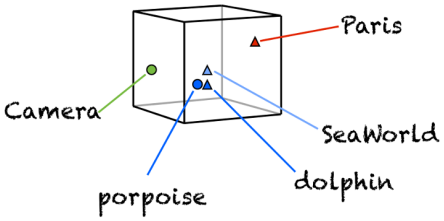


Evaluation of Methods for Generating Document Vectors Using Word Embedding

Jason Xie (jxieeducation@gmail.com)

Introduction

In this poster, we present three methods for creating document vectors that represent documents in dense high dimensional spaces with the property that similar documents have higher cosine similarities. We compare the effectiveness of these methods by training the resulting document vectors to perform the supervised task of sentiment analysis.

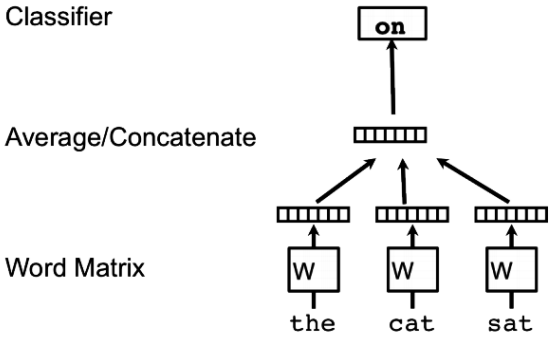
Process

Part 1. Construct word embeddings of words from 25,000 reviews in the IMDB movie dataset. Word vectors are created using Word2Vec (Mikolov et al., 2013) through skip-gram and hierarchical softmax.

Part 2. Construct document embeddings (described in detail in the methods section) from the word embeddings. We see documents as the sum of individual words.

Part 3. Perform supervised learning on the document vectors to predict whether IMDB movie reviews are positive or negative.

Word2Vec, the Shallow Neural Network



Methods

Method 1. Averaging Word Vectors

We first represent documents as the sum of its words. Here we naively average the vector representation of words in a document.

$$D_i = \frac{\sum_{j=0}^{N_i} W_{ij}}{N_i}$$

Method 2. TF-IDF Weighed Word Vectors

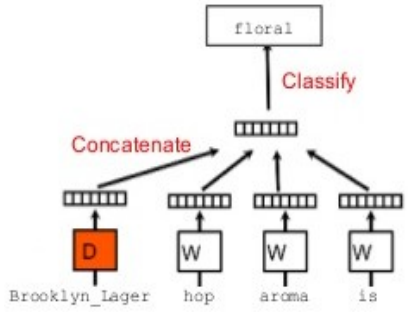
Instead of treating all words as if they are equal, we instead weight words based on their TF-IDF score. The intuition behind using the TF-IDF weights is to reflect how important a word is to a document in the document vector.

$$n_{ij} = tf_{i,j} \; x \; \log\left(\frac{N}{df_i}\right)$$

$$D_i = \frac{\sum_{j=0}^{N_i} n_{ij} W_{ij}}{N_i}$$

Method 3. Doc2Vec

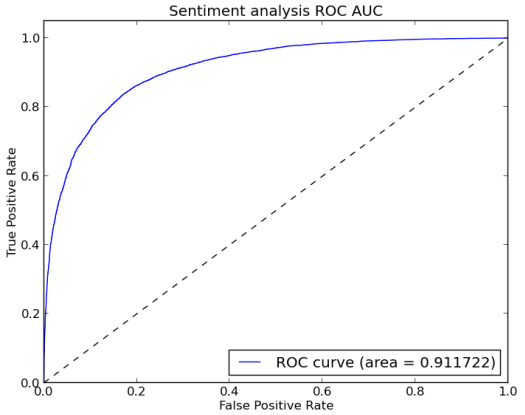
Doc2Vec (Le et al., 2014) constructs document vectors by including the document as a unique word during the Word2Vec training process.



Results

ROC AUC (%)	Logistic Regression (C = 0.7)	Random Forest (20 trees, 0.5 ft)	Neural Network (100x50x1)
Average	87.996%	81.527%	88.646%
TF-IDF	86.786%	81.524%	86.725%
Doc2Vec	90.761%	82.389%	91.172%

ROC AUC of Doc2Vec embeddings with a Shallow Neural Network



The results indicate that Doc2Vec is the best method to represent documents in vector space. This is unsurprising due to the fact that the document representations are learned explicitly during the back propagation process, instead of through mathematical derivation post hoc.

TF-IDF weighed word vectors performed worse than simple averaging. This is likely due to the fact that common sentimental words (e.g. happy, sad), which are important in sentiment analysis, are punished. On the other hand, the averaging method benefitted from the high number of words per review (239 words / review) and was able to better differentiate between positive and negative reviews.