CS 510 - Advanced Information Retrieval

Wenqi He (wenqihe2)

1 Classic Probabilistic Retrieval Model

1.1

$$
\begin{aligned}
score(Q, D) &\overset{rank}{=} p(R=1 \mid Q, D) \overset{rank}{=} \frac{p(R=1 \mid Q, D)}{p(R=0 \mid Q, D)} \\
&\overset{rank}{=} \frac{p(Q, D \mid R=1)p(R=1)}{p(Q, D \mid R=0)p(R=0)} \\
&\overset{rank}{=} \frac{p(Q, D \mid R=1)}{p(Q, D \mid R=0)} \\
&\overset{rank}{=} \frac{p(Q \mid R=1)p(D \mid Q, R=1)}{p(Q \mid R=0)p(D \mid Q, R=0)} \\
&\overset{rank}{=} \frac{p(D \mid Q, R=1)}{p(D \mid Q, R=0)} \\
&\overset{rank}{=} \frac{p(D \mid Q, R=1)}{p(D \mid Q, R=0)} \\
&= \frac{\prod_{w \in V} p(w \mid Q, R=1)^{c(w,D)}}{\prod_{w \in V} p(w \mid Q, R=0)^{c(w,D)}} = \prod_{w \in V} \left(\frac{p(w \mid Q, R=1)}{p(w \mid Q, R=0)} \right)^{c(w,D)} \\
&\overset{rank}{=} \log \prod_{w \in V} \left(\frac{p(w \mid Q, R=1)}{p(w \mid Q, R=0)} \right)^{c(w,D)} \\
&= \sum_{w \in V} c(w, D) \log \frac{p(w \mid Q, R=1)}{p(w \mid Q, R=0)}
\end{aligned}
$$

There are $2|V|$ parameters, since we have two unigram models, one for $R = 1$ and the other for $R = 0$.

1.2

$$\hat{p}(w \mid Q, R=0) = \frac{c(w, C)}{\sum_{w' \in V} c(w', C)} = \frac{c(w, C)}{\sum_{d \in C} |d|}$$

1.3

$$\hat{p}(w \mid Q, R=1) = \frac{c(w, Q)}{|Q|}$$

1.4

$$
\begin{aligned}
\hat{p}(w \mid Q, R=1) &= (1-\lambda)\frac{c(w, Q)}{|Q|} + \lambda p(w \mid C) \\
&= (1-\lambda)\frac{c(w, Q)}{|Q|} + \lambda \frac{c(w, C)}{\sum_{d \in C} |d|}
\end{aligned}
$$

1.5

$$
\begin{aligned}
score(Q,D) &\overset{rank}{=} \sum_{w\in V} c(w,D)\log\frac{(1-\lambda)p(w\mid Q)+\lambda p(w\mid C)}{p(w\mid C)}\\
&= \sum_{w\in V} c(w,D)\log\left(\lambda+(1-\lambda)\frac{c(w,Q)}{|Q|c(w,C)}\sum_{d\in C}|d|\right)
\end{aligned}
$$

Here $c(w,Q)$ captures TF, $1/c(w,C)$ captures IDF, and $1/|Q|$ captures document length normalization.

