多语种文本数据集

AISHELL-T0015

产品说明书

北京希尔贝壳科技有限公司
bd@aishelldata.com

目录

1 产品概述

AISHELL-T0015 多语种文本数据集，包含 34 个国家，共计 4726596 条文本。文本内容覆盖金融、体育、娱乐、科技、时事新闻、人名、街道名等日常生活领域。

2 产品目录结构

2.1 目录结构

数据目录结构	
数据目录结构	
| AISHELL-T0015.pdf	**(数据库简介)**
└─Languages	**(语种分类文件夹)**
| ├─German	**(语种分类)**
| | | | content.txt	**(文本文件)**
| ├─Chinese	**(语种分类)**

图表 2-1-1 数据目录结构

2.2 命名规则

CORPUS/USAGE/CONTRY/TEXT

e.g.AISHELL-T0015/Languages/Germany/content.txt

目录名称	内容	备注
CORPUS	AISHELL-T0015	语音数据库编号
USAGE	Languages	语种分类文件夹
CONTRY	German	语种分类
TEXT	content.txt	文本文件

图表 2-2-1 命名规则

3 文本信息

文本数据集包含 34 国语种，其中 29 国文本内容包含金融、体育、娱乐、科技、时事新闻以及日常生活领域，16 国文本包含人名、节目名、歌曲名以及街道名。文本内容经过专业的脱敏校对处理。

29 国具体条数参见下表：

语种	条数
丹麦	152303
乌克兰	113916
亚美尼亚	95297
俄国	169302
僧伽罗	80938
冰岛	121117
加泰罗尼亚	48504
南非	83081
印地	191774
希伯来	102027
德国	463545
意大利	277984
拉脱维亚	52058
挪威	239508
斯洛文尼亚	71176
斯瓦希里	15024
格鲁吉亚	130409
法国	339414
波兰	176833
瑞典	298322
祖鲁	161426

立陶宛	72577
罗马尼亚	63787
芬兰	321792
西班牙	280691
阿塞拜疆	49431
阿姆哈拉	50924
阿拉伯	153919
马拉亚拉姆	89116
合计	**4313892**

图表 3-3-1

16 国具体条数参见下表：

语种	人名	节目名	歌曲名	街道名	合计
丹麦	1739	2888	5767	5064	**15458**
俄国	1507	1497	6156	6090	**15250**
印地	2381	1643	13861	6086	**23971**
哥伦比亚	1783	872	2866	49	**5570**
埃及	1499	1500	5979	5942	**14920**
墨西哥	1506	1501	6061	6060	**15128**
德国	1500	1539	5989	6112	**15140**
意大利	4741	1861	6073	5541	**18216**
挪威	3200	1721	6116	6933	**17970**
智利	1552	928	3447	5936	**11863**
法国	1513	1502	6477	7380	**16872**
波兰	1503	1551	6260	6781	**16095**
祖鲁	1628	2744	6043	31589	**42004**
芬兰	1865	2411	5741	5901	**15918**
西班牙	1542	1509	5985	6022	**15098**

阿联酋	91	166	437	274	**968**
合计	**29550**	**25833**	**93258**	**111760**	**260401**

图表 3-3-2

4 版权声明