

a.

$$\begin{aligned}
Score(Q,D) &= P(R=1|Q,D) = O(R=1|Q,D) = \frac{P(R=1|Q,D)}{P(R=0|Q,D)} \\
&= \frac{P(Q,D|R=1)}{P(Q,D|R=0)} \times \frac{P(R=1)}{P(R=0)} \stackrel{\text{Document generation}}{=} \frac{P(Q,D|R=1)}{P(Q,D|R=0)} \\
&= \frac{P(D|Q,R=1) \times P(Q|R=1)}{P(D|Q,R=0) \times P(Q|R=0)} = \frac{P(D|Q,R=1)}{P(D|Q,R=0)} = \\
&= \prod_{t_i \in D} \frac{P(t_i|Q,R=1)}{P(t_i|Q,R=0)} = \prod_{w \in V} \frac{P(w|Q,R=1)^{c(w,D)}}{P(w|Q,R=0)^{c(w,D)}} \\
&\stackrel{RSJ}{=} \sum_{w \in V} \log \left( \frac{P(w|Q,R=1)^{c(w,D)}}{P(w|Q,R=0)^{c(w,D)}} \right) = \sum_{w \in V} \log \left( \left( \frac{P(w|Q,R=1)}{P(w|Q,R=0)} \right)^{c(w,D)} \right) \\
&\stackrel{3}{=} \sum_{w \in V} c(w,D) \times \log \left( \frac{P(w|Q,R=1)}{P(w|Q,R=0)} \right)
\end{aligned}$$

در این مدل، بایستی 3\*IVI پارامتر بدست آمده و 2\*IVI پارامتر تخمین زده شود. پارامترهای  $P(w|Q,R=1)$  و  $P(w|Q,R=0)$  به ازای هر کلمه در لغت نامه بایستی تخمین زده شوند که جمعاً 2\*IVI پارامتر میشوند.

b.

*Bernouli:*

$$P(w|Q,R=0) = \frac{\#(\text{non rel docs with } w) + 0.5}{\#(\text{non rel docs}) + 1} = \frac{\#(\text{non rel docs with } w) + 0.5}{n+1}$$

*Multinomial:*

$$P(w|Q,R=0) = \frac{\sum_{d \in \text{non rel docs with } w} \frac{c(w,d)}{|d|} + 0.5}{\#(\text{non rel docs}) + 1} = \frac{\sum_{d \in \text{non rel docs with } w} \frac{c(w,d)}{|d|} + 0.5}{n+1} = \frac{\sum_{d \in \text{non rel docs with } w} p(w,d) + 0.5}{n+1}$$

c.

*Bernouli:*

$$P(w|Q,R=1) = \frac{\#(\text{rel docs with } w) + 0.5}{\#(\text{rel docs}) + 1} = \frac{t_w + 0.5}{1+1}$$

$$\begin{cases} t_w = 1, & \text{if } Q \text{ contains } w \\ t_w = 0, & \text{otherwise} \end{cases}$$

*Multinomial:*

$$P(w|Q,R=1) = \frac{\sum_{d \in \text{rel docs with } w} \frac{c(w,Q)}{|Q|} + 0.5}{\#(\text{rel docs}) + 1} = \frac{\sum_{d \in Q} \frac{c(w,Q)}{|Q|} + 0.5}{1+1} = \frac{\sum_{w \in Q} P(w,Q) + 0.5}{2}$$

d.

*Bernouli:*

$$p(w|Q, R=1) = (1-\lambda) \frac{t_w + 0.5}{I+1} + \lambda \times p(w|REF)$$

$$\begin{cases} t_w = 1, & \text{if } Q \text{ contains } w \\ t_w = 0, & \text{otherwise} \end{cases}$$

*Multinomial:*

$$P(w|Q, R=1) = (1-\lambda) \frac{\sum_{w \in Q} P(w, Q) + 0.5}{2} + \lambda \times p(w|REF)$$

e.

$$score(Q, D) = \sum_{w \in V} c(w, D) \log \left( \frac{(1-\lambda) \frac{\sum_{w \in Q} p(w, Q) + 0.5}{2} + \lambda \times p(w|REF)}{\frac{\sum_{d \in \text{non rel docs with } w} p(w, d) + 0.5}{n+1}} \right)$$

از بین سه مکاشفهی گفتهشده، تنها TF را با داشتن  $c(w, D)$  برآورده می‌کند.

