Language Identification

Dataset used: https://downloads.tatoeba.org/exports/sentences.csv
The dataset contains sentences with language labels 300+ languages. For the purpose of our project, only 8 languages were chosen due to limited computational power. 10,000 (if available) sentences of each of these 8 languages were extracted for this task and split into train, validation and test sets. The languages chosen are German, English, French, Spanish, Italian, Portuguese, Arabic and Bengali. Arabic was specially chosen to see how the algorithms employed worked with a language that is written from right to left.

Preprocessing: All the sentences were vectorized using bigrams. The bigrams were extracted from all the sentences and a subset of these bigrams were used to create the feature space. For the training, test and validation sets, the feature matrix was created by using the frequency of these bigrams in those sentences. In the feature matrix, the rows denoted the sentences, the columns denoted the unique bigrams. The cells in the matrix represented the presence or absence of these bigrams. There were articles using trigrams but our models performed better using bigrams.

Models:
Three models were attempted for this task.

- ANN - Variations of ANN were trained varying number of epochs, batch size and number of nodes to find the ANN model with the highest accuracy.
- CNN
- Multinomial Naive Bayes

Results:

Model	Accuracy
ANN - [200,200,100] ; epoch = 2; batch = 10	99.31
CNN	99.45 (mean across 10 repetitions)
Multinomial Naive Bayes	99.35

All the models perform similarly well. By a very margin, the CNN model performs best followed by MultiNB followed by the best performing ANN.

Future Scope:

1. Fine-tune the parameters for ANN.
2. Try more dense and complex CNN networks.
3. More complex sentences which are a mix of two languages or more can be used for stress- testing these systems.

Resources:

- https://ieeexplore.ieee.org/document/6854622
- https://www.aclweb.org/anthology/W17-5043.pdf
- http://cloudmark.github.io/Language-Detection/
- Blogs on https://medium.com/