A4.2 - Entropy of English Letter

Jaan Tollander de Balsch - 452056

November 15, 2018

NOTE: The attached Python code contains the functions for computing the results that are obtained in this report.

In information theory entropy is defined as

$$S = -\sum_i P(x_i) \log_2 P(x_i), \tag{entropy}$$

where P is a probability mass function.

Then for English letter $l \in L$ associated with letter frequency, i.e. probability of occurrence $P(l)$, the entropy for English letter symbol space L can be calculated using the formula (entropy).

$$S = 4.175759791063625 \approx 4.18$$

Table 1: Constructing optimal multi-tap layout.

	$d=1$	$d=2$	...	$d=\lfloor n/m \rfloor$
k_1	l_1	l_{m+1}	...	
k_2	l_2	l_{m+2}	...	
$\vdots$	$\vdots$	$\vdots$	$\ddots$	
k_m	l_m	l_{2m}	...	

One design metric for creating an efficient Multi-tap text entry system is the expected number of key presses

$$\sum_{l \in L} P(l) \cdot d(l), \tag{expected-key-presses}$$

where $P(l)$ the frequency of the letter l and d measures how many key presses are required for inputting letter l. An optimal design aims to minimize the expected number of key presses.

Optimal layout for a set of m keys can be constructed by first sorting the letters by frequency in decreasing order $l_1, l_2, ..., l_n$ where $P(l_1) > P(l_2) > ... > P(l_n)$ and then by assigning first m letters to first places $d = 1$ on keys, second m letters to second places $d = 2$ on keys and so on until we run out of letters, see table 1 . This algorithm will minimize the expected number of key presses because higher frequencys $P(l)$ always have smaller the values d therefore minimizing (expected-key-presses).

The results from running the algorithm are:

1) Optimal layout with keys $(1, 2, ..., 9, 0)$ where $m = 10$ has expected number of key presses ≈ 1.28.

2) Optimal layout with keys $(2, 3, ..., 9)$ where $m = 8$ has an expected number of key presses ≈ 1.46.

3) The default layout with keys $(2, 3, ..., 9))$ where $m = 8$ has expected number of key presses ≈ 2.15.

By comparing the values in the results 2 and 3, by using the optimal layout the expected number of key presses can be lowered by $1 - 1.46/2.15 \approx 32\%$ which is almost a one third!