سوال ۱: $\arg\max_i I(Y, X_i) = \arg\max_i [H(Y) - H(Y|X_i)] = \arg\min_i H(Y|X_i)$

$H(Y) = -\sum_y P(Y=y)\log_2 P(Y=y)$, $H(Y|X_i) = \sum_x P(X_i=x) H(Y|X_i=x)$

First Level:

$H(Y|\text{نرخ شکایت از فرستنده}) = -\frac{1}{2}[\frac{3}{7}\log\frac{3}{7} + \frac{4}{7}\log\frac{4}{7}] - \frac{1}{2}[\frac{6}{7}\log\frac{6}{7} + \frac{1}{7}\log\frac{1}{7}]$

$= \frac{1}{2}[\frac{3}{7}\times 1,22 + \frac{4}{7}\times 0,8] + \frac{1}{2}[\frac{6}{7}\times 0,22 + \frac{1}{7}\times 2,8] = \frac{1}{2}\times 0,98 + \frac{1}{2}\times 0,58 = 0,78$

$H(Y|\text{طول بدنه ایمیل}) = -\frac{5}{14}[\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}] - \frac{5}{14}[\frac{2}{5}\log\frac{2}{5} + \frac{3}{5}\log\frac{3}{5}] - \frac{4}{14}[\frac{4}{4}\log\frac{4}{4} + \frac{0}{4}\log\frac{0}{4}]$

$= \frac{5}{14}[\frac{3}{5}\times 0,73 + \frac{2}{5}\times 1,32] + \frac{5}{14}[\frac{2}{5}\times 1,32 + \frac{3}{5}\times 0,73] + \frac{4}{14}[1\times 0 + 0] = 2\times\frac{5}{14}\times 0,94 = 0,68$

$H(Y|\text{فرمت ایمیل}) = -\frac{8}{14}[\frac{6}{8}\log\frac{6}{8} + \frac{2}{8}\log\frac{2}{8}] - \frac{6}{14}[\frac{3}{6}\log\frac{3}{6} + \frac{3}{6}\log\frac{3}{6}]$

$= \frac{8}{14}[\frac{3}{4}\times 0,41 + \frac{1}{4}\times 2] + \frac{6}{14}[\frac{1}{2}\times 1 + \frac{1}{2}\times 1] = \frac{6}{7}\times 0,8 + \frac{3}{7}\times 1 = 0,88$

$H(Y|\text{دامنه ایمیل}) = -\frac{6}{14}[\frac{4}{6}\log\frac{4}{6} + \frac{2}{6}\log\frac{2}{6}] - \frac{4}{14}[\frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4}] - \frac{4}{14}[\frac{2}{4}\log\frac{2}{4} + \frac{2}{4}\log\frac{2}{4}]$

$= \frac{6}{14}[\frac{2}{3}\times 0,58 + \frac{1}{3}\times 1,58] + \frac{4}{14}[\frac{3}{4}\times 0,41 + \frac{1}{4}\times 2] + \frac{4}{14}[\frac{1}{2}\times 1 + \frac{1}{2}\times 1]$

$= \frac{3}{7}\times 0,91 + \frac{2}{7}\times 0,8 + \frac{2}{7}\times 1 = 0,90$

کمترین مقدار متعلق به طول بدنه ایمیل می باشد پس آن انتخاب می شود.

Second Level: ایمیل ها با طول بدنه کوتاه

$H(Y|\text{نرخ شکایت از فرستنده}) = -\frac{2}{5}[\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}] - \frac{3}{5}[\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}]$

$= + \frac{2}{5}[\frac{1}{2}\times 1 + \frac{1}{2}\times 1] + \frac{3}{5}[\frac{2}{3}\times 0,58 + \frac{1}{3}\times 1,58] = \frac{2}{5} + \frac{3}{5}\times 0,91 = 0,946$

$H(Y|\text{فرمت ایمیل}) = -\frac{3}{5}[\frac{3}{3}\log\frac{3}{3} + \frac{0}{3}\log\frac{0}{3}] - \frac{2}{5}[\frac{2}{2}\log\frac{2}{2} + \frac{0}{2}\log\frac{0}{2}] = 0$

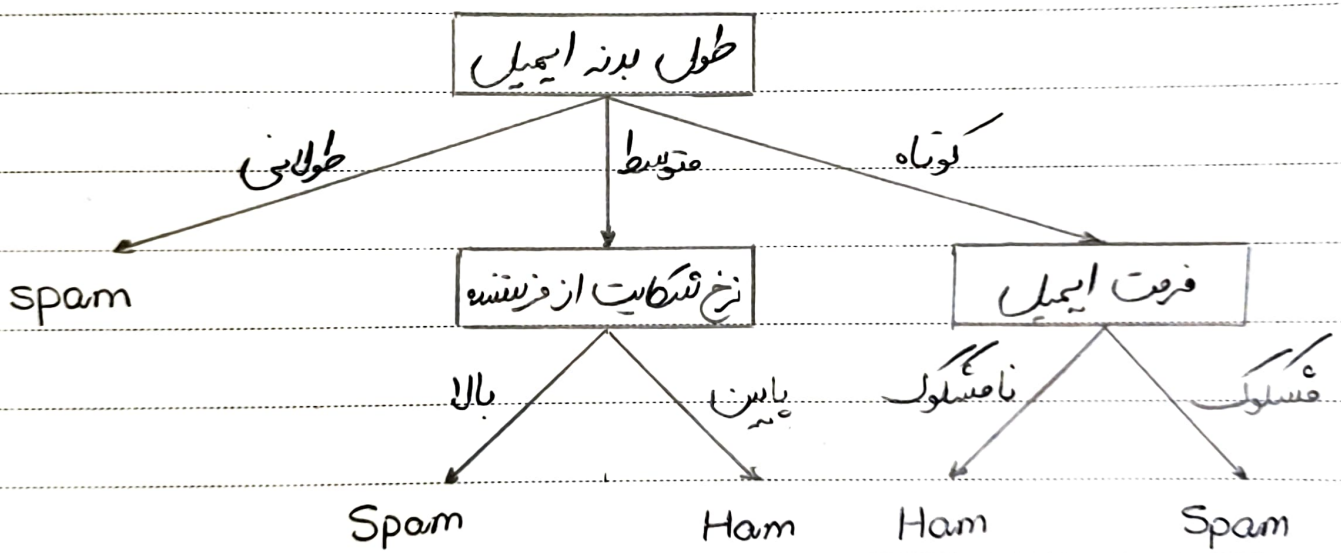
قطعاً از این کمتر می شود. پس فرمت ایمیل را انتخاب می کنیم

Second Level : متوسط

$H(Y|\text{نرخ شکایت از فرستنده}) = -\frac{3}{5}\left[\frac{0}{3}\log\frac{0}{3} + \frac{3}{3}\log\frac{3}{3}\right] - \frac{2}{5}\left[\frac{2}{2}\log\frac{2}{2} + \frac{0}{2}\log\frac{0}{2}\right] = 0$ ←

قطعاً از این کمتر نمی‌شود. پس نرخ شکایت از فرستنده را انتخاب می‌کنیم.

Second Level : طولانی ← چونکه همه spam هستند نیازی به شاخه جدید نیست



Spam_۶ Ham_۵ Spam_۴ Spam_۳ Ham_۲ Spam_۱