Objective

The goal is to analyze sentiments of a text or a corpus based on the context information.

Background

Broadly there are three ways to perform Context Aware Sentiment Analysis, Dictionary Based Approach and Corpus Based Approach.

• Dictionary Approach: a target word is identified and then compared with words from general purpose lexicon for similarity or dissimilarity and the polarity is assigned to the word. This approach fails for domain specific analysis.

• Corpus Approach: the basic idea of this approach is to find the similar relatedness of the words and then decide the sentiment for the target word. To find the similar relatedness word co-occurrences, contextual similarity can be used.

• Machine Learning Approach: Doc2Vec, a shallow NN for CASA.

Doc2Vec

**Document to Vector (doc2Vec):** is a technique used to produce word embeddings. Doc2Vec is a Neural Network based approach and is a shallow network with two layers only.

1. Input to Doc2Vec: one-hot-encoding of the text.
2. Output of Layer 1: weight matrix representing the words.
3. Output Softmax Layer: produces an output in the range [0-1]

Goal is to learn the weights from layer 1 not the final output. Doc2Vec has two models DBOW and DM.

Data Preprocessing

The standard NLP preprocessing which includes

(i) Stemming
(ii) Removing Punctuations
(iii) Tokenize.

The Doc2Vec take care of Stop words automatically. Preprocessing by Doc2Vec.

(i) Making Bi-grams, treat ‘Boston Globe’ as single word.
(ii) Sub-Sampling frequent words.

Negative Sampling is used to speed up the process of learning optimal weights.

Solution

To analyze the sentiments using Doc2Vec as the base technique.

Model Evaluation

To evaluate the performance of the developed model **Accuracy** is used as the performance metric.

\[
Accuracy = \frac{(TN+TP)}{(TN+TP+FP+FN)} \tag{1}
\]

Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Doc2Vec</th>
<th>TF-IDF</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMDB</td>
<td>31.2MB</td>
<td>85%</td>
<td>74%</td>
<td>79%</td>
</tr>
<tr>
<td>Groupon</td>
<td>350KB</td>
<td>87%</td>
<td>83%</td>
<td>76%</td>
</tr>
<tr>
<td>Linkedin</td>
<td>15KB</td>
<td>84%</td>
<td>58%</td>
<td>78%</td>
</tr>
<tr>
<td>Yelp</td>
<td>31MB</td>
<td>81%</td>
<td>70%</td>
<td>76%</td>
</tr>
</tbody>
</table>

Conclusion

As it is evident from results Doc2Vec has surpassed all State-of-the-Art algorithms. The results might seem not that good, but, given the fact that Doc2Vec scales very good in comparison to others, results are way better.

Future Work

The overall performance of Doc2Vec can be improved further. Two obvious ways to do that.

⇒ Modify the Doc2Vec algorithm.
⇒ Improve the quality of data.

The standard procedure for data pre-processing using NLP and the ability of Doc2Vec to pre-process the data are limited. An intelligent approach to clean the data further can improve the overall accuracy more.

References