**80.Reuter\_50\_50. Authorship identification in online Writeprint which is a new research field of pattern recognition**

**http://archive.ics.uci.edu/ml/datasets/Reuter\_50\_50**

1. 数据库网址

http://archive.ics.uci.edu/ml/datasets/Reuter\_50\_50

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 E. Alpaydin, C. Kaynak, from Department of Computer Engineering,Bogazici University at July,1998.【2.用于[什么实验目的]】We used preprocessing programs made available by NIST to extract normalized bitmaps of handwritten digits from a preprinted form.【3】

【4】The database has 5620 samples, respectively belong to optdigits.tra with 3823 samples and optidigits.tes with 1797 samples. The categories of network system include seven categories, as shown in Table 1.

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

|  |  |  |  |
| --- | --- | --- | --- |
| Invasion Categories | optdigits.tra | optdigits.tes | Total Number of Samples |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| Total number of samples in total |  |  | 2500 |

|  |  |
| --- | --- |
| **Abstract**: The dataset is used for authorship identification in online Writeprint which is a new research field of pattern recognition. |  |

**Source:**

Dataset creator and donator: ZhiLiu, e-mail: liuzhi8673 '@' gmail.com, institution: National Engineering Research Center for E-Learning, Hubei Wuhan, China 

**Data Set Information:**

The dataset is the subset of RCV1. These corpus has already been used in author identification experiments. In the top 50 authors (with respect to total size of articles) were selected. 50 authors of texts labeled with at least one subtopic of the class CCAT(corporate/industrial) were selected.That way, it is attempted to minimize the topic factor in distinguishing among the texts. The training corpus consists of 2,500 texts (50 per author) and the test corpus includes other 2,500 texts (50 per author) non-overlapping with the training texts.

**Attribute Information:**

Attributes of the dataset are character n-grams(n=1-5)