variable have been kept
- **Outlier detection**: the samples which didn’t contain at least 10 words between the ones selected (i.e. the ones which didn’t have at least 10 non-zero features) have been removed
- **Rescaling**: the data have been rescaled inside the range [0,1]
- **Grid search**: the model hyperparameters have been optimized through a grid search approach
- **Confusion Matrix**: the confusion matrix has been inspected to study the rate between the false negatives and the false positives.

The model selected was a Random Forest Classifier, which is a very powerful ensemble algorithm based on Decision Trees.

The algorithm reached an average classification accuracy of 87% on the test set.

A further approach which can be implemented to study deeper the dataset could be to build word embeddings through deep neural networks. This would likely bring to better results, but requires a higher amount of time to be implemented.

18 December 2020

Davide Garbelotto