<<返回新闻数据

搜狐新闻数据(SogouCS) 版本：2012

介绍：

来自搜狐新闻2012年6月—7月期间国内，国际，体育，社会，娱乐等18个频道的新闻数据，提供URL和正文信息

格式说明：

数据格式为

<doc>

<url>页面URL</url>

<docno>页面ID</docno>

<contenttitle>页面标题</contenttitle>

<content>页面内容</content>

</doc>

注意：content字段去除了HTML标签，保存的是新闻正文文本

相关任务：

文本分类

事件检测跟踪

新词发现

命名实体识别

自动摘要

相关资源：

全网新闻数据 互联网语料库 Reuters-21578 20 Newsgroups Web KB

成果列表：

Automatic Online News Issue Construction in Web Environment

Canhui Wang, Min Zhang, Shaoping ma, Liyun Ru, the 17th International World Wide Web Conference (WWW08), Beijing, April, 2008.

下载：

下载前请仔细阅读“搜狗实验室数据使用许可协议”

Please read the "License for Use of Sogou Lab Data" carefully before downloading.

迷你版(样例数据, 110KB)：tar.gz格式，zip格式

完整版(648MB)：tar.gz格式，zip格式

历史版本：2008版(6KB)：完整版(同时提供硬盘拷贝,65GB)：tar.gz格式

迷你版(样例数据, 1KB)：tar.gz格式

精简版(一个月数据, 347MB)：tar.gz格式

特别版(王灿辉WWW08论文数据, 647KB)：tar.gz格式

反馈：

在线上反馈留下您的宝贵意见和建议。

在资源下载FAQ中查找您遇到的资源下载问题的答案