2 Language Models

2.1

$$
\begin{aligned}
score(Q,D) &\overset{rank}{=} p(Q\mid D) = \prod_{w\in V} p(w\mid D)^{c(w,Q)} = \prod_{w\in Q} p(w\mid D)^{c(w,Q)}\\
&= \prod_{w\in Q, w\in D}\left((1-\lambda)\frac{c(w,D)}{|D|}+\lambda p(w\mid C)\right)^{c(w,Q)} \prod_{w\in Q, w\notin D}\left(\lambda p(w\mid C)\right)^{c(w,Q)}\\
&= \prod_{w\in Q, w\in D}\left((1-\lambda)\frac{c(w,D)}{|D|}+\lambda p(w\mid C)\right)^{c(w,Q)} \frac{\prod_{w\in Q}\left(\lambda p(w\mid C)\right)^{c(w,Q)}}{\prod_{w\in Q,w\in D}\left(\lambda p(w\mid C)\right)^{c(w,Q)}}\\
&= \prod_{w\in Q\cap D}\left(\frac{(1-\lambda)\frac{c(w,D)}{|D|}+\lambda p(w\mid C)}{\lambda p(w\mid C)}\right)^{c(w,Q)} \prod_{w\in Q}\left(\lambda p(w\mid C)\right)^{c(w,Q)}\\
&\overset{rank}{=} \prod_{w\in Q\cap D}\left(\frac{(1-\lambda)\frac{c(w,D)}{|D|}+\lambda p(w\mid C)}{\lambda p(w\mid C)}\right)^{c(w,Q)} = \prod_{w\in Q\cap D}\left(1+\frac{(1-\lambda)c(w,D)}{\lambda p(w\mid C)|D|}\right)^{c(w,Q)}\\
&\overset{rank}{=} \log\prod_{w\in Q\cap D}\left(1+\frac{(1-\lambda)c(w,D)}{\lambda p(w\mid C)|D|}\right)^{c(w,Q)}\\
&= \sum_{w\in Q\cap D} c(w,Q)\log\left(1+\frac{(1-\lambda)c(w,D)}{\lambda p(w\mid C)|D|}\right)
\end{aligned}
$$

2.2

The query vector is $q=(q_1,q_2,...q_{|V|}), q_i=c(w_i,Q)$. The document vector is

$$
d=(d_1,d_2,...d_{|V|}), d_i=\log\left(1+\frac{(1-\lambda)c(w_i,D)}{\lambda p(w_i\mid C)|D|}\right)
$$

The similarity function is dot product: $Sim(q,d)=q\cdot d$. For the document vector, $c(w_i,D)$ captures TF, $1/p(w_i\mid C)$ captures IDF, and $1/|D|$ captures document length normalization.

2.3

2.3.1 Jelinek-Mercer Smoothing

$$
\begin{aligned}
score(Q, D') &= \sum_{w \in Q \cap D'} c(w, Q) \log\left(1 + \frac{(1-\lambda)c(w, D')}{\lambda p(w \mid C)|D'|}\right) \\
&= \sum_{w \in Q \cap D} c(w, Q) \log\left(1 + \frac{(1-\lambda)kc(w, D)}{\lambda p(w \mid C)k|D|}\right) \\
&= \sum_{w \in Q \cap D} c(w, Q) \log\left(1 + \frac{(1-\lambda)c(w, D)}{\lambda p(w \mid C)|D|}\right) = score(Q, D)
\end{aligned}
$$

2.3.2 Dirichlet Prior Smoothing

$$
\begin{aligned}
score(Q, D) &\overset{rank}{=} \prod_{w \in Q \cap D} \left(1 + \frac{c(w, D)}{\mu p(w \mid C)}\right)^{c(w,Q)} \prod_{w \in Q} \left(\frac{\mu}{|D| + \mu} p(w \mid C)\right)^{c(w,Q)} \\
&\overset{rank}{=} \sum_{w \in Q \cap D} c(w, Q) \log\left(1 + \frac{c(w, D)}{\mu p(w \mid C)}\right) + \sum_{w \in Q} c(w, Q) \log\left(\frac{\mu}{|D| + \mu} p(w \mid C)\right) \\
&\overset{rank}{=} \sum_{w \in Q \cap D} c(w, Q) \log\left(1 + \frac{c(w, D)}{\mu p(w \mid C)}\right) + \sum_{w \in Q} c(w, Q) \log \frac{\mu}{|D| + \mu} \\
&= \sum_{w \in Q \cap D} c(w, Q) \log\left(1 + \frac{c(w, D)}{\mu p(w \mid C)}\right) + |Q| \log \frac{\mu}{|D| + \mu}
\end{aligned}
$$

$$
\begin{aligned}
score(Q, D') &= \sum_{w \in Q \cap D'} c(w, Q) \log\left(1 + \frac{c(w, D')}{\mu p(w \mid C)}\right) + |Q| \log \frac{\mu}{|D'| + \mu} \\
&= \sum_{w \in Q \cap D} c(w, Q) \log\left(1 + \frac{kc(w, D)}{\mu p(w \mid C)}\right) + |Q| \log \frac{\mu}{k|D| + \mu}
\end{aligned}
$$

