Delta TFIDF: an Improved Feature Space for Text Analysis

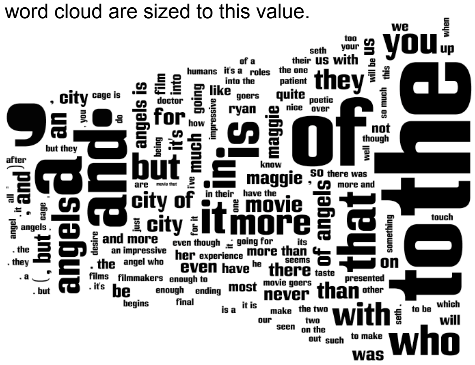
Justin Martineau, Tim Finin, Shamit Patel and Anupam Joshi
University of Maryland, Baltimore County

Bag of words text classification techniques represent documents as a collection of words with associated counts. Machine learning algorithms, such as support vector machines, use these feature spaces to classify new documents using their similarity to a set of manually labeled documents. We describe an efficient way to weight these feature scores to improve classification accuracy.

Representing A Movie Review for *City of Angels* as a Bag of Words

Raw Term Frequency Weights:

Most bag of words representations use the raw word count as the feature value. The words in this word cloud are sized to this value.

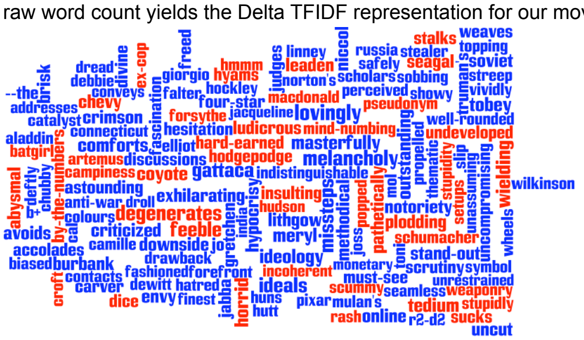


TFIDF Weights:

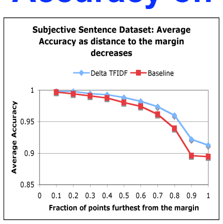
Raw word counts place too much emphasis on stop words. To correct, we can weight the raw score with TFIDF to boost the value of rare words in documents. Words are sized to their TFIDF value.

Delta TFIDF Weights:

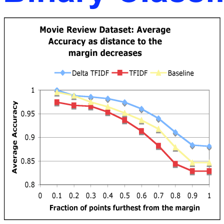
TFIDF values rare words, but sentimental words are common in sentiment analysis corpora. Delta TFIDF values words that are rarer in one type of labeled document than the other. Delta TFIDF calculates *inverse document frequency* (IDF) scores separately in the positive and negative training sets. Subtracting the positive IDF score from the negative for a word determines the sentiment orientation. The word cloud on the left shows positive words for movies in blue and negative in red. Word size indicates strength. Multiplying the Delta IDF score by the raw word count yields the Delta TFIDF representation for our movie review of the city of angels as depicted on the right.



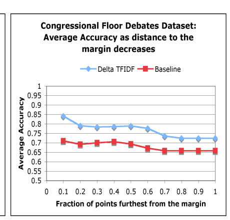
Accuracy on Binary Classification Tasks using SVMs:



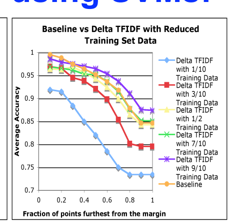
Distance from an SVM's margin is a good proxy for opinion strength and confidence. Delta TFIDF works on both subjectivity detection and sentiment orientation determinations.



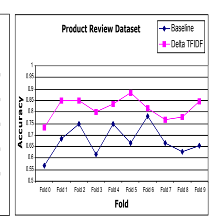
TFIDF underperforms the baseline because sentimental words are common in a corpus composed entirely of movie reviews.



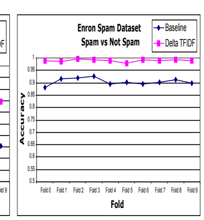
Congressional debate transcripts are more difficult than movie reviews. Party affiliation is a better indicator of how a congress person will vote than the content of their speeches.



Delta TFIDF using only half as much training data as the baseline method operates about as well on even the weakest sentimental expressions.



Delta TFIDF works on a variety of domains and is not limited to just movie reviews.



Delta TFIDF works on a variety of binary text classification problems and is not limited to sentiment analysis.

UMBC

AN HONORS UNIVERSITY IN MARYLAND