۱.

a.

$$Score(Q,D) \overset{r}{=} P(R=1|Q,D) \overset{r}{=} O(R=1|Q,D) \overset{r}{=} \frac{P(R=1|Q,D)}{P(R=0|Q,D)}$$

$$= \frac{P(Q,D|R=1)}{P(Q,D|R=0)} \times \frac{P(R=1)}{P(R=0)} \overset{r}{=} \frac{P(Q,D|R=1)}{P(Q,D|R=0)} \overset{Document\ \ generation}{=}$$

$$\frac{P(D|Q,R=1)\times P(Q|R=1)}{P(D|Q,R=0)\times P(Q|R=0)} \overset{r}{=} \frac{P(D|Q,R=1)}{P(D|Q,R=0)} =$$

$$\prod_{t_i \in D} \frac{P(t_i|Q,R=1)}{P(w|Q,R=0)} \overset{1}{=} \prod_{w \in V} \frac{P(w|Q,R=1)^{c(w,D)}}{P(w|Q,R=0)^{c(w,D)}}$$

$$\overset{RSJ}{=} \sum_{w \in V} log\left(\frac{P(w|Q,R=1)^{c(w,D)}}{P(w|Q,R=0)^{c(w,D)}}\right) \overset{2}{=} \sum_{w \in V} log\left(\left(\frac{P(w|Q,R=1)}{P(w|Q,R=0)}\right)^{c(w,D)}\right)$$

$$\overset{3}{=} \sum_{w \in V} c(w,D) \times log\left(\frac{P(w|Q,R=1)}{P(w|Q,R=0)}\right)$$

در این مدل، بایستی 3*|V| پارامتر بدست آمده و 2*|V| پارامتر تخمین زده شود.
پارامتر های P(w|Q,R=1) و P(w|Q,R=1) به ازای هر کلمه در لغت نامه بایستی تخمین زده شوند که جمعا 2*|V| پارامتر میشود.

b.

Bernouli:

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

$$\frac{\#(non\ \ rel\ \ docs\ \ with\ \ w) + 0.5}{n+1}$$

Multinominal:

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

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

c.

Bernouli:

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

$$\frac{t_w\ +0.5}{1+1}$$

$$\left\{\begin{array}{ll} t_w = 1, & if\ \ Q\ \ contains\ \ w \\ t_w = 0, & otherwise \end{array}\right\}$$

Multinominal:

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

$$\frac{\sum_{d \in Q} \frac{c(w,Q)}{|Q|} +0.5}{1+1} = \frac{\sum_{w \in Q} P(w,Q) +0.5}{2}$$

d.

Bernouli:

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

$$\left\{\begin{array}{ll} t_w = 1, & if\ \ Q\ \ contains\ \ w \\ t_w = 0, & otherwise \end{array}\right\}$$

Multinominal:

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

e.

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

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

۲.

a.

$$Score_{Query\ Generation\ by\ JM\ smoothing}(Q,D)\overset{0}{=}$$

$$\sum_{w\in Q\cap D} c(w,Q)\log\left(\frac{P_{Seen}(w|D)}{\alpha_d\times p(w|REF)}\right)\overset{1}{=}$$

$$\sum_{w\in Q\cap D} c(w,Q)\log\left(\frac{(1-\lambda)\times c(w,D)/|D|+\lambda\times p(w|REF)}{\lambda\times p(w|REF)}\right)\overset{2}{=}$$

$$\sum_{w\in Q\cap D} c(w,Q)\log\left(1+\frac{c(w,D)/|D|}{\mu\times p(w|REF)}\right)\overset{3}{=}$$

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

تساوی 0 : براساس تعریف نوشته‌شده‌است.

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∩D| که مولفه‌ی iام آن برابر است با $c(w_i, q)$
بردار سند: برداری با طول |Q∩D| که مولفه‌ی iام آن برابر است با

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

تابع شباهت: ضرب نقطه‌ای، dot product
وزن ترم در بردار سند:

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

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 آورده‌می‌کند.
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

$$Score_{JM}(Q,D') \overset{0}{=} \sum_{w \in Q \bigcap D'} c(w,Q) log(1 + \frac{(1-\lambda) \times c(w,D')}{\lambda \times p(w|REF) \times |D'|}) \overset{1}{=}$$

$$\sum_{w \in Q \bigcap D'} c(w,Q) log(1 + \frac{(1-\lambda) \times k * c(w,D)}{\lambda \times p(w|REF) \times k|D|}) \overset{2}{=}$$

$$\sum_{w \in Q \bigcap D} c(w,Q) log(1 + \frac{(1-\lambda) \times c(w,D)}{\lambda \times p(w|REF) \times |D|}) \overset{3}{=} Score_{JM}(Q,D)$$

$$Score_{JM}(Q,D') = Score_{JM}(Q,D)$$

تساوی 0 : براساس تعریف نوشته‌شده‌است.

تساوی 1 : با جایگذاری مقادیر c(w, D') و |D'| بدست‌آمده 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

$$Score_{Dir}(Q,D') \overset{0}{=}$$

$$\sum_{w \in Q \cap D'} c(w,Q) log\left(\frac{c(w,D') + \mu \times p(w|REF)}{\mu \times p(w|REF)} \right) \overset{1}{=}$$

$$\sum_{w \in Q \cap D} c(w,Q) log\left(\frac{k \times c(w,D) + \mu \times p(w|REF)}{\mu \times p(w|REF)} \right) \overset{2}{\geq}$$

$$\sum_{w \in Q \cap D} c(w,Q) log\left(\frac{c(w,D) + \mu \times p(w|REF)}{\mu \times p(w|REF)} \right) \overset{3}{=} Score_{Dir}(Q,D)$$

$$Score_{Dir}(Q,D') \geq Score_{Dir}(Q,D)$$

تساوی 0 : براساس تعریف نوشته‌شده‌است.

تساوی 1 : با جایگذاری مقادیر c(w, D') و بدست‌آمده است که نسبت به مقادیر معادل در سند D ، k برابر می‌شوند.

نامساوی 2 : با ساده‌سازی 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 خواهدشد. که همان مورد خواسته‌شده 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 تقریبا مشابه است، اما نتایج KL-Divergence از دیگر نتایج، در تمامی سه معیار بهتربوده‌است.

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 مفیدنخواهد 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.
به طور کلی، روش‌های هموارسازی، دقت و فراخوانی را بالا میبرند، اما اگر پارامترها به صورت مناسب تنظیم نشوند، روش‌های هموارسازی می‌توانند دقت و فراخوانی را کاهش دهند. با این وجود، با توجه به پارامترهای داده‌شده در تمرین، دقت در۵ و ۱۰ سند اول و همچنین map، با هر سه روش هموارسازی پایین می‌آید. اما
دسته‌بندی نشدن پرس‌وجو ها، پارامتر مناسب برای هر نوع پرس‌وجو
انواع پرس‌وجو
به این ترتیب، در بخش‌بعدی، پارامتر روش دیریشله را تنظیم می‌کنیم.
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.

پس از تنظیم پارامتر در هموارسازی دیریشله، به مقداری از map بهتر از مقدار در بخش الف، بدون هموارسازی دست یافتیم.

نمودار خواسته شده نیز، طبق انتظار نزولی است.
شیب نمودار از ریکال صفر به ۰.۱ ناگهان شدت دارد اما در دیگر موارد شیب ثابت مانده‌است.
اما از بین روش‌های هموار سازی، روش Dirichlet prior از دیگر روش‌های هموارسازی کمی بهتر عمل کرده است.