$$
\begin{aligned}
score(Q, D') - score(Q, D) &= \sum_{w \in Q \cap D} c(w, Q) \log \frac{\mu p(w \mid C) + kc(w, D)}{\mu p(w \mid C) + c(w, D)} - |Q| \log \frac{k|D| + \mu}{|D| + \mu} \\
&= \sum_{w \in Q} c(w, Q) \left(\log \frac{\mu p(w \mid C) + kc(w, D)}{\mu p(w \mid C) + c(w, D)} - \log \frac{k|D| + \mu}{|D| + \mu}\right)
\end{aligned}
$$

for each word w,

$$
\frac{\mu p(w \mid C) + kc(w, D)}{\mu p(w \mid C) + c(w, D)} > \frac{k|D| + \mu}{|D| + \mu} \quad \Longleftrightarrow \quad \tilde{p}(w \mid D) = \frac{c(w, D)}{|D|} > p(w \mid C)
$$

Therefore, if every query term has a higher empirical probability in D than expected based on the whole collection, then the score for D' would be higher. Similarly, if every term has a lower empirical probability than expected, then D' would be lower. Otherwise, whether the score increases or decreases depends on how many times each term occurs in the query and how each term's empirical probability compares to its probability predicted by the background model.

3 KL-divergence Retrieval Function

$$
\begin{aligned}
score(Q,D) = -D(\theta_Q||\theta_D) &= \sum_{w\in V} p(w \mid \theta_Q) \log \frac{p(w \mid \theta_D)}{p(w \mid \theta_Q)} \\
&= \sum_{w\in V} p(w \mid \theta_Q) \log p(w \mid \theta_D) - \sum_{w\in V} p(w \mid \theta_Q) \log p(w \mid \theta_Q) \\
&\overset{rank}{=} \sum_{w\in V} p(w \mid \theta_Q) \log p(w \mid \theta_D) \\
&= \sum_{w\in V} \frac{c(w,Q)}{|Q|} \log p(w \mid \theta_D) \\
&\overset{rank}{=} \sum_{w\in V} c(w,Q) \log p(w \mid \theta_D)
\end{aligned}
$$

which is the query likelihood retrieval function.

4 Divergence Minimization Feedback Model

The objective function is

$$
\begin{aligned}
f(\theta) &= \frac{1}{n}\sum_{i=1}^{n} D(\theta \parallel \theta_{E_i}) - \lambda D(\theta \parallel \theta_C) \\
&= \frac{1}{n}\sum_{i=1}^{n} \Big[H(\theta,\theta_{E_i}) - H(\theta)\Big] - \lambda\Big[H(\theta,\theta_C) - H(\theta)\Big] \\
&= \frac{1}{n}\sum_{i=1}^{n} \Big[H(\theta,\theta_{E_i}) - \lambda H(\theta,\theta_C)\big] - (1-\lambda)H(\theta) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left[-\sum_{w\in V} p(w \mid \theta) \log p(w \mid \theta_{E_i}) + \lambda \sum_{w\in V} p(w \mid \theta) \log p(w \mid \theta_C)\right] + (1-\lambda) \sum_{w\in V} p(w \mid \theta) \log p(w \mid \theta) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\sum_{w\in V} p(w \mid \theta)\Big[\log p(w \mid \theta_{E_i}) - \lambda \log p(w \mid \theta_C)\Big] + (1-\lambda)\sum_{w\in V} p(w \mid \theta) \log p(w \mid \theta)
\end{aligned}
$$

The objective function with constraint is:

$$
\begin{aligned}
g(\theta,\mu) &= f(\theta) + \mu\Big[\sum_{w\in V} p(w \mid \theta) - 1\Big] \\
&= -\frac{1}{n}\sum_{i=1}^{n}\sum_{w\in V} p(w \mid \theta)\Big[\log p(w \mid \theta_{E_i}) - \lambda \log p(w \mid \theta_C)\Big] + (1-\lambda)\sum_{w\in V} p(w \mid \theta) \log p(w \mid \theta) + \mu\Big[\sum_{w\in V} p(w \mid \theta) - 1\Big]
\end{aligned}
$$

