

UDC 004.4
Golub T. V.
assistant ZNTU

THE INFLUENCE OF THE TRAINING SET SIZE ON THE FEATURE VECTOR SIZE FOR THE TEXT DOCUMENT CLASSIFICATION

One of the factors of the text document classification process complexity is the size of the feature vector, which is used as input to assess the ownership of a category document.

The feature vector is defined on the basis of the training sample processing results presented in the text document form. The text in this case is processed by a number of preliminary steps, including the removal of stop words, stemming, and others. Then we weigh the obtained terms to evaluate their degree of importance.

Statistically, the size of the training set affects the size of the feature vector and, as a result, the complexity and quality of the classification. A feature of the prize vector for a separate category is that to classify a text document into this category, it is necessary to match the terms of the analyzed document with the elements of this vector. So, there are a finite number of key terms that describe a separate category. Therefore, a significant increase in the size of the training set may not provide a significant qualitative increase in the size of the feature vector.

Table 1 shows the dependence of the size of the feature vector on the size of the training sample separately for each category. The number of words in the training collection is indicated in the first column for each category. The resulting

size of the feature vector obtained as a result of the analysis of the training set presented in the second column separate for each category.

Table 1 - Dependence of the feature vector size on the training set size

1 category		2 category		3 category		4 category	
Word summa	Term vector	Word summa	Term vector	Word summa	Term vector	Word summa	Term vector
10336	2036	12475	1273	8629	2084	6715	1955
14138	2491	13859	1649	14004	2681	10392	2439
19973	3305	15760	1806	16238	2956	14648	2758
23207	3734	16743	1955	19947	3536	15243	2894
29333	4448	17459	2074	21641	3804	19427	3457
32982	4931	24714	2796	24017	4183	23158	3885

Figure 1 shows a plot of the normalized size of the feature vector for separate categories relative to the first category, taken as a reference, from the size of text documents of the training set. This calculation is presented in table 2.

Table 2 – Normalized values of dependence of the feature vector size on the training set size

1 category		2 category			3 category			4 category		
Word summa	Term vector	Word summa	Term vector	Norm. vector	Word summa	Term vector	Norm. vector	Word summa	Term vector	Norm. vector
10336	2036	12475	1273	1536	8629	2084	1740	6715	1955	1270
14138	2491	13859	1649	1616	14004	2681	2656	10392	2439	1793
19973	3305	15760	1806	1425	16238	2956	2403	14648	2758	2023
23207	3734	16743	1955	1410	19947	3536	3039	15243	2894	1901
29333	4448	17459	2074	1234	21641	3804	2806	19427	3457	2290
32982	4931	24714	2796	2095	24017	4183	3046	23158	3885	2728

As can be seen from Table 1 and Figure 1, the size of the training set (Word sum) increases and the increase in the size of the feature vector decreases (lines of 1-4 categories). This indicates the inexpediency of a significant increase in the size of the training sample to increase the set of the terms variety of the feature vector.

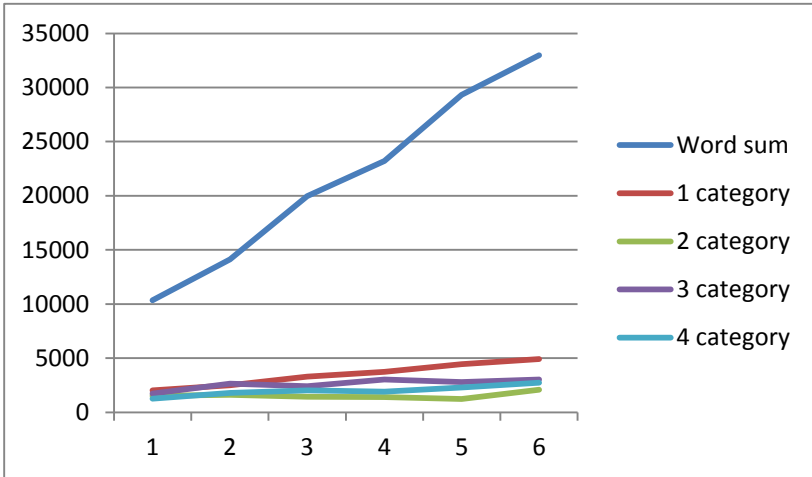


Figure 1 - Dependence of the term vector size on the learning set size

Therefore, further study of the determination of the sampling set optimal size is appropriate.