Statistical Methods in NLP 1
Assignment 1: Exploring Entropy and Language Modeling

Andrew McIsaac

December 23, 2021

1 Entropy of a Text

Code for this question can be found in the file `entropy.py`.

1.1 English

Table 1: Entropy of English

Text	Entropy			Avg. Perplexity
Original	5.2874			39.0553
	Min	Max	Average	
char 10%	4.7262	4.7367	4.7308	26.5533
char 5%	5.0522	5.0622	5.0564	33.2764
char 1%	5.2473	5.2529	5.2504	38.0654
char 0.1%	5.2825	5.2842	5.2835	38.9484
char 0.01%	5.2868	5.2873	5.2871	39.0454
char 0.001%	5.2874	5.2875	5.2874	39.0549
word 10%	5.4508	5.4602	5.4572	43.9313
word 5%	5.3769	5.3836	5.3800	41.6440
word 1%	5.3053	5.3087	5.3072	39.5928
word 0.1%	5.2890	5.2899	5.2894	39.1089
word 0.01%	5.2875	5.2878	5.2877	39.0609
word 0.001%	5.2874	5.2876	5.2875	39.0560

1.2 Czech

1.3 Independent Languages L_1 and L_2

2 Cross-Entropy and Language Modeling

Table 2: Entropy of Czech

Text	Entropy			Avg. Perplexity
Original	4.7478			26.8685
	Min	Max	Average	
char 10%	3.9976	4.0117	4.0048	16.0534
char 5%	4.3338	4.3443	4.3383	20.2283
char 1%	4.6547	4.6610	4.6578	25.2430
char 0.1%	4.7384	4.7399	4.7390	26.7039
char 0.01%	4.7467	4.7471	4.7469	26.8516
char 0.001%	4.7477	4.7478	4.7478	26.8671
word 10%	4.6335	4.6444	4.6378	24.8945
word 5%	4.6965	4.7022	4.6992	25.9768
word 1%	4.7383	4.7417	4.7394	26.7126
word 0.1%	4.7464	4.7474	4.7468	26.8508
word 0.01%	4.7475	4.7478	4.7477	26.8662
word 0.001%	4.7478	4.7479	4.7478	26.8686