Taking the derivatives and setting them to 0:

$$
\begin{aligned}
\frac{\partial}{\partial p(w \mid \theta)} g(\theta, \mu) &= -\frac{1}{n} \sum_{i=1}^{n} \Big[\log p(w \mid \theta_{E_i}) - \lambda \log p(w \mid \theta_C) \Big] + (1-\lambda) \Big[\log p(w \mid \theta) + 1 \Big] + \mu = 0 \\
\log p(w \mid \theta) &= -1 - \frac{\mu}{1-\lambda} + \frac{1}{(1-\lambda)n} \sum_{i=1}^{n} \Big[\log p(w \mid \theta_{E_i}) - \lambda \log p(w \mid \theta_C) \Big] \\
p(w \mid \theta) &= \exp\left(-1 - \frac{\mu}{1-\lambda} \right) \exp\left(\frac{1}{(1-\lambda)n} \sum_{i=1}^{n} \Big[\log p(w \mid \theta_{E_i}) - \lambda \log p(w \mid \theta_C) \Big] \right) \\
&= \exp\left(-1 - \frac{\mu}{1-\lambda} \right) \exp\left(\frac{1}{(1-\lambda)n} \sum_{i=1}^{n} \log p(w \mid \theta_{E_i}) - \frac{\lambda}{1-\lambda} \log p(w \mid \theta_C) \right) \\
&= C \exp\left(\frac{1}{(1-\lambda)n} \sum_{i=1}^{n} \log p(w \mid \theta_{E_i}) - \frac{\lambda}{1-\lambda} \log p(w \mid \theta_C) \right) \\
\frac{\partial}{\partial \mu} g(\theta, \mu) &= \sum_{w \in V} p(w \mid \theta) - 1 = \sum_{w \in V} C \exp\left(\frac{1}{(1-\lambda)n} \sum_{i=1}^{n} \log p(w \mid \theta_{E_i}) - \frac{\lambda}{1-\lambda} \log p(w \mid \theta_C) \right) - 1 = 0 \\
C &= \left[\sum_{w \in V} \exp\left(\frac{1}{(1-\lambda)n} \sum_{i=1}^{n} \log p(w \mid \theta_{E_i}) - \frac{\lambda}{1-\lambda} \log p(w \mid \theta_C) \right) \right]^{-1}
\end{aligned}
$$

Finally, the solution is

$$
p(w \mid \theta^*) = \frac{\exp\left(\frac{1}{(1-\lambda)n} \sum_{i=1}^{n} \log p(w \mid \theta_{E_i}) - \frac{\lambda}{1-\lambda} \log p(w \mid \theta_C) \right)}{\sum_{w' \in V} \exp\left(\frac{1}{(1-\lambda)n} \sum_{i=1}^{n} \log p(w' \mid \theta_{E_i}) - \frac{\lambda}{1-\lambda} \log p(w' \mid \theta_C) \right)}
$$

5 Deriving the EM Algorithm for PLSA

5.1

5.1.1

$$
\log p(D \mid \Theta) = \sum_{i=1}^{|D|} \log p(w_i \mid \Theta) = \sum_{i=1}^{|D|} \log \big[\lambda p(w_i \mid H) + (1-\lambda) p(w_i \mid T) \big]
$$

5.1.2

The complete log likelihood function is:

$$
\log p(D, Z \mid \Theta) = \sum_{i=1}^{|D|} \log p(w_i, z_i \mid \Theta) = \sum_{i=1}^{|D|} \log \big[p(z_i) p(w_i \mid z_i) \big] = \sum_{i=1}^{|D|} \big[\log p(z_i) + \log p(w_i \mid z_i) \big]
$$

The Q-function is:

$$
\begin{aligned}
Q(D,\Theta) &= E\left[\sum_{i=1}^{|D|}\big[\log p(Z_i)+\log p(w_i \mid Z_i)\big]\middle| D,\Theta^{(t)}\right] = \sum_{i=1}^{|D|} E\Big[\log p(Z_i)+\log p(w_i \mid Z_i)\Big| D,\Theta^{(t)}\Big] \\
&= \sum_{i=1}^{|D|}\sum_z p(Z_i = z \mid D,\Theta^{(t)})\big[\log p(z)+\log p(w_i \mid z)\big] = \sum_{i=1}^{|D|}\sum_z p(Z_{w_i} = z \mid \Theta^{(t)})\big[\log p(z)+\log p(w_i \mid z)\big] \\
&= \sum_{w\in V} c(w,D)\sum_z p(Z_w = z \mid \Theta^{(t)})\big[\log p(z)+\log p(w \mid z)\big] \\
&= \sum_{w\in V} c(w,D)\Big[p(Z_w = H \mid \Theta^{(t)})\big[\log\lambda+\log p(w \mid H)\big] + p(Z_w = T \mid \Theta^{(t)})\big[\log(1-\lambda)+\log p(w \mid T)\big]\Big]
\end{aligned}
$$

