CCL 语料库语料分类分布情况汇总

2014-08-03

詹卫东

1　CCL 语料库：现代汉语、古代汉语字数统计

表 1：CCL 语料库字数、文件数统计

类别	字节数	汉字数	字符数
古代（gudai）文件夹	399,602,896 byte	163,662,943(token) 18,898(type)	195,526,348(token) 19,745(type)
现代（xiandai）文件夹	1,170,199,473 byte	509,913,589(token) 10,645(type)	592,412,339(token) 11,399(type)
合计	1,569,802,369 byte	141 个文件夹	3067 个文件

2　CCL 古代汉语语料库规模

表 2：古代汉语各朝代语料

朝代	字节数	百分比
01 周	292, 382	0.1783%
02 春秋	942, 293	0.5745%
03 战国	2, 521, 690	1.5375%
04 西汉	1, 013, 037	0.6176%
05 东汉	2, 585, 081	1.5761%
06 六朝	5, 796, 356	3.5340%
07 隋	1, 788	0.0011%
08 唐	9, 002, 907	5.4890%
09 五代	1, 555, 366	0.9483%
10 北宋	31, 982, 012	19.4992%
11 南宋	2, 843, 677	1.7338%
12 元	961, 884	0.5865%
13 明	21, 038, 301	12.8269%
14 清	48, 109, 077	29.3317%
15 民国	35, 371, 339	21.5656%
合计	164, 017, 190	100%

表 3：　古代汉语语料杂类

类别	字节数	百分比
二十五史	70, 496, 299	29. 9238%
全元曲	5, 751, 855	2. 4415%
全唐诗	8, 789, 487	3. 7309%
全宋词	3, 890, 311	1. 6513%
十三经注疏	16, 828, 127	7. 1431%
大藏經	72, 455, 219	30. 7554%
笔记	46, 666, 829	19. 8089%
蒙学读物	394, 627	0. 1675%
诸子百家	7, 001, 200	2. 9718%
辞书	1, 464, 667	0. 6217%
道藏	1, 847, 085	0. 7840%
合计	235, 585, 706	100%

3　CCL 语料库现代汉语语料

表 4：CCL 现代汉语语料库：现代/当代规模

分类	字节数
现代	15,250,163
当代	1,154,949,310
合计	1,170,199,473

表 5： CCL 现代汉语语料库：当代(1949 ——) 语料规模

类别	字节数	百分比
当代\口语	3,081,723	0.2668%
当代\史传	8,799,888	0.7619%
当代\应用文	48,286,885	4.1809%
当代\报刊	839,973,730	72.7282%
当代\文学	85,241,162	7.3805%
当代\电视电影	21,359,547	1.8494%
当代\相声小品	3,480,086	0.3013%
当代\网络语料	54,680,142	4.7344%
当代\翻译作品	90,046,147	7.7965%
当代	1,154,949,310	100%

表 6： CCL 现代汉语语料库：现代 (—— 1949) 语料规模

类别	字节数	百分比
现代\戏剧	1,197,572	7.8528%
现代\文学	14,052,591	92.1472%
现代	15,250,163	100.0000%

4　中文学术文献语料库（CWAC）规模

表 7：CWAC 的类别（大类）与规模

类别	字节数	百分比
Corpus\Arts	5,134,643	24.8582%
Corpus\Commerce	5,390,517	26.0970%
Corpus\Law	4,847,149	23.4664%
Corpus\Science	5,283,403	25.5784%
合计	20,655,712	100%

表 8： CWAC 的类别（细类）与规模

大类	小类	字节数	百分比
Corpus\Arts			
	Corpus\Arts\Education	764,104.00	3.6992%
	Corpus\Arts\History	632,219.00	3.0607%
	Corpus\Arts\Linguistics	755,423.00	3.6572%
	Corpus\Arts\Philosophy	746,945.00	3.6162%
	Corpus\Arts\Politics	776,343.00	3.7585%
	Corpus\Arts\Psychology	733,694.00	3.5520%
	Corpus\Arts\Sociology	725,915.00	3.5144%
	小计	5,134,643.00	24.8582%
Corpus\Commerce			
	Corpus\Commerce\Accounting	950,805.00	4.6031%
	Corpus\Commerce\Economics	887,334.00	4.2958%
	Corpus\Commerce\Finance	817,230.00	3.9564%
	Corpus\Commerce\Industrial_Relations	682,686.00	3.3051%
	Corpus\Commerce\Management	764,224.00	3.6998%
	Corpus\Commerce\Marketing	695,367.00	3.3665%
	Corpus\Commerce\Public_Policy	592,871.00	2.8703%
	小计	5,390,517.00	26.0970%
Corpus\Law			
	Corpus\Law\Constitutional_Law	709,767.00	3.4362%
	Corpus\Law\Criminal_Law	799,142.00	3.8689%
	Corpus\Law\Family_Law_and_Medico-Legal	544,653.00	2.6368%
	Corpus\Law\International_Law	615,807.00	2.9813%
	Corpus\Law\Pure_Commercial_Law	864,241.00	4.1840%
	Corpus\Law\Quasi-Commercial_Law	560,303.00	2.7126%
	Corpus\Law\Rights_and_Remedies	753,236.00	3.6466%
	小计	4,847,149.00	23.4664%
Corpus\Science			
	Corpus\Science\Biology	754,248.00	3.6515%
	Corpus\Science\Chemistry	739,871.00	3.5819%
	Corpus\Science\Computer_Science	882,276.00	4.2713%
	Corpus\Science\Geography	842,892.00	4.0807%
	Corpus\Science\Geology	711,296.00	3.4436%
	Corpus\Science\Mathematics	649,892.00	3.1463%
	Corpus\Science\Physics	702,928.00	3.4031%
	小计	5,283,403.00	25.5784%
合计		20,655,712.00	100%