- 자연 후보 (스팸) $\Rightarrow \{X_1, X_2, X_3\}$

ex(1)

	$Y=0$	$Y=1$	
$X_1=0$	30	10	40
$X_1=1$	10	30	40
	40	40	80

$\Rightarrow$ X_1은 스팸메일인지 아닌지 구별하는데 도움되는 변수이며, 여기서 말하는 것이 "조건부 엔트로피"이다.

$$H[Y|X_1] = p(X_1=0)\cdot H[Y|X_1=0] + p(X_1=1)\cdot H[Y|X_1=1]$$

$$= \frac{40}{80} \times 0.81 + \frac{40}{80} \times 0.81 = \underline{0.81}$$

ex(2)

	$Y=0$	$Y=1$	
$X_2=0$	20	40	60
$X_2=1$	20	0	20
	40	40	80

$-\frac{20}{20}\log_2\frac{20}{20} - 0\log_2 0 = -\log_2 1 - 0 = 0$

$-\frac{20}{60}\log_2\frac{20}{60} - \frac{40}{60}\log_2\frac{40}{60} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.92$

$$H[Y|X_2] = p(X_2=0)\cdot H[Y|X_2=0] + p(X_2=1)\cdot H[Y|X_2=1]$$

$$= \frac{60}{80} \times 0.92 + \frac{20}{80} \times 0 = \boxed{0.69}$$

ex3

	$Y=0$	$Y=1$	
$X_3=0$	0	40	40
$X_3=1$	40	0	40
	40	40	80

$-\frac{0}{40}\log_2\frac{0}{40} - \frac{40}{40}\log_2\frac{40}{40} = 0$

$-\frac{40}{40}\log_2\frac{40}{40} - \frac{0}{40}\log_2\frac{0}{40} = 0$

$$H[Y|X_3] = p(X_3=0)\cdot H[Y|X_3=0] + p(X_3=1)\cdot H[Y|X_3=1]$$

$$= \frac{40}{80} + 0 + \frac{40}{80} + 0 = \boxed{0}$$

- 결론 : X_3 keyword가 엔트로피 0에 가장 가까이 스팸 메일 구별에 가장 좋은 keyword!