Taking the partial derivative w.r.t. λ and setting it to zero:

$$
\begin{aligned}
\frac{\partial}{\partial\lambda}Q(D,\Theta) &= \sum_{w\in V} c(w,D)\left[\frac{p(Z_w = H \mid \Theta^{(t)})}{\lambda} - \frac{p(Z_w = T \mid \Theta^{(t)})}{1-\lambda}\right] \\
&= \frac{1}{\lambda}\sum_{w\in V} c(w,D)p(Z_w = H \mid \Theta^{(t)}) - \frac{1}{1-\lambda}\sum_{w\in V} c(w,D)p(Z_w = T \mid \Theta^{(t)}) = 0 \\
\lambda^{(t+1)} &= \boxed{\frac{\sum_{w\in V} c(w,D)p(Z_w = H \mid \Theta^{(t)})}{\sum_{z'}\sum_{w\in V} c(w,D)p(Z_w = z' \mid \Theta^{(t)})} = \frac{1}{|D|}\sum_{w\in V} c(w,D)p(Z_w = H \mid \Theta^{(t)})}
\end{aligned}
$$

Using Bayes' Rule,

$$
\begin{aligned}
p(Z_w = z \mid \Theta^{(t)}) &= p(z \mid w,\Theta^{(t)}) \\
&\propto p(z \mid \Theta^{(t)})p(w \mid z,\Theta^{(t)}) = p^{(t)}(z)p^{(t)}(w \mid z) \\
p(Z_w = H \mid \Theta^{(t)}) &= \boxed{\frac{\lambda^{(t)}p(w \mid H)}{\lambda^{(t)}p(w \mid H)+(1-\lambda^{(t)})p(w \mid T)}} \\
p(Z_w = T \mid \Theta^{(t)}) &= \boxed{\frac{(1-\lambda^{(t)})p(w \mid T)}{\lambda^{(t)}p(w \mid H)+(1-\lambda^{(t)})p(w \mid T)}}
\end{aligned}
$$

5.2

We have the same Q-function as before,

$$
Q(D,\Theta) = \sum_{w\in V} c(w,D)\Big[p(Z_w = H \mid \Theta^{(t)})\big[\log\lambda+\log p(w \mid H)\big] + p(Z_w = T \mid \Theta^{(t)})\big[\log(1-\lambda)+\log p(w \mid T)\big]\Big]
$$

Applying the constraint, the Lagrangian is:

$$
L(D,\Theta) = Q(D,\Theta) - \mu\left(\sum_{w\in V} p(w \mid H) - 1\right)
$$

Taking derivatives w.r.t. $p(w \mid H)$, μ:

$$
\begin{aligned}
\frac{\partial}{\partial p(w \mid H)} L(D, \Theta) &= c(w, D) \frac{p(Z_w = H \mid \Theta^{(t)})}{p(w \mid H)} - \mu = 0 \\
\frac{\partial}{\partial \mu} L(D, \Theta) &= -\left(\sum_{w \in V} p(w \mid H) - 1 \right) = 0 \\
p^{(t+1)}(w \mid H) &= \boxed{\frac{c(w, D) p(Z_w = H \mid \Theta^{(t)})}{\sum_{w' \in V} c(w', D) p(Z_{w'} = H \mid \Theta^{(t)})}}
\end{aligned}
$$

Using Bayes' Rule,

$$
\begin{aligned}
p(Z_w = z \mid \Theta^{(t)}) &= p(z \mid w, \Theta^{(t)}) \\
&\propto p(z \mid \Theta^{(t)}) p(w \mid z, \Theta^{(t)}) = p^{(t)}(z) p^{(t)}(w \mid z) \\
p(Z_w = H \mid \Theta^{(t)}) &= \boxed{\frac{0.9 p^{(t)}(w \mid H)}{0.9 p^{(t)}(w \mid H) + 0.1 p(w \mid T)}}
\end{aligned}
$$

