Retrieval 평가와 메트릭

Retrieval Metric
- ① Rank-Unaware (순서를 고려하지 않는 Metric)
  - ⇝ Recall, Precision, F1
- ② Rank-aware (순서를 고려하는 Metric) ; 전통적인 검색 시스템에서 사용되는 방법
  - ⇝ mAP, mRR, NDCG

평가방식 ⇝ 'Retrieval contents'와 'Retrieval GT'의 ID를 비교해서 채점하는 방식 채택

'Retrieval contents' → 검색 결과
'Retrieval GT' → 정답

⇝ 즉, 정답에 존재한 ID를 가진 내용을 잘 가지고 있는지 검사하는 것.

		실제 정답	
		True	False
분류 결과	True	True Positive	False Positive
	False	False Negative	True Negative

- **Precision**
  - ⇝ $\frac{\text{Retrieved contents 중 정답 수}}{\text{Retrieved contents의 전체 수}}$ ⇝ 검색해온 전체 수 (= Top k)
  - ⇝ 'Positive 정답률', 'PPV(Positive Predictive Value)'라고도 불림
  - 즉, True라고 예측한 것 중, 실제 True인 것의 비율

- **Recall (Sensitivity, hit rate)** → Retrieval 평가에서는 '정답을 가져오긴 했는지'가 매우 중요하기 때문에 Recall도 많이 봄
  - ⇝ $\frac{\text{GT 중 맞은 정답 수}}{\text{Retrieved GT의 전체수}}$ ⇝ 정답 검색에 고려된 GT 수 / ⇝ 전체 GT 수
  - ⇝ 실제 True인 것 중, True라고 예측한 것의 비율
  - ⇝ 'Auto RAG' 프레임워크에서는 GT가 기본적으로 2개인 리스트임
  - ⇝ 정답을 맞히더라도 답변을 한 숫자가 X 패널티
  - ⇝ Top-k를 1개면 Recall도 올라가지만 그만큼 다른 오답도 같이 가져와서 '환각' 확률이 ↑

ex)

Retrieved Contents (ID)	
test-1	o
Pred-1	x
test-2	o
Pred-	x

Retrieval GT	
[test-1, test-2]	
o	o
[tes-3]	
x	

「여러 ID가 1개만 있어도 해당 ID는 정답처리」

test-1 or test-2
AND
test-3

⇒ $\frac{\text{GT중 정답 처리수}}{\text{전체 gt 수}} = \frac{1}{2} = 0.5$

o F1 Score

↝ 'Precision'과 'Recall'을 대체로 반영해
이를 보정하기 위해 'F1-score'사용

↝ $2 \times \frac{Precision \times Recall}{Precision + Recall}$ (Precision과 Recall의 조화평균)

↝ 0에서 사이의 값, 1에 가까운 수록 성능이 좋다.

※ Top-k를 50으로 선정 후, 검색결과 임베딩 모델별 변화 비교
↝ 1, 2, … 부터 50까지 Top-k를 늘리면 Recall이 증가할 확률 (거의 1에 가깝게 도달)
↝ 즉, 많이 가져와서 정답이 포함될 수 있도록 한 다음,
결과를 ReRanker에 넣어서 검색 결과 재 조정.

⇒ 이렇게 해보려고함.

---

o MRR (Mean Reciprocal Rank) ; Mean RR

↝ 정답 단락(Passage)가 몇 번째로 추천했는 지(RR)에 대한 평균

ex)

Retrieved Contents(ID)	
ID-1	X
ID-2	O
ID-3	X
ID-4	O

Retrieval GT
ID-2
ID-4

① (GT ID-2의 RR)
→ Contents에서 2번째 순서로 등장
→ $\frac{1}{\text{맞춘 수치 까지의 수}} = \frac{1}{2} = 0.5$

② (GT ID-4의 RR)
→ Contents에서 4번째 순서로 등장
→ $\frac{1}{4} = 0.25$

③ RR의 평균(MRR) 계산
$\frac{0.5+0.25}{2} = 0.375$

○ mAP (Mean Average Precision) ; Rank-unaware의 Precision과 다른 Precesion

↝ AP(Average Precision)의 평균 (평균의 평균)

step 1) 각 gt 당 precision 계산 (precision)

step 2) gt들의 Precision의 평균 계산 (Average Precision)

step 3) AP의 평균 계산 (mAP)

ex)

Retrieved Contents (ID)	
test-1	o
Pred-1	x
test-2	o
Pred-	x

Retrieval GT: [test-1, test-2] o o, [tes-3] x

step 1)

① GT 'test-1'이 첫번째로 등장 ⇒ $\frac{1}{1}=1$

② test-1이 첫번째, test-2가 세번째에 등장

step 2) $\frac{1+2/3}{2} = \frac{5}{6}$, $\frac{0}{4} = 0$

step 3) $\frac{5/6+0}{2} = \frac{5}{12}$ (mAP)

⇒ $\frac{\text{포함된 GT 수}}{\text{반환된 개체의 수}} = \frac{2}{3} = 0.66$

○ NDCG (Normalized Discounted Cumulative Gain)

↝ 상위 순위에 더 높은 가중치를 부여, 결과의 순위 품질을 0~1로 평가

step 1) CG (Cumulated Gain) ↝ 관련성의 누적 합, 순위고려X ($CG_p = \sum_{i=1}^{p} rel_i$)

Auto RAG에서는 0/1 만 사용 (관련성 정수 개념 미활용X)

step 2) DCG (Discounted Cumulated Gain) ↝ 관련성을 크기로 쓰고 순위도 내어 합한 값 ($DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)}$)

갈수록 관련성에 중가, 관련성 영향력이 후순위 일수록 ↓

step 3) IDCG (Ideal Discounted Cumulated Gain) ↝ DCG의 가장 이상적인 값 (실제 정답 결과와는 상관X)

↝ 즉 검색 순위 결과가 가장 좋은 때의 PCG

step 4) NDCG (Normalized Discounted Cumulative Gain) ↝ $\frac{DCG}{IDCG}$, 이상적 상태에 얼마나 가까운가?

0~1사이, 1에 가까울수록 좋음.