

Test Set	Documents	Languages	Encoding	Document Length (bytes)	Avg. # characters
EuroGov	1500	10	1	17460.5 ± 39353.4	17037.3
TCL	3174	60	12	2623.2 ± 3751.9	1686.1
Wikipedia	4963	67	1	1480.8 ± 4063.9	1314.2