CS598: CXZ	**Fall 2016**
Problem Set 4	
chen zhang, NetId: czhang49	*Handed In: October 9, 2016*

1 KL-Divergence Retrieval Function

In KL-Divergence, $score(D, Q) = -D(q||d)$, and we have the following equations:

$$\begin{aligned} -D(q||d) &= \sum_{w \in V} -p(w|\theta_q) \log (\frac{p(w|\theta_q)}{p(w|\theta_d)}) \\ &\propto \sum_{w \in V} p(w|\theta_q) \log p(w|\theta_d) \end{aligned}$$

As a special case, if $p(w|\theta_q) = \frac{c(w,Q)}{|Q|}$, then we have :

$$\begin{aligned} -D(q||d) &\propto \sum_{w \in V} \frac{c(w,Q)}{|Q|} \log p(w|\theta_d) \\ &\propto \sum_{w \in V} c(w,Q) \log p(w|\theta_d) \\ &\propto \log \prod_{w \in V} p(w|\theta_d)^{c(w,Q)} \end{aligned}$$

which is exactly the query likelihood retrieval function in the logarithmic form.

2 Language Models for Boolean Queries

(a)

In the basic multinomial query likelihood retrieval model, the semantic structure is Conjunctive. Since in the likelihood equation :

$$p(q = q_1, q_2 \dots q_m) = \prod_{j=1}^{m} p(q_j|d),$$

we use multiplication for the probability of the query words, so if one query word is missing then the whole probability is zero, which is the property of conjunctive relation.

(b)

Now that with the basic multinomial query likelihood model, we have a conjunctive relation, we can modify that to include the disjunctive part. For $Q = Q_1 \quad AND \quad Q_2 \ldots AND \quad Q_k$

$$p(Q) = \prod_{i=1}^{m} p(Q_i),$$

where in each of the Q_i where $Q_i = w_{i,1} \ldots OR \quad w_{i,n_i}$, we can have the following relation:

$$p(Q_i) = \sum_{j=1}^{|Q_i|=n_i} p(w_{i,j}|d).$$

So the final form of the relation is :

$$p(Q) = \prod_{i=1}^{m} [\sum_{j=1}^{|Q_i|=n_i} p(w_{i,j}|d)]$$

3 Mixture Models

(a)

The general formula for a word to appear is as follows:

$$p(w) = \lambda p(w|H) + (1-\lambda)p(w|T), \quad \lambda = 0.8.$$

For the word 'the' to appear as the first word, plug in the values we have:

$$p('the') = 0.8 * 0.3 + (1-0.8) * 0.3 = 0.24 + 0.06 = 0.3.$$

(b)

Because of the independence, the probability of observing 'the' in the first or second word is the same.

$$p('the' = 2nd) = p('the' = 1st) = 0.3$$

(c)

We can write the following formula for determination of whether a given word is from H:

$$\begin{aligned} p(H|w) &= \frac{p(H,w)}{p(w)} \\ &= \frac{p(w|H)p(H)}{p(w,H) + p(w,T)} \\ &= \frac{p(w|H)p(H)}{p(w|H)p(H) + p(w|T)p(T)}, \end{aligned}$$

where $p(H) = \lambda = 0.8$ and $p(T) = 1 - \lambda = 0.2$. Plug in the values we have

$$\begin{aligned} p(H|'data') &= \frac{\lambda p('data'|H)}{\lambda p('data'|H) + (1-\lambda)p('data'|T)} \\ &= \frac{0.8 * 0.1}{0.8 * 0.1 + 0.2 * 0.1} \\ &= 0.8 \end{aligned}$$

(d)

The least-frequently-occurred word should have the smallest $p(w)$, which can be calculated as:

$$\begin{aligned} &p(w) = \lambda p(w|H) + (1-\lambda)p(w|T), \quad \lambda = 0.8. \\ &p(w) = 0.8 * p(w|H) + 0.2 * p(w|T). \end{aligned}$$

