1.

(a)

$P(\text{"sensational"} \mid \text{"pop"}) = \frac{1}{1} = 1$,

$P(\text{"pop"} \mid \text{"the"}) = \frac{0}{2} = 0$

$P(\text{"sensation"} \mid \text{ricky}) = \frac{0}{1} = 0$

(b)

$P(\text{"pop martin"})$ should be higher.

MLE-estimated unigram model : $P(\text{"pop martian"}) = \frac{1}{10} \cdot \frac{1}{10} = \frac{1}{100} = P(\text{"pop martin"}) = \frac{1}{10} \cdot \frac{1}{10} = \frac{1}{100}$

" bigram " : $P(\text{"pop martian"}) = \frac{1}{10} \cdot \frac{0}{1} = 0 = P(\text{"pop martin"}) = \frac{1}{10} \cdot \frac{0}{1} = 0$

Alternative:

If we build the model with unigram and consider document seperatedly with equal occurrence, i.e. $P(d_1) = P(d_2) = \frac{1}{2}$

$P(w_1 \dots w_n) = P(w_1 \dots w_n \mid d_1) P(d_1) + P(w_1 \dots w_n \mid d_2) P(d_2)$.

$\Rightarrow P(\text{"pop martian"}) = 0 \cdot \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{6} \cdot 0 \cdot \frac{1}{2} = 0 < P(\text{"pop martin"}) = 0 \cdot 0 \cdot \frac{1}{2} + \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{72}$

$\Rightarrow$ Agree with the judgement.

2.

(a) $P(q \mid d_1) = \frac{2}{10} \cdot \frac{3}{10} = \frac{3}{50}$

$P(q \mid d_2) = \frac{7}{10} \cdot \frac{1}{10} = \frac{7}{100}$ $\Rightarrow P(q \mid d_2) > P(q \mid d_1)$

(b) $P(q \mid d_1) = \left(\frac{10}{10+10} \cdot \frac{2}{10} + \frac{10}{10+10} \cdot \frac{8000}{10000} \right) \left(\frac{10}{10+10} \cdot \frac{3}{10} + \frac{10}{10+10} \cdot \frac{1000}{10000} \right)$

$= \frac{1}{2} \cdot \frac{1}{5} = \frac{1}{10}$

$P(q \mid d_2) = \left(\frac{10}{10+10} \cdot \frac{7}{10} + \frac{10}{10+10} \cdot \frac{8000}{10000} \right) \left(\frac{10}{10+10} \cdot \frac{1}{10} + \frac{10}{10+10} \cdot \frac{1000}{10000} \right)$

$= \frac{3}{4} \cdot \frac{1}{10} = \frac{3}{40}$

$\Rightarrow P(q \mid d_1) > P(q \mid d_2)$

(c)

The rank in (b), $P(q \mid d_1) > P(q \mid d_2)$ is more reasonable because $P(w_1 \mid REF)$ is very high suggests that w_1 occurs in more documents than w_3 and is likely to be a stopword. Hence, its occurrence should be less important to the relevance of queries and documents. On the other hand, w_3 has a smaller $P(w_3 \mid REF)$ so its more important than w_1 considering relevance. Therefore, $rank(d_1)$ should be higher than $rank(d_2)$, and (b)'s ranking is more reasonable.

(Its like relevance is positive correlated to inverse document frequency in VSM)