DeepNews: Scoring News Articles By Quality
Susannah Meyer (smeyer7) and Harper Carroll (hcarroll)
CS230, Spring 2018

Abstract
In the age of digital consumerism, consumers are fed countless news stories with no baseline for evaluating which articles are worthy of their attention. The news feeds of the public are constantly bombarded with shallow, copy-and-pasted news articles with the intention of driving up traffic for publishers’ revenue.

We investigated how to assign news articles scores based on their quality in order to identify articles of high value that contribute more uniquely to a consumer’s new experience. We built an RNN model with a single-layer LSTM unit and a fully connected layer to assign scores between 0 and 1 to a dataset of news articles.

We found that such a model does a decent job of scoring articles given a similarity threshold of 0.5. However, there is room for improvement likely due to a number of constraints relating to data, labels, and model complexity.

Data

Data Collection: Our dataset consists of ~65,000 articles from the Dallas Morning News with associated scores as listed below. Our dataset is courtesy of UN Fellow Frederic Filoux, and score labels were aggregated by Matter Economics.

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>URL</td>
<td>Score averaging all scores below</td>
</tr>
<tr>
<td>Reach Score</td>
<td>Volume of page views, non direct and non-referrer page views</td>
</tr>
<tr>
<td>Quality Score</td>
<td>Average scroll depth, average time per page</td>
</tr>
<tr>
<td>Audience Score</td>
<td>Proportion and volume of: known and local page views, direct and interstitial referer page views, page views from users in the top 2 engagement buckets</td>
</tr>
<tr>
<td>Yield Score</td>
<td>Volume of ad revenue, conversions from an article, pages on the path to conversion</td>
</tr>
</tbody>
</table>

Labeling: We used the Overall Score listed above as the ground truth for the score of an article and transformed scores to be confined within a range from 0 to 1. It should be noted that the above scores do not correlate with text quality but rather a number of user engagement metrics. Our model sought to test whether such scores could be reliable in scoring text quality.

Model

Processing:
We extracted the main text from each given article URL and tokenized this main text in order to transform input articles to lists of token word IDs.

Word Representations:
We used GloVe’s 100-dimensional word embeddings to represent the list of word IDs for each article into embedding vectors for each word. The features for each input to our model is comprised of the article’s word IDs and the embedding vector for each word.

Model architecture:
We implemented a TensorFlow dynamic RNN model with a single LSTM layer with 100 hidden units and a fully connected layer with a sigmoid activation.

Performance:
A single example is classified as correctly scored according to $|y - y| < 0.1$

Results

Our training, dev, and test sets consist of a randomized shuffle of articles from our entire dataset.

- Training set article count: 35,000
- Dev set article count: 10,000
- Test set article count: 5,000

Our results show that our model outputs a high performance rate on the training set and outputs lower performance rates on the test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training error</th>
<th>Test error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.066</td>
<td>0.256</td>
<td></td>
</tr>
</tbody>
</table>

Discussion

Our results show that our model outperforms the training set and leaves room for improvement in terms of performance for the test set. We have drawn a number of interpretations from these results.

First, the extraction of main text from the URLs of each article in our dataset relied on the Python Scrapy Extractor library to parse through the HTM of each article and keep only plain text within the article’s body. The extractor was unable to parse articles which fully relied on multimedia, and it was impossible for us to manually verify the parsing of those articles for which extraction succeeded. This might lead to unstable results.

Next, the dataset of our model provided scores related to user engagement as our ground-truth labels. While our model was trained on the training data, our results might show that extracting features using only an article’s text from a natural language processing standpoint are inadequate for predicting such scores.

Finally, the use of regularization techniques and the addition of layers to increase model complexity might reduce overfitting.

Future Directions

There are a number of significant applications that our project might be extended to explore.

First, an improved model might take advantage of human-labeled quality scores for better ground-truth labeling as opposed to the use of Matter Economics user-engagement scores.

With functioning quality scores, we would propose integration into applications including smart advertising to match the revenue of advertisements with the quality of an article, as well as integrating personalized news experiences based on qualities and ensuring smart curation of news aggregators to remove low-quality sources.