Plug in all the five words, we have the following results:

$$\begin{aligned} &p('the') = 0.8 * 0.3 + 0.2 * 0.3 = 0.3 \\ &p('computer') = 0.8 * 0.1 + 0.2 * 0.2 = 0.12 \\ &p('data') = 0.8 * 0.1 + 0.2 * 0.1 = 0.1 \\ &p('baseball') = 0.8 * 0.2 + 0.2 * 0.1 = 0.18 \\ &p('game') = 0.8 * 0.2 + 0.2 * 0.1 = 0.18 \end{aligned}$$

The smallest is $p('data') = 0.1$, so the least-frequently-occurred word should be 'data'.

(e)

The maximum likelihood estimator is:

$$p(w|H) = \frac{c(w, D)}{|D|},$$

where $c(w, D)$ is the count of a word in document D, and $|D|$ is the length of document D. With this, we can have the following estimator as:

$$\begin{aligned} p('computer'|H) &= \frac{3}{10} = 0.3 \\ p('game'|H) &= \frac{2}{10} = 0.2 \end{aligned}$$

4 EM Algorithm

(a)

$$p(w_i|\theta_2) = p(w_i|\lambda) = \lambda p(w_i|\theta_1) + (1-\lambda)p(w_i|C)$$

(b)

$$\log p(\theta_2) = \sum_{i=1}^{N} \log p(w_i|\theta_2) = \sum_{i=1}^{N} \log \left[\lambda p(w_i|\theta_1) + (1-\lambda)p(w_i|C)\right],$$

where N is the length of document D_2.

(c)

There are N binary hidden variables in total do we need for computing this maximum likelihood estimate using the EM algorithm, with N being the length of document D_2. Because for every one of the words, we need a variable to denote whether this word comes from document D_1 or the background collection C.

(d)

In order to better represent the Lagrange Multiplier, let's denote $\beta = 1 - \lambda$, with the constraint that $\beta + \lambda = 1$. $z_i = 0$ denotes a word w_i comes from document D_1, and $z_i = 1$ denotes that a word w_i comes from background collection C.
The E step in the EM algorithm:

$$p(z_i = 0|\lambda, \beta) = \frac{\lambda p(w_i|\theta_1)}{\lambda p(w_i|\theta_1) + \beta p(w_i|C)}$$

The M step in the EM algorithm: First let's rewrite the objective function with the introduction of Lagrange Multiplier ξ:

$$\log p(\theta_2) = \sum_{i=1}^{N} \log p(w_i|\theta_2) = \sum_{i=1}^{N} \log \left[\lambda p(w_i|\theta_1) + \beta p(w_i|C)\right] + \xi(\lambda + \beta - 1)$$

Taking derivative w.r.t. λ and β, and set to zero we have:

$$\frac{\partial \log p(\theta_2)}{\partial \lambda} = \sum_{i=1}^{N} \frac{p(w_i|\theta_1)}{\lambda p(w_i|\theta_1) + \beta p(w_i|C)} + \xi = 0$$

$$\frac{\partial \log p(\theta_2)}{\partial \beta} = \sum_{i=1}^{N} \frac{p(w_i|C)}{\lambda p(w_i|\theta_1) + \beta p(w_i|C)} + \xi = 0$$

Multiply the two equation with λ and β and sum the resulting two equations, we have $\xi = -N$. Plug this result back into the first equation and multiply by λ, we have

$$\sum_{i=1}^{N} \frac{\lambda p(w_i|\theta_1)}{\lambda p(w_i|\theta_1) + \beta p(w_i|C)} - N\lambda = 0$$

Then we have

$$\begin{aligned}\lambda^n &= \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda^{(n-1)}p(w_i|\theta_1)}{\lambda^{(n-1)}p(w_i|\theta_1)+\beta^{(n-1)}p(w_i|C)} \\ &= \frac{1}{N}\sum_{i=1}^{N}p(z_i=0|\lambda^{(n-1)},\beta^{(n-1)}),\end{aligned}$$

which is the update.