5.3

5.3.1 Mixture Model

The complete log likelihood function is:

$$
\begin{aligned}
\log p(X, Z \mid \Theta) &= \sum_{d \in S_1} \log p(d, z_d \mid \Theta) + \sum_{d \in S_2} \log p(d, z_d \mid \Theta) \\
&= \sum_{d \in S_1} \sum_{i=1}^{|d|} \log p(w_{d,i}, z_{d,i} \mid \Theta) + \sum_{d \in S_2} \sum_{i=1}^{|d|} \log p(w_{d,i}, z_{d,i} \mid \Theta) \\
&= \sum_{d \in S_1} \sum_{i=1}^{|d|} \Big[\log p(z_{d,i} \mid d) + \log p(w_{d,i} \mid z_{d,i}) \Big] + \sum_{d \in S_2} \sum_{i=1}^{|d|} \Big[\log p(z_d \mid d) + \log p(w_{d,i} \mid z_d) \Big]
\end{aligned}
$$

If we define $p(Z_{d,w} = z \mid \Theta^{(t)})$ as a shorthand for $p(Z = z \mid w, d, \Theta^{(t)})$, and use the fact that

$$
p(Z_{d,i} = z \mid W_{d,i} = w) = \frac{p(Z_{d,i} = z) p(W_{d,i} = w \mid Z_{d,i} = z)}{p(W_{d,i} = w)} = \frac{p(z \mid d) p(w \mid z, d)}{p(w \mid d)} = p(z \mid w, d) = p(Z_{d,w} = z)
$$

The Q-function is:

$$Q(X,\Theta) = E\left[\sum_{d\in S_1}\sum_{i=1}^{|d|}\Big[\log p(Z_{d,i} \mid d) + \log p(w_{d,i} \mid Z_{d,i})\Big] + \sum_{d\in S_2}\sum_{i=1}^{|d|}\Big[\log p(z_d \mid d) + \log p(w_{d,i} \mid z_d)\Big]\,\middle|\, X, \Theta^{(t)}\right]$$

$$= \sum_{d\in S_1}\sum_{i=1}^{|d|} E\Big[\log p(Z_{d,i} \mid d) + \log p(w_{d,i} \mid Z_{d,i})\Big| X, \Theta^{(t)}\Big] + \sum_{d\in S_2}\sum_{i=1}^{|d|}\Big[\log p(z_d \mid d) + \log p(w_{d,i} \mid z_d)\Big]$$

$$= \sum_{d\in S_1}\sum_{i=1}^{|d|}\sum_z p(Z_{d,i} = z \mid X, \Theta^{(t)})\Big[\log p(z \mid d) + \log p(w_{d,i} \mid z)\Big] + \sum_{d\in S_2}\sum_{i=1}^{|d|}\Big[\log p(z_d \mid d) + \log p(w_{d,i} \mid z_d)\Big]$$

$$= \sum_{d\in S_1}\sum_{i=1}^{|d|}\sum_z p(Z_{d,i} = z \mid W_{d,i} = w_{d,i}, \Theta^{(t)})\Big[\log p(z \mid d) + \log p(w_{d,i} \mid z)\Big] + \sum_{d\in S_2}\sum_{i=1}^{|d|}\Big[\log p(z_d \mid d) + \log p(w_{d,i} \mid z_d)\Big]$$

$$= \sum_{d\in S_1}\sum_{i=1}^{|d|}\sum_z p(Z_{d,w_{d,i}} = z \mid \Theta^{(t)})\Big[\log p(z \mid d) + \log p(w_{d,i} \mid z)\Big] + \sum_{d\in S_2}\sum_{i=1}^{|d|}\Big[\log p(z_d \mid d) + \log p(w_{d,i} \mid z_d)\Big]$$

$$= \sum_{d\in S_1}\sum_{w\in V} c(w,d)\sum_z p(Z_{d,w} = z \mid \Theta^{(t)})\Big[\log p(z \mid d) + \log p(w \mid z)\Big] + \sum_{d\in S_2}\sum_{w\in V} c(w,d)\Big[\log p(z_d \mid d) + \log p(w \mid z_d)\Big]$$

$$= \sum_{d\in S_1}\sum_{w\in V} c(w,d)\sum_z p(Z_{d,w} = z \mid \Theta^{(t)})\Big[\log p(z \mid d) + \log p(w \mid z)\Big] + \sum_{d\in S_2}\sum_{w\in V} c(w,d)\sum_z \delta[z = z_d]\Big[\log p(z \mid d) + \log p(w \mid z$$

