* the answer to why from previous page:
we are looking for maximizing the $P(X|y)$ likelihood.

to max $P(X|y) = \frac{P(y|X)}{P(y)}$ we can maximiz $P(y|X)$

because $0<P(y)<1$. And it will vanish when taking FOC of lagrange.

Now the optimization problem, by taking log:

$$\log P(X,y) = \log \prod_{i=1}^{m} \left(P(y^i) \prod_{k=1}^{n} \theta_{c,k}^{x_k^i}\right) = \sum_{i=1}^{m} \log \left(P(y^i) \prod_{k=1}^{n} \theta_{c,k}^{x_k^i}\right) =$$

$$= \sum_{i=1}^{m} \left\{ \log (P(y^i)) + \sum_{k=1}^{n} \log \theta_{c,k}^{x_k^i} \right\}$$

with constaint $\sum_{k=1}^{n} \theta_{c,k} = 1 \quad \forall c$

$\theta_{c,k}^{x_k^i}$ is the likelihood of word x_k from sentence i that belongs to category c.

y^i is the catigory indicator. Here c is the category here (it can be c)

c is variable for category. here is 1 or 0

the lagrangian optimization problem will be:

$$L(P(X,y)) = \sum_{i=1}^{m} \left\{ \log (P(y^i)) + \sum_{k=1}^{n} \log \theta_{c,k}^{x_k^i} \right\} + \lambda \left(\sum_{k=1}^{n} \theta_{c,k} - 1 \right)$$