Text classifier

Jacek Multan, index: 248964

June 2022

1 Data representation

For data representation BoW (Bag of Words) was used. It contains information about most used words in texts. To make it work properly data preprocessing was necessary. Firstly, all special characters were replaced by a space. Dog's is the same word as dog. Secondly, all one letter long characters were deleted as after replacing special character there were many single letter words that didn't mean anything. In the following steps multiple spaces were replaced with just one and all capital letters were changed to lower case. "Dog" is the same word as "dog" so we want to treat it as the same. Finally lemmatizer was used to convert words like "dogs" into "dog". As the last step vectorizer was used to count words in the documents. It was made to remember 2000 words that were used in maximum of 80% documents with minimal occurency of 5. At the end it was normalized using Term Frequency-Inverse Document Frequency.

2 Initial classification experiment

Data was split into train and test sets. Train set contained 70% of initial text files and test set 30%. During the experiment Random Forest Classifier was used. It is made out of set of decision trees. Initially it was used with n_estimators parameter set to 1 and random_state=0. The accuracy of the trained model was 0,70. Precision and recall of classes were as follows:

class	precision	recall
business	0.58	0.64
entertainment	0.60	0.68
politics	0.77	0.72
sport	0.81	0.82
tech	0.81	0.82

3 Optimization attempts

During the optimization the same classifier was used. Only the n_estimators parameter was changed. For n_estimators equal to

- 10 accuracy was equal to 0,91
- 50 accuracy was equal to 0,962
- 100 accuracy was equal to 0,968
- 500 accuracy was equal to 0,964
- 1000 accuracy was equal to 0,966
- 2000 accuracy was equal to 0,966

Accuracy was the highest for n_estimators=100. Recall and precision for this parameter were as follows:

class	precision	recall
business	0.93	0.97
entertainment	0.99	0.99
politics	0.98	0.93
sport	0.98	0.99
tech	0.97	0.96

3.1 Conclusion

Model recognizes entertainment and sport without much trouble. It sometimes confuses sport with politics. Tech category is recognized a bit worse than sport but it has pretty high accuracy anyway. The best results were achieved by setting n_estimators parameter as 100 so not too high but not too low.

4 Conclusion

- For text classifier method of representation and preprocessing of the data is very important. Bag of Words takes into account only number of words used and doesn't worry about semantics. That's why it's necessary to convert similar words into one (for example plural into singular).
- Random forest classifier works well in classifying texts into different categories. It must have well processed data and tuned parameters to do so. The highest accuracy achieved during experiments was 0,968 and it's quite high. Merging business and politics into one category and then using next classifier to distinguish between them could be good idea.