۲.

a.

$$\begin{aligned}
& \text{Score}_{\text{Query Generation by JM smoothing}}(Q,D) = \sum_{w \in Q \cap D} c(w,Q) \log \left( \frac{P_{\text{Scen}}(w|D)}{\alpha_d \times p(w|REF)} \right) = \quad 0 \\
& \sum_{w \in Q \cap D} c(w,Q) \log \left( \frac{(1-\lambda) \times c(w,D) / |D| + \lambda \times p(w|REF)}{\lambda \times p(w|REF)} \right) = \quad 1 \\
& \sum_{w \in Q \cap D} c(w,Q) \log \left( 1 + \frac{c(w,D) / |D|}{\mu \times p(w|REF)} \right) = \quad 2 \\
& \sum_{w \in Q \cap D} c(w,Q) \log \left( 1 + \frac{c(w,D)}{\mu \times p(w|REF) \times |D|} \right) = \quad 3
\end{aligned}$$

تساوی 0: براساس تعریف نوشته شده است.

تساوی 1: با جایگذاری مقادیر احتمالات در هموارسازی JM بدست آمده است

تساوی 2: با سادهسازی صورت و مخرج

تساوی 3: با سادهسازی مقدار طول سند

لذا مقدار مورد نظر بدست آمد.

b.

بردار پرسوجو: برداری با طول  $|Q \cap D|$  که مولفهی آلم آن برابر است با  $c(w_i, q)$

بردار سند: برداری با طول  $|Q \cap D|$  که مولفهی آلم آن برابر است با

$$\log \left( \frac{(1-\lambda) * c(w_i, D)}{\lambda * |D| * p(w_i | REF)} + 1 \right)$$

تابع شباهت: ضرب نقطه‌ای، dot product

وزن ترم در بردار سند:

$$\log \left( \frac{(1-\lambda) * c(w_i, D)}{\lambda * |D| * p(w_i | REF)} + 1 \right)$$

بله، مکاشفات وزندهی TF-IDF و هموارسازی طول سند را به ترتیب به خاطر وجود مولفه‌های

$$\frac{1}{p(w_i|REF)} * c(w_i, D) = TF-IDF$$

و

$$\frac{c(w_i, D)}{|D|}$$

بر آورده می‌کند.

چون

$$\frac{1}{p(w_i|REF)} = \frac{\# REF docs}{\# REF docs with w_i} \propto \frac{N+1}{df} = IDF$$

C.

درستی برای Jelinek-Mercer

$$\begin{aligned} Score_{JM}(Q, D') &= \sum_{w \in Q \cap D'} c(w, Q) \log \left( 1 + \frac{(1-\lambda) \times c(w, D')}{\lambda \times p(w|REF) \times |D'|} \right) = \\ & \sum_{w \in Q \cap D'} c(w, Q) \log \left( 1 + \frac{(1-\lambda) \times k * c(w, D)}{\lambda \times p(w|REF) \times k |D|} \right) = \\ & \sum_{w \in Q \cap D} c(w, Q) \log \left( 1 + \frac{(1-\lambda) \times c(w, D)}{\lambda \times p(w|REF) \times |D|} \right) = Score_{JM}(Q, D) \\ Score_{JM}(Q, D') &= Score_{JM}(Q, D) \end{aligned}$$

تساوی 0: براساس تعریف نوشته شده است.

تساوی 1: با جایگذاری مقادیر  $c(w, D')$  و  $|D'|$  بدست آمده است که هر کدام نسبت به مقادیر معادل در سند  $D$ ،  $k$  برابر می‌شوند.

تساوی 2: با ساده‌سازی  $k$  از صورت و مخرج

همچنین با توجه به آنکه سند  $D'$  همان سند  $D$  است و تنها  $k$  برابر شده است، مجموعه کلمات موجود در آن تغییر نکرده و به همین خاطر زیر وند سیگما به  $D$  تغییر یافته است.

تساوی 3: براساس تعریف

لذا برای هر پرسمان، امتیاز سند با  $k$  برابر کردن سند، ثابت به می‌ماند.