The Lagrangian is,

$$L(C,\Theta) = Q(C,\Theta) - \sum_{d\in S_1}\lambda_d\left(\sum_z p(z \mid d) - 1\right) - \sum_z \mu_z\left(\sum_{w\in V} p(w \mid z) - 1\right)$$

Taking derivatives w.r.t. $p(z \mid d)$ and λ_d for $d \in S_1$:

$$\frac{\partial}{\partial p(z \mid d)} L(C,\Theta) = \sum_{w\in V}\frac{c(w,d)p(Z_{d,w} = z \mid \Theta^{(t)})}{p(z \mid d)} - \lambda_d = 0$$

$$\frac{\partial}{\partial \lambda_d} L(C,\Theta) = -\left(\sum_z p(z \mid d) - 1\right) = 0$$

$$p^{(t+1)}(z \mid d) = \boxed{\frac{\sum_{w\in V} c(w,d)p(Z_{d,w} = z \mid \Theta^{(t)})}{\sum_{z'}\sum_{w\in V} c(w,d)p(Z_{d,w} = z' \mid \Theta^{(t)})} = \frac{1}{|d|}\sum_{w\in V} c(w,d)p(Z_{d,w} = z \mid \Theta^{(t)})}$$

Taking derivatives w.r.t. $p(w \mid z)$ and μ_z:

$$\frac{\partial}{\partial p(w \mid z)} L(C,\Theta) = \sum_{d\in S_1}\frac{c(w,d)p(Z_{d,w} = z \mid \Theta^{(t)})}{p(w \mid z)} + \sum_{d\in S_2}\frac{c(w,d)\delta[z = z_d]}{p(w,z)} - \mu_z = 0$$

$$\frac{\partial}{\partial \mu_z} L(C,\Theta) = -\left(\sum_{w\in V} p(w \mid z) - 1\right) = 0$$

$$p^{(t+1)}(w \mid z) = \boxed{\frac{\sum_{d\in S_1} c(w,d)p(Z_{d,w} = z \mid \Theta^{(t)}) + \sum_{d\in S_2} c(w,d)\delta[z = z_d]}{\sum_{w'\in V}\Big[\sum_{d\in S_1} c(w',d)p(Z_{d,w'} = z \mid \Theta^{(t)}) + \sum_{d\in S_2} c(w',d)\delta[z = z_d]\Big]}}$$

Using Bayes' Rule,

$$
\begin{aligned}
p(Z_{d,w} = z \mid \Theta^{(t)}) &= p(z \mid w, d, \Theta^{(t)}) \\
&\propto p(z \mid d, \Theta^{(t)}) p(w \mid z, d, \Theta^{(t)}) = p^{(t)}(z \mid d) p^{(t)}(w \mid z) \\
p(Z_{d,w} = z \mid \Theta^{(t)}) &= \boxed{\frac{p^{(t)}(z \mid d) p^{(t)}(w \mid z)}{\sum_{z'} p^{(t)}(z' \mid d) p^{(t)}(w \mid z')}}, z \in \{Seattle, Chicago\}
\end{aligned}
$$

5.3.2 Clustering Model

The complete log likelihood is:

$$
\log p(X, Z \mid \Theta) = \sum_{d \in S_1} \left[\log p(z_d \mid d) + \sum_{w \in V} c(w,d) \log p(w \mid z_d) \right] + \sum_{d \in S_2} \left[\log p(z_d \mid d) + \sum_{w \in V} c(w,d) \log p(w \mid z_d) \right]
$$

