**64.Reuters RCV1 RCV2 Multilingual, Multiview Text Categorization Test collection**

1. 数据库网址

http://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2+Multilingual,+Multiview+Text+Categorization+Test+collection

2. 数据库描述

【1.[数据集名称]数据集由[机构名或人名]采集；】The data used in our experiments were collected by Massih-Reza Amini and Cyril Goutte, National Research Council Canada, Interactive Language Technology group.【2.用于[什么实验目的]】We used to classify the features from different languages.【3】This test collection contains feature characteristics of documents originally written in five different languages and their translations, over a common set of 6 categories. We focused on six relatively populous categories: C15, CCAT, E21, ECAT, GCAT, M11. For each language and each class, we sampled up to 5000 documents from the RCV1 (for English) or RCV2 (for other languages). Documents belonging to more than one of our 6 classes were assigned the label of their smallest class. This resulted in 12-30K documents per language, and 11-34K documents per class. In experiments that we conducted in cite{AUG09}, we considered each document available in a given language as the observed view for an example and all translated documents were used as the other views for that example, generated using Machine Translation. Results shown in this study were averaged over 10 random samples of 10 labeled examples per view for training, and 20% of the collection for testing.【4】The database has 111740 samples, respectively belong to training with 78218 samples and testing with 33522 samples. As shown in Table 1.

Table 1 Category Distribution of Dataset [根据数据库绘制]

|  |  |  |  |
| --- | --- | --- | --- |
| Categories | Training | Testing | Total Number of Samples |
| English | 13131 | 5627 | 18758 |
| French | 18654 | 7994 | 26648 |
| German | 20967 | 8986 | 29953 |
| Italian | 16827 | 7212 | 24039 |
| Spanish | 8639 | 3703 | 12342 |
| Total number of samples in total | 78218 | 33522 | 111740 |

3. 精简描述

The Data in our experiment were collected by Massih-Reza Amini and Cyril Goutte, National Research Council Canada, Interactive Language Technology group. The dataset includes have 111740 samples, which used to classify the features from different languages. Through which, we divided the dataset into two part, training data set with 78218 samples and forecasting data set with 33522 samples.