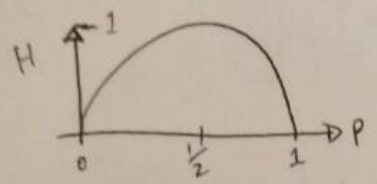
HW#2:

1) a) $p(W = w_i) = \frac{1}{N}$ $\quad H(w) = -\sum_{w \in V} p(w) \log p(w)$



$$H(w) = -\sum_{w \in V} \frac{1}{N} \log \frac{1}{N} = \sum_{w \in V} \frac{1}{N} \log N$$

$$= N\left(\frac{1}{N} \log N\right) = \log N$$

If the number of unique words $N = 1$ (i.e all words are identical) then $H(w) = \log 1 = 0$

$\therefore$ minimum $H(w) = 0$; maximum $H(w) = \log N$

b) Sample minimum $H(w)$ article (e.g. maximally homogeneous) $= \{w_1, w_1, w_1, w_1, w_1\}$

Sample maximum $H(W)$ article (e.g. maximally heterogeneous) $= \{w_2, w_1, w_3, w_4, w_5, w_6, w_3, w_5\}$

c) Two article which has $H(w) = 0$ probably mens that the documents themselves contain only one unique word each.

For example; $D_1 = \{w_1, w_1, w_1 \ldots, w_1\}$ and $D_2 = \{w_2, w_2 \ldots w_2\}$

in that case, when combining the documents, the most distinct set that you would get is $\{w_1, w_2\}$ so the maximum entropy for A_3 is $\log 2 = 0.69$