**61.Reuters Transcribed Subset**

1. 数据库网址

http://archive.ics.uci.edu/ml/datasets/Reuters+Transcribed+Subset

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 Shourya Roy and Shantanu Godbole.【2.用于[什么实验目的]】We used to distinguish the reuters’s text.【3】This dataset is created by reading out 200 files from the 10 largest Reuters classes and using an Automatic Speech Recognition system to create corresponding transcriptions. This data was created by selecting 20 files each from the 10 largest classes in the Reuters-21578 collection. The files were read out by 3 Indian speakers and an Automatic Speech Recognition (ASR) system was used to generate the transcripts. Such a dataset will be really helpful to study the effect of speech recognition noise on text mining algorithms. There are 10 directories labeled by the topic name. Each contains 20 files of transcriptions.【4】The database has 200 samples, respectively belong to training with 160 samples and testing with 40 samples. As shown in Table 1.

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

|  |  |  |  |
| --- | --- | --- | --- |
| Categories | Training | Testing | Total Number of Samples |
| R10 | 30 | 20 | 50 |
| R90 | 30 | 20 | 50 |
| 20NG | 30 | 20 | 50 |
| NB100 | 30 | 20 | 50 |
| Total number of samples in total | 120 | 80 | 200 |

3. 精简描述

The Data in our experiment were collected by Shourya Roy and Shantanu Godbole. The dataset includes have 200 samples, which used to distinguish the reuters’s text. Through which, we divided the dataset into two part, training data set with 120 samples and forecasting data set with 80 samples.