**63.NSF Research Award Abstracts 1990-2003**

1. 数据库网址

https://archive.ics.uci.edu/ml/datasets/NSF+Research+Award+Abstracts+1990-2003

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 Michael J. Pazzani, from ICS Department, School of Computer Science, UCI.【2.用于[什么实验目的]】We used to classify the bag-of-word data from the files.【3】This data set consists of (a) 129,000 abstracts describing NSF awards for basic research, (b) bag-of-word data files extracted from the abstracts, (c) a list of words for indexing the bag-of-word data. The bag-of-word data was produced by automatically processing the abstracts with a text analyzer called NSFAbst, built using VisualText. While most fields of the output are very accurate, the authors were not extracted from the Investigator: field with 100% accuracy, due to wide variability in that field. The word list came from a separate process, and may not include all the words of interest in the abstracts.【4】The database has 129000 samples, respectively belong to traing with 120000 samples and testing with 9000 samples. As shown in Table 1.

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

|  |  |  |
| --- | --- | --- |
| Training | Testing | Total Number of Samples |
| 120000 | 9000 | 129000 |

3. 精简描述

The Data in our experiment were collected by Michael J. Pazzani, from ICS Department, School of Computer Science, UCI. The dataset includes have 129000 samples, which used to classify the bag-of-word data from the files. Through which, we divided the dataset into two part, training data set with 120000 samples and forecasting data set with 9000 samples.