A Full derivation

A.1 Optimizing objective

We want to choose S_k^* that maximizes the objective:

$$\text{Exp-}n\text{-Call@}k(S_k, \mathbf{q}) = \mathbb{E}[R_k \geq n | s_1, \ldots, s_k, \mathbf{q}]$$

By taking a greedy approach, we select s_k^* given S_{k-1}^*:

$$\begin{aligned}
s_k^* &= \arg\max_{s_k} \mathbb{E}[R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}] \\
&= \arg\max_{s_k} P(R_k \geq n | S_{k-1}^*, s_k, \mathbf{q}) && (1) \\
&= \arg\max_{s_k} \sum_{T_k} \Big(P(t|\mathbf{q})\, P(t_k|s_k) \Big(\prod_{i=1}^{k-1} P(t_i|s_i^*) \Big) \cdot P(R_k \geq n | T_k, S_{k-1}^*, s_k, \mathbf{q}) \Big) && (2) \\
&= \arg\max_{s_k} \sum_{T_k} P(t|\mathbf{q})\, P(t_k|s_k) \Big(\prod_{i=1}^{k-1} P(t_i|s_i^*) \Big) \cdot \Big(\underbrace{P(r_k \geq 0 | R_{k-1} \geq n, t_k, t)}_{1} P(R_{k-1} \geq n | T_{k\text{-}1}) \\
&\qquad + P(r_k = 1 | R_{k-1} = n-1, t_k, t) P(R_{k-1} = n-1 | T_{k\text{-}1}) \Big) && (3) \\
&= \arg\max_{s_k} \Bigg(\sum_{T_{k-1}} \Big[\underbrace{\sum_{t_k} P(t_k|s_k)}_{1} \Big] P(t|\mathbf{q}) \Big(\prod_{i=1}^{k-1} P(t_i|s_i^*) \Big) P(R_{k-1} \geq n | T_{k\text{-}1}) + \\
&\qquad \sum_{T_{k\text{-}1}} \Big[\sum_{t_k} P(t_k|s_k) P(r_k = 1 | t_k, t) \Big] P(t|\mathbf{q}) \Big(\prod_{i=1}^{k-1} P(t_i|s_i^*) \Big) P(R_{k-1} = n-1 | T_{k\text{-}1}) \Bigg) \\
&= \arg\max_{s_k} \sum_t P(t|\mathbf{q}) P(t_k = t | s_k) \Big[\sum_{t_1, \ldots, t_{k-1}} P(R_{k-1} = n-1 | T_{k\text{-}1}) \prod_{i=1}^{k-1} P(t_i|s_i^*) \Big] && (4) \\
&= \arg\max_{s_k} \sum_t P(t|\mathbf{q}) P(t_k = t | s_k) P(R_{k-1} = n-1 | S_{k-1}^*) && (5)
\end{aligned}$$

Note:
(1) Since $(R_k \geq n)$ can only be zero or one in probability.
(2) Marginalize out T_k.
(3) Split $(R_k \geq n)$ into two disjoint events $(r_k \geq 0, R_{k-1} \geq n)$, $(r_k = 1, R_{k-1} = n-1)$, conditioned on R_{k-1}.
(4) Drop the first line as it does not involve s_k and has no influence in determining s_k^*.
Note that $\sum_{t_k} P(t_k|s_k) P(r_k = 1 | t_k, t) = \sum_{t_k} P(t_k|s_k) \mathbb{I}[t_k = t] = P(t_k = t | s_k)$, where t is implicitly conditioned and is not explicitly shown here.
(5) This objective is recursively defined.