The Q-function is:

$$
\begin{aligned}
Q(X, \Theta) &= E \left[\sum_{d \in S_1} \left[\log p(Z_d \mid d) + \sum_{w \in V} c(w,d) \log p(w \mid Z_d) \right] + \sum_{d \in S_2} \left[\log p(z_d \mid d) + \sum_{w \in V} c(w,d) \log p(w \mid z_d) \right] \middle| X, \Theta^{(t)} \right] \\
&= \sum_{d \in S_1} E \left[\log p(Z_d \mid d) + \sum_{w \in V} c(w,d) \log p(w \mid Z_d) \middle| X, \Theta^{(t)} \right] + \sum_{d \in S_2} \left[\log p(z_d \mid d) + \sum_{w \in V} c(w,d) \log p(w \mid z_d) \right] \\
&= \sum_{d \in S_1} \sum_z p(Z_d = z \mid d, \Theta^{(t)}) \left[\log p(z \mid d) + \sum_{w \in V} c(w,d) \log p(w \mid z) \right] + \sum_{d \in S_2} \sum_z \delta[z = z_d] \Bigg[\log p(z \mid d) + \sum_{w \in V} c(w,d) \log p(w \mid z
\end{aligned}
$$

The Lagrangian is,

$$
L(C, \Theta) = Q(C, \Theta) - \sum_{d \in S_1} \lambda_d \left(\sum_z p(z \mid d) - 1 \right) - \sum_z \mu_z \left(\sum_{w \in V} p(w \mid z) - 1 \right)
$$

Using Bayes' Rule,

$$
\begin{aligned}
p(Z_d = z \mid d, \Theta^{(t)}) &\propto p(Z_d = z \mid \Theta^{(t)}) p(d \mid Z_d = z, \Theta^{(t)}) = p^{(t)}(z \mid d) \prod_{w \in V} p^{(t)}(w \mid z)^{c(w,d)} \\
p(Z_d = z \mid d, \Theta^{(t)}) &= \boxed{\frac{p^{(t)}(z \mid d) \prod_{w \in V} p^{(t)}(w \mid z)^{c(w,d)}}{\sum_{z'} p^{(t)}(z' \mid d) \prod_{w \in V} p^{(t)}(w \mid z')^{c(w,d)}}}
\end{aligned}
$$

Taking derivatives w.r.t. $p(z \mid d)$ and λ_d for $d \in S_1$:

$$
\begin{aligned}
\frac{\partial}{\partial p(z \mid d)} L(C, \Theta) &= \frac{p(Z_d = z \mid d, \Theta^{(t)})}{p(z \mid d)} - \lambda_d = 0 \\
\frac{\partial}{\partial \lambda_d} L(C, \Theta) &= -\left(\sum_z p(z \mid d) - 1 \right) = 0 \\
p^{(t+1)}(z \mid d) &= \frac{p(Z_d = z \mid d, \Theta^{(t)})}{\sum_{z'} p(Z_d = z' \mid d, \Theta^{(t)})} = p(Z_d = z \mid d, \Theta^{(t)}) \\
&= \boxed{\frac{p^{(t)}(z \mid d) \prod_{w \in V} p^{(t)}(w \mid z)^{c(w,d)}}{\sum_{z'} p^{(t)}(z' \mid d) \prod_{w \in V} p^{(t)}(w \mid z')^{c(w,d)}}}
\end{aligned}
$$

Taking derivatives w.r.t. $p(w \mid z)$ and μ_z:

$$\frac{\partial}{\partial p(w \mid z)} L(C, \Theta) = \sum_{d \in S_1} p(Z_d = z \mid d, \Theta^{(t)}) \frac{c(w,d)}{p(w \mid z)} + \sum_{d \in S_2} \delta[z = z_d] \frac{c(w,d)}{p(w \mid z)} - \mu_d = 0$$

$$\frac{\partial}{\partial \mu_z} L(C, \Theta) = -\left(\sum_{w \in V} p(w \mid z) - 1 \right) = 0$$

$$p^{(t+1)}(w \mid z) = \boxed{\frac{\sum_{d \in S_1} c(w,d) p(Z_d = z \mid d, \Theta^{(t)}) + \sum_{d \in S_2} c(w,d) \delta[z = z_d]}{\sum_{w' \in V} \left[\sum_{d \in S_1} c(w',d) p(Z_d = z \mid d, \Theta^{(t)}) + \sum_{d \in S_2} c(w',d) \delta[z = z_d] \right]}}$$