Test data statistics

Text corpus	File	Size in MB	Alphabet size	Lines
canterbury[1]	alice29.txt	0.145	74	3,609
	asyoulik.txt	0.119	68	4,123
	cp.html	0.023	86	646
	fields.c	0.011	90	432
	grammar.lsp	0.004	76	95
	kennedy.xls	0.982	256	886
	lcet10.txt	0.407	84	7,520
	plrabn12.txt	0.460	81	10,700
	ptt5	0.489	159	1
	sum	0.036	255	95
	xargs.1	0.004	74	113
largecanterbury[2]	bible.txt	3.860	63	30,384
	E.coli	4.424	4	1
	world192.txt	2.359	94	65,120
silesia[3]	dickens	9.720	100	200,784
	mozilla	48.848	256	141,536
	mr	9.509	256	118,957
	nci	31.999	62	840,552
	ooffice	5.867	256	13,278
	osdb	9.618	256	29,564
	reymont	6.320	256	20,138
	samba	20.606	256	595,584
	sao	6.916	256	17,031
	webster	39.538	98	930,839
	xml	5.098	104	57,511
	x-ray	8.082	256	362,032
pizzachili[4]	sources	201.097	230	7,065,007
	pitches	53.246	133	84
	proteins	1,129.200	27	4,210,908
	dna	385.216	16	1,867
	english	2,108.000	239	43,894,766
	dblp.xml	282.417	97	7,619,950
repetitive[5]	Escherichia-Coli	107.468	15	1
	cere	439.917	5	1
	coreutils	195.772	236	6,443,719
	einstein.de.txt	88.461	117	545,357
	einstein.en.txt	445.963	139	2,379,632
	influenza	147.636	15	1
	kernel	246.011	160	9,550,967
	para	409.380	5	1
	world-leaders	44.792	89	103,041
genomes[6]	hg38	2,908.070	4	1
	mm10	2,529.890	4	1
	rn6	2,603.400	4	1

[1] http://www.data-compression.info/Corpora/CanterburyCorpus/index.html
[2] http://www.data-compression.info/Corpora/CanterburyCorpus/index.html
[3] http://sun.aei.polsl.pl/ sdeor/index.php?page=silesia
[4] http://pizzachili.dcc.uchile.cl/texts.html
[5] http://pizzachili.dcc.uchile.cl/repcorpus.html
[6] http://hgdownload.soe.ucsc.edu/downloads.html