By similar reasoning, the probability needed in (5) is recursively defined as

$$P(R_k = n | S_k, t) = \begin{cases}
n \geq 1, k > 1: & \big(1 - P(t_k = t | s_k)\big) P(R_{k-1} = n | S_{k-1}, t) \\
& \quad + P(t_k = t | s_k) P(R_{k-1} = n-1 | S_{k-1}, t) \\
n = 0, k > 1: & \big(1 - P(t_k = t | s_k)\big) P(R_{k-1} = 0 | S_{k-1}, t) \\
n = 1, k = 1: & P(t_1 = t | s_1) \\
n = 0, k = 1: & 1 - P(t_1 = t | s_1) \\
n > k: & 0
\end{cases}$$

For expected n-call@k where $n \leq k/2$, by unrolling its recursive definition in (5), the explicit objective is

$$s_k^* = \arg\max_{s_k} \sum_t \Bigg(P(t|\mathbf{q})\, P(t_k = t | s_k) \sum_{j_1, \ldots, j_{n-1}} \prod_{l \in \{j_1, \ldots, j_{n-1}\}} P(t_l = t | s_l^*) \prod_{\substack{i=1 \\ i \notin \{j_1, \ldots, j_{n-1}\}}}^{k-1} \big(1 - P(t_i = t | s_i^*)\big) \Bigg) \tag{6}$$

where $j_1, \ldots, j_{n-1} \in \{1, \ldots, k-1\}$ satisfy that $j_i < j_{i+1}$ (i.e., an ordered permutation of $n-1$ result set indices).

Similarly, for expected n-call@k where $n > k/2$, the explicit objective is

$$s_k^* = \arg\max_{s_k} \sum_t \Bigg(P(t|\mathbf{q})\, P(t_k = t | s_k) \sum_{j_n, \ldots, j_{k-1}} \prod_{l \in \{j_n, \ldots, j_{k-1}\}} \big(1 - P(t_l = t | s_l^*)\big) \prod_{\substack{i=1 \\ i \notin \{j_n, \ldots, j_{k-1}\}}}^{k-1} P(t_i = t | s_i^*) \Bigg) \tag{7}$$

where $j_n, \ldots, j_{k-1} \in \{1, \ldots, k-1\}$ satisfy that $j_i < j_{i+1}$ (i.e., an ordered permutation of $k-n$ result set indices).

A.2 Relation to MMR: expected n-call@k when $n > k/2$

Assuming that $\forall i\ P(t_i|s_i) \in \{0,1\}$ and $P(t|\mathbf{q}) \in \{0,1\}$. It is possible to write

$$\prod_{l\in\{j_n,\ldots,j_{k-1}\}}\left(1-P(t_l{=}t|s_l^*)\right) = 1-\left(1-\prod_{l\in\{j_n,\ldots,j_{k-1}\}}\left(1-P(t_l{=}t|s_l^*)\right)\right) = 1-\left(\max_{l\in\{j_n,\ldots,j_{k-1}\}} P(t_l{=}t|s_l^*)\right)$$

This allows us to rewrite (7)

$$\begin{aligned} s_k^* = \arg\max_{s_k} \sum_t \Bigg(& P(t|\mathbf{q})P(t_k{=}t|s_k) \sum_{j_n,\ldots,j_{k-1}} \prod_{\substack{i=1\\ i\notin\{j_n,\ldots,j_{k-1}\}}}^{k-1} P(t_i{=}t|s_i^*) \\ & - P(t|\mathbf{q})P(t_k{=}t|s_k) \sum_{j_n,\ldots,j_{k-1}} \prod_{\substack{i=1\\ i\notin\{j_n,\ldots,j_{k-1}\}}}^{k-1} P(t_i{=}t|s_i^*) \max_{l\in\{j_n,\ldots,j_{k-1}\}} P(t_l{=}t|s_l^*) \Bigg) \end{aligned} \tag{8}$$

