TI – zad 3

Kamil Kowalczyk, 136742, L1, czwartek 13:30

Do rozwiązania wszystkich zadań, na wejściu podano korpus Wikipedii.

Zad 1

Entropia przybliżenia zerowego rzędu:

$$H(X) = -37 * \left(\frac{1}{37} * log_2\left(\frac{1}{37}\right)\right) \approx 5.21$$

Zad 2

Wynik działania programu został przedstawiony na poniższym zrzucie ekranu.

filename	S1	SC1	SC2	SC3	W1	WC1	WC2	WC3
norm_wiki_en.txt	4.29	3.52	3.02	2.48	11.54	6.39	2.18	0.48
	100.00	82.01	70.39	57.87	100.00	55.35	18.85	4.20
norm_wiki_eo.txt	4.18	3.34	2.87	2.39	11.56	6.56	2.48	0.63
	100.00	79.97	68.76	57.28	100.00	56.72	21.49	5.48
norm_wiki_et.txt	4.17	3.51	3.13	2.61	13.75	5.42	0.90	0.12
	100.00	84.10	75.17	62.61	100.00	39.46	6.58	0.85
norm_wiki_ht.txt	4.15	3.11	2.27	1.49	8.17	3.19	1.31	0.81
	100.00	75.10	54.83	35.99	100.00	39.10	16.06	9.95
norm_wiki_la.txt	4.23	3.45	2.82	2.15	11.97	4.40	1.17	0.39
	100.00	81.60	66.78	50.90	100.00	36.76	9.75	3.24
norm_wiki_nv.txt	3.87	2.95	2.37	1.80	9.15	3.86	1.72	0.90
	100.00	76.06	61.10	46.33	100.00	42.21	18.78	9.82
norm_wiki_so.txt	4.04	3.30	2.84	2.37	11.73	5.40	1.61	0.41
	100.00	81.67	70.40	58.77	100.00	46.02	13.71	3.49
sample1.txt	4.27	2.92	2.00	1.54	7.75	7.49	4.41	0.60
	100.00	68.24	46.81	36.02	100.00	96.61	56.87	7.68
sample2.txt	4.13	3.24	2.86	2.33	11.50	5.37	1.57	0.51
	100.00	78.49	69.33	56.38	100.00	46.71	13.69	4.41
sample3.txt	3.99	3.05	2.47	1.94	8.02	7.35	3.78	0.86
	100.00	76.39	61.79	48.58	100.00	91.58	47.13	10.71
sample4.txt	3.93	3.18	2.63	2.02	9.06	5.95	2.63	1.26
	100.00	81.02	66.86	51.50	100.00	65.67	29.03	13.95
sample5.txt	4.25	4.23	4.23	4.18	17.13	3.44	0.23	0.00
	100.00	99.42	99.37	98.23	100.00	20.11	1.37	0.02

Zrzut prezentuje tabelę. W pierwszej kolumnie znajduje się nazwa pliku wejściowego. Każdy plik wejściowy jest opisany dwoma wierszami. W pierwszym znajdują się miary entropii, oraz entropii warunkowych wyliczonych dla pliku. Kolumny S1 i W1 reprezentują entropię pierwszego rzędu odpowiednio dla znaków (signs) i dla słów (words). Następnie kolumny opisane jako SCk i WCk oznaczają entropie warunkowe k-tego rzędu odpowiednio dla znaków i słów. Pod każdym wierszem z wynikami dla konkretnego pliku wejściowego znajduje się wiersz pomocniczy, zawierający wartości procentowe. Ukazują one tempo zmieniania się entropii, wraz z rosnącym rzędem. Wartością stuprocentową są zawsze klasyczne wartości entropii. Występujące po nich wartości w kolumnach ukazują jaką część pierwotnej entropii stanowi wyliczona nowa entropia warunkowa k-tego rzędu. Ze względu na to, że dane zostały zgromadzone zarówno dla znaków jak i dla słów, w tabeli kolumny S1, SC1, SC2 i SC3, dotyczą znaków i ich wartości procentowe są wyliczane niezależnie w stosunku do kolumn W1, WC1, WC2, i WC3, które odpowiadają słowom.

Odpowiedzi na pytanie, czy plik zawiera język naturalny:

sample1.txt - FAŁSZ

sample2.txt - PRAWDA

sample3.txt - FAŁSZ

sample4.txt - PRAWDA

sample5.txt - FAŁSZ

Plik sample5.txt jest oczywistym przykładem pliku, który nie zawiera języka naturalnego. Widać to po tempie zmniejszania się wartości entropi warunkowych dla znaków przy rosnących rzędach, które w językach naturalnych jest szybsze. Analogiczna sytuacja występuje w plikach sample1.txt i sample3.txt, z tą różnicą, że tu tempo zmniejszania się entropi warunkowych słów (a nie znaków) jest zbyt niskie. Plik sample2.txt posiada zbliżone wyniki do plików z językami naturalnymi, dlatego bez wątpliwości można go zaklasyfikować do grupy poprawnych plików. Natomiast ciekawym przypadkiem jest sample4.txt. Dla znaków, jego entropie mają podobną tendencję do poprawnych plików. Jednak w przypadku słów, tempo zmniejszania się wartości entropi warunkowych jest zauważalnie wolniejsze. Mimo to, ze względu na poprawne wyniki entropi dotyczących znaków i brak większych anomalii został on uznany za plik zawierający język naturalny.