Information Retrieval Research Project

Author: Pathikrit Ghosh
Instructor: Dr. Rajendra Roul(Professor,BITS Pilani KK Birla Goa Campus)

Proposed Feature Fast selection technique :

feature selection of terms by variance(in their tf-idf values) and then by idf

Algorithm

1. Consider a given corpus consisting of number of classes of documents.
2. Preprocess the dataset by stemming/lemmatizing and eliminating stopwords.(stemming used here)
3. Perform document elimination by calculating the centroid of tf-idf vectors of documents and eliminating documents by using a threshold value.(done to remove outliers from a given class of documents).
4. Calculate tf-idf and idf values for each of the terms in the corpus.
5. Feature selection -
    1. Calculate the variance of tf-idf values of all keywords.
    2. Variance for each term is calculated as

$$\text{Var(X)} = \frac{1}{n}\sum_{1}^{n}\left(x_i - \mu\right)^2$$

where n is total number of documents, x_i is the tf-idf value of a term in ith document and μ is the average tf-idf value of the term in all the documents.

    3. The term with top K variance values will be selected as features.
    4. Terms are further eliminated by using their idf values. The term with top K' idf values will be selected from the previously selected features.
6. The final reduced feature vector of each class is then used to train Linear SVM classifier for text classification. Using the output prediction generated by a classifier and the known class label of the test data, calculate the precision, recall, F-measure and accuracy to quantify the performance of the classifier.

Experimental Results

Results were computed by varying two parameters K and K' for variance based feature selection and then further idf-based. Thus top M% results were computed as (K% x K'%). The threshold for document elimination was kept at 95%. The score given here is average f1-score(out of 100) among all classes.

Classic Dataset

	15x70	**25x85**	**45x90**	**60x85**	**80x90**	**90x90**	**90x100**
Chi2	86	92	95	96	95	95	95
Feature selection	94	87	88	88	88	86	95

WebKB Dataset(taking 500 documents from each class)

	15x75	**25x75**	**45x90**	**60x85**	**80x90**	**90x90**	**90x100**
Chi2	88	89	89	89	89	89	89
Feature selection	88	89	90	90	90	90	91