Assuming m relevant documents are already selected in the $k-1$ collection, then the top term (specifically $\prod_i$) is non-zero $\binom{m}{n-1}$ times. For the bottom term, it takes $n-1$ relevant documents to satisfy its $\prod_i$, and one additional relevant document to satisfy the $\max_l$ making it non-zero $\binom{m}{n}$ times. Factoring out the max element from the bottom and pushing the $\sum_t$ inwards (all legal due to the $\{0,1\}$ subtopic probability assumption), (8) becomes

$$\begin{aligned} s_k^* &= \arg\max_{s_k} \left[\sum_t P(t|\mathbf{q})P(t_k{=}t|s_k)\binom{m}{n-1}\right] - \left[\sum_t P(t|\mathbf{q})P(t_k{=}t|s_k)\binom{m}{n}\underbrace{\max_{s_i\in S_{k-1}^*} P(t_i{=}t|s_i)}_{1}\right] \\ &= \arg\max_{s_k} \binom{m}{n-1}\underbrace{\sum_t P(t|\mathbf{q})P(t_k{=}t|s_k)}_{\text{relevance: } \mathrm{Sim}_1(s_k,\mathbf{q})} - \binom{m}{n}\max_{s_i\in S_{k-1}^*}\underbrace{\sum_t P(t_i{=}t|s_i)P(t|\mathbf{q})P(t_k{=}t|s_k)}_{\text{diversity: } \mathrm{Sim}_2(s_k,s_i,\mathbf{q})} \qquad (9) \\ &= \arg\max_{s_k} \frac{n}{m+1}\mathrm{Sim}_1(s_k,\mathbf{q}) - \frac{m-n+1}{m+1}\max_{s_i\in S_{k-1}^*}\mathrm{Sim}_2(s_k,s_i,\mathbf{q}) \qquad (10) \end{aligned}$$

Note:
(9) We can rearrange "$\sum_t P(t|\mathbf{q})\max_{s_i}\cdots$" as "$\max_{s_i}\sum_t P(t|\mathbf{q})\cdots$" since the $\sum_t P(t|\mathbf{q})$ 'selects' the only t for which $P(t|\mathbf{q}) = 1$.
(10) Normalize by dividing the equation by $\binom{m}{n-1} + \binom{m}{n} = \binom{m+1}{n}$ (Pascal's rule).
The result is the same as the case where $n \le k/2$.

The reason that we do not remove the max term in (9) is that this allows us to compare the objective with MMR directly. Also, leaving the max term suggests an approximate form for the case where the subtopic probabilities are non-deterministic (not strictly 0 or 1), and approaches (9) as the probabilities become more deterministic.

In practice, under the greedy approach of the expected n-call@k in selecting S_k^*, we expect that there are already n relevant documents chosen in the set $S_{k-1}^* = \{s_1^*,\ldots,s_{k-1}^*\}$ (where $n << k$). In expectation, $m = n$ and hence the optimizing objective can be thought to be

$$s_k^* = \arg\max_{s_k} \frac{n}{n+1}\mathrm{Sim}_1(s_k,\mathbf{q}) - \frac{1}{n+1}\max_{s_i\in S_{k-1}^*}\mathrm{Sim}_2(s_k,s_i,\mathbf{q}) \tag{11}$$

From (11), it is simple to see that the diversification level decreases with n.

