CS598: CXZ	**Fall 2016**
Problem Set 3	
chen zhang, NetId: czhang49	*Handed In: September 24, 2016*

1 Classic Probabilistic Retrieval Model

1.1 (a)

For document generation, using multinomial model, the score can be written as:

$$\begin{aligned} score(Q,D) &= \frac{P(D|Q,R=1)}{P(D|Q,R=0)} \\ &= \frac{\prod_{j=1}^{|V|} P(w_j|Q,R=1)^{c(w_j|Q,R=1)}}{\prod_{j=1}^{|V|} P(w_j|Q,R=0)^{c(w_j|Q,R=0)}} \\ &= -\sum_{x\in\Omega} P(x)\log_2 P(x) \end{aligned}$$

And then we have

$$\begin{aligned} score(Q,D) &\propto \log \frac{P(D|Q,R=1)}{P(D|Q,R=0)} \\ &\propto \log \frac{\prod_{j=1}^{|V|} P(w_j|Q,R=1)^{c(w_j|Q,R=1)}}{\prod_{j=1}^{|V|} P(w_j|Q,R=0)^{c(w_j|Q,R=0)}} \end{aligned}$$

Since the occurrence of a word in the document is independent of the query, we have

$$c(w_j|Q,R=0) = c(w_j|Q,R=1) = c(w_j,D)$$

Then we have

$$\begin{aligned} score(Q,D) &\propto c(w_j,D)\log \frac{\prod_{j=1}^{|V|} P(w_j|Q,R=1)}{\prod_{j=1}^{|V|} P(w_j|Q,R=0)} \\ &\propto c(w_j,D)\log \prod_{j=1}^{|V|} \frac{P(w_j|Q,R=1)}{P(w_j|Q,R=0)} \\ &\propto \sum_{w\in V} c(w,D)\log \frac{P(w|Q,R=1)}{P(w|Q,R=0)} \end{aligned}$$

There's too many equations for this HW, I will write the answers very clearly starting from 1.(b). I apologize for that.

(b) $P(w|Q, R=0) = \frac{c(w|C)}{|C|}$

with $|C|$ representing the total number of words in $C = D_1, \ldots D_n$

(c) $P(w|Q, R=1) = \frac{c(w|q)}{|q|}$

with $|q|$ representing the total number of words in q

(d) $p(w|Q, R=1) = (1-\lambda)\frac{c(w|q)}{|q|} + \lambda P(w|REF)$.

(e)

$$score(Q,D) \propto \sum_{w \in V} c(w,D) \log \frac{P(w|Q,R=1)}{P(w|Q,R=0)}.$$

$$= \sum_{\substack{w \in V \\ w \in q}} c(w,D) \log P(w|Q,R=1) + \sum_{\substack{w \in V \\ w \notin q}} c(w,D) P(w|Q,R=1)$$

$$- \sum_{w \in V} c(w,D) \log P(w|Q,R=0)$$

$$\overset{rank}{=} \sum_{\substack{w \in V \\ w \in q}} c(w,D) \log\left[(1-\lambda)\frac{c(w|q)}{|q|} + \lambda p(w|REF)\right]$$

$$- \sum_{\substack{w \in V \\ w \in q}} c(w,D) \log\left[\lambda p(w|REF)\right]$$

$$+ \sum_{w \in V} c(w,D) \log\left[\lambda p(w|REF)\right] - \sum_{w \in V} c(w,D) \log p(w|Q,R=0)$$

$$= \sum_{\substack{w \in V \\ w \in q}} c(w,D) \log\left[1 + \frac{(1-\lambda)c(w|q)}{|q| \cdot \lambda p(w|REF)}\right]$$

(A) (B)

$$+ \sum_{w \in V} c(w,D) \log\left[\lambda p(w|REF)\right]$$

$$- \sum_{w \in V} c(w,D) \log p(w|Q,R=0).$$

(A) captures TF, (B) captures IDF,

Document length normalization not captured

2. (a) For Jelinek-Mercer,

$$p(w|d) = (1-\lambda)\frac{c(w,d)}{|d|} + \lambda p(w|REF)$$

$$\log p(q|d) = \sum_{i=1}^{m} \log p(q_i|d) = \sum_{i=1}^{|V|} c(w_i,q)\log p(w_i|d)$$

$$= \sum_{w \in q} c(w,q)\log p(w|d).$$

plug in $p(w|d)$ we have

$$\log p(q|d) = \sum_{\substack{w \in q \\ w \in d}} c(w,q)\log\left[(1-\lambda)\frac{c(w,d)}{|d|} + \lambda P(w|REF)\right]$$

$$+ \sum_{\substack{w \in q \\ w \notin d}} c(w,q)\log\left[\lambda P(w|REF)\right]$$

$$= \sum_{\substack{w \in q \\ w \in d}} c(w,q)\log\left[(1-\lambda)\frac{c(w,d)}{|d|} + \lambda P(w|REF)\right]$$

$$- \sum_{\substack{w \in q \\ w \in d}} c(w,q)\log\left[\lambda P(w|REF)\right]$$

$$+ \sum_{w \in q} c(w,q)\log\left[\lambda P(w|REF)\right]$$

$$= \sum_{w \in Q \cap D} c(w,Q) \log\left[1 + \frac{(1-\lambda)\, c(w,D)}{\lambda\, p(w|REF)\,|D|}\right]$$

$$+ \underbrace{\sum_{w \in Q} c(w,q) \log\left[\lambda\, p(w|REF)\right]}_{\text{independent of document}}$$

$$\propto \sum_{w \in Q \cap D} c(w,Q) \log\left[1 + \frac{(1-\lambda)\, c(w,D)}{\lambda\, p(w|REF)\,|D|}\right]$$

(b) $q = (w_{1,q}, w_{2,q} \cdots w_{n,q})$

if word in query then $w_{i,q} = 1$, else $w_{i,q} = 0$

$d = (w_{1,d}, w_{2,d} \cdots w_{t,d})$

if word in doc then $w_{j,d} = 1$, else $w_{j,d} = 0$.

Weight $w_{t,d} = \log\left[1 + \frac{(1-\lambda)\, c(w,D)}{\lambda\, p(w|REF) \times |D|}\right]$ This captures TF & IDF with $c(w,D)$ and $p(w|REF)$

(c) Since for Jelinek-Mercer:

$$\text{score} \propto \sum_{w \in Q \cap D} c(w,Q) \log\left(1 + \frac{(1-\lambda)\, c(w,D)}{\lambda\, p(w|REF) \times |D|}\right)$$

$\frac{c(w,D)}{|D|}$ = const when doubling doc

& $c(w,Q)$ will increase $\Rightarrow$

score should increase when doubling doc $\Rightarrow$

$\Rightarrow$ no over constraint

For Dirichlet prior

$$\text{Score} \propto \sum_{\substack{w \in d \\ w \in q}} \log\left[1 + \frac{c(w,d)}{\mu\, p(w|C)}\right] + n \log \frac{\mu}{|d| + \mu}$$

(A) (B)

when double doc, the increase in (A) is more than the decrease in (B) $\Rightarrow$

$\Rightarrow$ score increases $\Rightarrow$ no over constraints.