درستی برای Dirichlet

$$\begin{aligned}
Score_{Dir}(Q, D') &= \sum_{w \in Q \cap D'} c(w, Q) \log \left( \frac{c(w, D') + \mu \times p(w|REF)}{\mu \times p(w|REF)} \right) \quad 1 \\
&\sum_{w \in Q \cap D} c(w, Q) \log \left( \frac{k \times c(w, D) + \mu \times p(w|REF)}{\mu \times p(w|REF)} \right) \geq \quad 2 \\
&\sum_{w \in Q \cap D} c(w, Q) \log \left( \frac{c(w, D) + \mu \times p(w|REF)}{\mu \times p(w|REF)} \right) = Score_{Dir}(Q, D) \quad 3 \\
Score_{Dir}(Q, D') &\geq Score_{Dir}(Q, D)
\end{aligned}$$

تساوی 0 : براساس تعریف نوشته شده است.

تساوی 1 : با جایگذاری مقادیر  $c(w, D')$  و بدست آمده است که نسبت به مقادیر معادل در سند  $D$ ،  $k$  برابر می شوند.  
نامساوی 2 : با ساده سازی  $k$  از صورت، و اینکه  $k$  بزرگتر مساوی ۱ بوده و تابع لگاریتم صعودی است.  
همچنین با توجه به آنکه سند  $D'$  همان سند  $D$  است و تنها  $k$  برابر شده است، مجموعه کلمات موجود در آن تغییر نکرده و به همین خاطر زیر وند سیگما به  $D$  تغییر یافته است.  
تساوی 3 : براساس تعریف  
لذا برای هر پرسمان، امتیاز سند با  $k$  برابر کردن سند، بزرگتر مساوی خواهد شد. که همان مورد خواسته شده در سوال است.

۳.

الف.

	TF-IDF	Okapi	KL-Divergence
P_5	0.4161	0.4107	0.4510
P_10	0.4107	0.3933	0.4262
map	0.2601	0.2610	0.2643

در این بین، نتایج TF-IDF و Okapi تقریباً مشابه است، اما نتایج KL-Divergence از دیگر نتایج بهتر است. پیشرفت ها در تمامی سه معیار وجود داشته است.

ب.

	Jelinek-Mercer	Bayesian/Dirichlet prior	Abs. Discount
P_5	0.4081	0.4416	0.4309
P_10	0.3987	0.4168	0.4081
map	0.2441	0.2638	0.2538

در صورتی که پارامترها تنظیم نشوند، هموارسازی کارساز مفید نخواهد بود، به همین علت پیشرفت چشمگیری در نتایج قسمت سوم، نسبت به قسمت دوم، حتی با وجود هموارسازی نداریم. حتی نتایج مقداری بدتر شده اند که می‌تواند به علت تنظیم نشدن پارامترها باشد. همچنین نوع پرسوجو ها از قبیل بلند و کوتاه بودن آن‌ها، در تاثیر هموارسازی موثر است.

ج.

Dirichlet Prior	500	750	800	900	1000	1250	1300	1500	1750	2000	3000
P_5	0.4497	0.4523	0.4523	0.4510	0.4510	0.4523	0.4497	0.4483	0.4376	0.4416	0.4389
P_10	0.4195	0.4221	0.4248	0.4242	0.4262	0.4255	0.4255	0.4208	0.4154	0.4168	0.4114
map	0.2602	0.2629	0.2633	0.2642	0.2643	0.2645	0.2646	0.2647	0.2643	0.2638	0.2609

در این قسمت با معیار قرار دادن مقدار پیش فرض 2000 آغاز کردیم و با روش نصف کردن آزمایشات را انجام دادیم، در هر بازه‌ای که مقدار مناسب تری و به طور صعودی پیشرفت در مقادیر مقایسه بدست آوردیم، آن بازه را نصف کردیم تا به مقدار بهینه برسیم.

مقیاس ارزیابی مقدار بهینه برای یافتن بهترین مقدار در Dirichlet prior را معیار map قرار دادیم. نمودار خواسته شده نیز، طبق انتظار نزولی است.

## Precision-Recall Curve