B Additional derivation

B.1 Alternative derivation for expected 2-call@k

$$
\begin{aligned}
s_k^* =& \arg\max_{s_k} \mathbb{E}\big[R_k \geq 2 \,\big|\, S_{k-1}^*, s_k, \mathbf{q}\big] \\
=& \arg\max_{s_k} \mathbb{E}\Big[(r_1 = 1 \wedge r_2 = 1) \vee (r_1 = 0 \wedge r_2 = 1 \wedge r_3 = 1) \vee (r_1 = 1 \wedge r_2 = 0 \wedge r_3 = 1) \vee \\
& \quad (r_1 = 0 \wedge r_2 = 0 \wedge r_3 = 1 \wedge r_4 = 1) \vee (r_1 = 0 \wedge r_2 = 1 \wedge r_3 = 0 \wedge r_4 = 1) \vee \\
& \quad (r_1 = 1 \wedge r_2 = 0 \wedge r_3 = 0 \wedge r_4 = 1) \vee \cdots \vee \\
& \quad (r_1 = 0 \wedge \cdots \wedge r_{k-2} = 0 \wedge r_{k-1} = 1 \wedge r_k = 1) \vee \\
& \quad (r_1 = 0 \wedge \cdots \wedge r_{k-3} = 0 \wedge r_{k-2} = 1 \wedge r_{k-1} = 0 \wedge r_k = 1) \vee \cdots \vee \\
& \quad (r_1 = 1 \wedge r_2 = 0 \wedge \cdots \wedge r_{k-1} = 0 \wedge r_k = 1) \Big| S_{k-1}^*, s_k, \mathbf{q} \Big] \\
=& \arg\max_{s_k} \mathbb{E}\Big[(r_1 = 1 \wedge r_2 = 1) \vee (r_1 = 0 \wedge r_2 = 1 \wedge r_3 = 1) \vee (r_1 = 1 \wedge r_2 = 0 \wedge r_3 = 1) \vee \\
& \quad (r_1 = 0 \wedge r_2 = 0 \wedge r_3 = 1 \wedge r_4 = 1) \vee (r_1 = 0 \wedge r_2 = 1 \wedge r_3 = 0 \wedge r_4 = 1) \vee \\
& \quad (r_1 = 1 \wedge r_2 = 0 \wedge r_3 = 0 \wedge r_4 = 1) \vee \cdots \vee \\
& \quad \bigvee_{j=1}^{k-1} \left(r_k = 1 \wedge \bigwedge_{\substack{i=1 \\ i \neq j}}^{k-1} r_i = 0 \wedge r_j = 1 \right) \Big| S_{k-1}^*, s_k, \mathbf{q} \Big] \\
=& \arg\max_{s_k} \sum_{j=1}^{k-1} P\left(r_k = 1 \wedge \bigwedge_{\substack{i=1 \\ i \neq j}}^{k-1} r_i = 0 \wedge r_j = 1 \,\middle|\, S_{k-1}^*, s_k, \mathbf{q} \right) \\
=& \arg\max_{s_k} \sum_{j=1}^{k-1} \left(\sum_{t_1, \ldots, t_k, t} P(t|\mathbf{q})\, P(t_k|s_k) \mathbb{I}[t_k = t]\, P(t_j|s_j^*) \mathbb{I}[t_j = t] \prod_{\substack{i=1 \\ i \neq j}}^{k-1} P(t_i|s_i^*) \mathbb{I}[t_i \neq t] \right) \\
=& \arg\max_{s_k} \sum_{t} P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j=1}^{k-1} \left(P(t_j = t|s_j^*) \prod_{\substack{i=1 \\ i \neq j}}^{k-1} (1 - P(t_i = t|s_i^*)) \right)
\end{aligned}
$$

Assuming that $\forall i\ P(t_i|s_i) \in \{0, 1\}$ and $P(t|\mathbf{q}) \in \{0, 1\}$, the objective becomes:

$$
\begin{aligned}
s_k^* =& \arg\max_{s_k} \sum_{t} P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j=1}^{k-1} \left(P(t_j = t|s_j^*) \prod_{\substack{i=1 \\ i \neq j}}^{k-1} (1 - P(t_i = t|s_i^*)) \right) \\
=& \arg\max_{s_k} \sum_{t} P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j=1}^{k-1} \left(P(t_j = t|s_j^*) \left[1 - \left(1 - \prod_{\substack{i=1 \\ i \neq j}}^{k-1} (1 - P(t_i = t|s_i^*)) \right) \right] \right) \\
=& \arg\max_{s_k} \sum_{t} P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j=1}^{k-1} \left[P\big(t_j = t|s_j^*\big) - P\big(t_j = t|s_j^*\big) \left(1 - \prod_{\substack{i=1 \\ i \neq j}}^{k-1} (1 - P(t_i = t|s_i^*)) \right) \right] \\
=& \arg\max_{s_k} \sum_{t} P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j=1}^{k-1} P(t_j = t|s_j^*) \\
& - \sum_{t} P(t|\mathbf{q})\, P(t_k = t|s_k) \sum_{j=1}^{k-1} P(t_j = t|s_j^*) \max_{\substack{i \in [1, k-1] \\ i \neq j}} P(t_i = t|s_i^*)
\end{aligned}
$$

Noting that this is of the same form as (8), albeit much simpler.