Language Identification for Text Documents

**Problem**
- Problem definition: Identifying the language of short documents within 1-2 sentences in a group of languages.
- Application areas:
  - New cloud machine translation in social media
  - Improving search relevancy
- Data description:
  - Sentence-level data with an approximate English sentence length (approx. 15 words)

**Data Visualization (t-SNE plots)**
- Clusters of highly similar languages
- Clusters of less similar languages
- More interesting visualizations are available at: http://SeeYourLanguage.info

**Baseline Results**
- Logistic Regression was used to set a baseline performance.
- Character n-gram models were used as out-of-the-box baselines.
- The n-gram model was trained on a dataset of approximately 200K sentences.
- The n-gram model achieved an accuracy of 90.7% on the development set.

**Recurrent Neural Networks (RNN)**
- A recurrent neural network (RNN) was trained on the dataset to identify the language of short documents.
- The RNN architecture used was a 1-layer Long Short-Term Memory (LSTM) network.
- The RNN was trained on a dataset of approximately 200K sentences.
- The RNN achieved an accuracy of 91.5% on the development set.

**Results**

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy on Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN (word 2-gram)</td>
<td>0.9119</td>
</tr>
<tr>
<td>RNN (char 8-gram)</td>
<td>0.9505</td>
</tr>
<tr>
<td>LR</td>
<td>0.9425</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy on Validation Data</th>
<th>Accuracy on Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN (char 2-gram)</td>
<td>0.9123</td>
<td>0.9112</td>
</tr>
<tr>
<td>RNN (char 6-gram)</td>
<td>0.9208</td>
<td>0.9333</td>
</tr>
<tr>
<td>RNN (char 4-gram)</td>
<td>0.9377</td>
<td>0.9427</td>
</tr>
<tr>
<td>RNN (char 5-gram)</td>
<td>0.947</td>
<td>0.9314</td>
</tr>
<tr>
<td>RNN (word unigram)</td>
<td>0.9351</td>
<td>0.9209</td>
</tr>
<tr>
<td>Ensemble of RNNs (LR)</td>
<td>0.9537</td>
<td>0.9512</td>
</tr>
</tbody>
</table>

**Error Analysis**
- The error analysis shows the distribution on the validation set.
- The error analysis shows the distribution on the test set.

**Failure modes in classifying the (bs, hr, sr) language group**
- The LR classifier performance was evaluated for the (bs, hr, sr) language group by looking at increasing fraction of the documents in the samples.
- The classifier's performance was evaluated on the validation set.
- The classifier's performance was evaluated on the test set.

**Made possible by**
- AWS Educate
- Microsoft Azure

---

*Training gradients from batches of a document and adding a middle prediction could significantly improve the classification.
- Removing question from documents will improve the robustness of the classifier.*