

**Fuzzy Bag-of-Words Model for Document Representation**

**Abstract:**

One key issue in text mining and natural language processing is how to effectively represent documents using numerical vectors. One classical model is the Bag-of-Words (BoW). In a BoW-based vector representation of a document, each element denotes the normalized number of occurrence of a basis term in the document. To count the number of occurrence of a basis term, BoW conducts exact word matching, which can be regarded as a hard mapping from words to the basis term. BoW representation suffers from its intrinsic extreme sparsity, high dimensionality, and inability to capture high-level semantic meanings behind text data. To address the aforementioned issues, we propose a new document representation method named fuzzy Bag-of-Words (FBoW) in this paper. FBoW adopts a fuzzy mapping based on semantic correlation among words quantified by cosine similarity measures between word embeddings. Since word semantic matching instead of exact word string matching is used, the FBoW could encode more semantics into the numerical representation. In addition, we propose to use word clusters instead of individual words as basis terms and develop fuzzy Bag-of-WordClusters (FBoWC) models. Three variants under the framework of FBoWC are proposed based on three different similarity measures between word clusters and words, which are named as FBoWCmean, FBoWCmax, and FBoWCmin , respectively. Document representations learned by the proposed FBoW and FBoWC are dense and able to encode high-level semantics. The task of document categorization is used to evaluate the performance of learned representation by the proposed FBoW and FBoWC methods. The results on seven real-word document classification datasets in comparison with six document representation learning methods have shown that our methods FBoW and FBoWC achieve the highest classification accuracies.