

테스트 계획 및 결과보고서 자체 LLM모델

- 1. 테스트 계획
 - 목적
 - 평가 방법
 - 테스트 방법
- 2. 테스트 진행
 - 테스트 프로세스
- 3. 테스트 결과
 - 유사도 평균 스코어
 - 히스토그램
 - 결과 정리
- 4. 결론 및 계획
 - 결론
 - 계획

1. 테스트 계획

목적

- 모델의 성능 평가를 위해 테스트 케이스 답과 예측값을 비교하고 평가

평가 방법

- 의미적으로 유사한 방법을 측정하기 위한 `sentence_transformers` 사용
- 임베딩 모델은 한국어 모델 중 하나로 `jhgan/ko-sroberta-multitask` 사용

테스트 방법

- 테스트 케이스 154건으로 진행
- 각 테스트의 질문에 대한 추론과 답의 실제값과 예측값 비교

▼ 테스트 케이스 예시

```
1 {
2   'answer_ground_truth': '항해사면허 중 4급부터 6급까지의 상선면허 또는 어선면허를 받으려는 자의 승무경력은 관련 법령에서 정
3   'answer_prediction': '「선박직원법 시행령」 별표 1의3 제1호의 비고 제3호 본문에 따라, 항해사면허 중 4급부터 6급까지의 면
4   'cot_ground_truth': '「선박직원법」 제4조에 따르면, 항해사 면허를 취득하려는 자는 해당 면허의 종류에 따라 상선 또는 어선어
5   'cot_prediction': '「선박직원법」 제4조제2항 및 시행령 제4조제1항은 해기사 면허를 부여할 때 승무경력을 기준으로 하며, 특히
6 }
```

2. 테스트 진행

테스트 프로세스

- 각 데이터의 항목을 임베딩
- 실제값과 예측값의 임베딩값을 가지고 코사인 유사도 계산
- 측정된 유사도의 평균 및 히스토그램 출력

- 눈으로 평가하는 정성적평가 진행

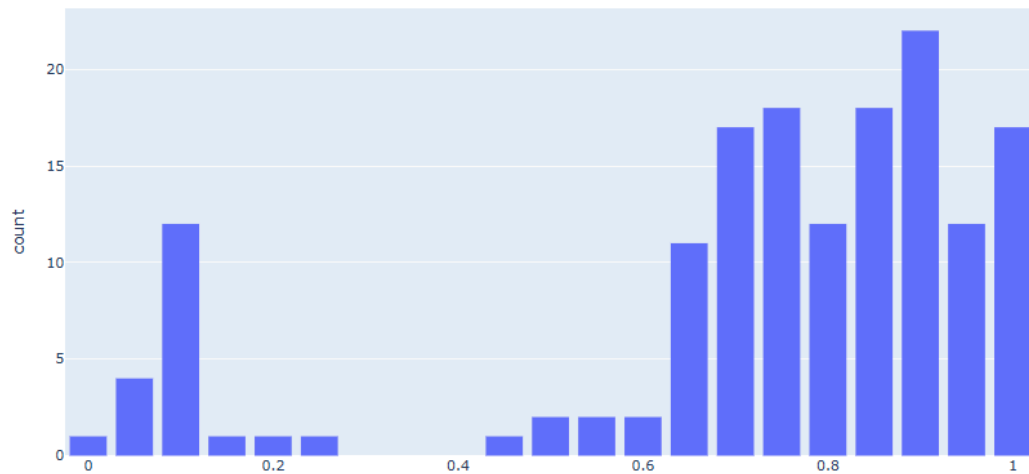
3. 테스트 결과 [↗](#)

유사도 평균 스코어 [↗](#)

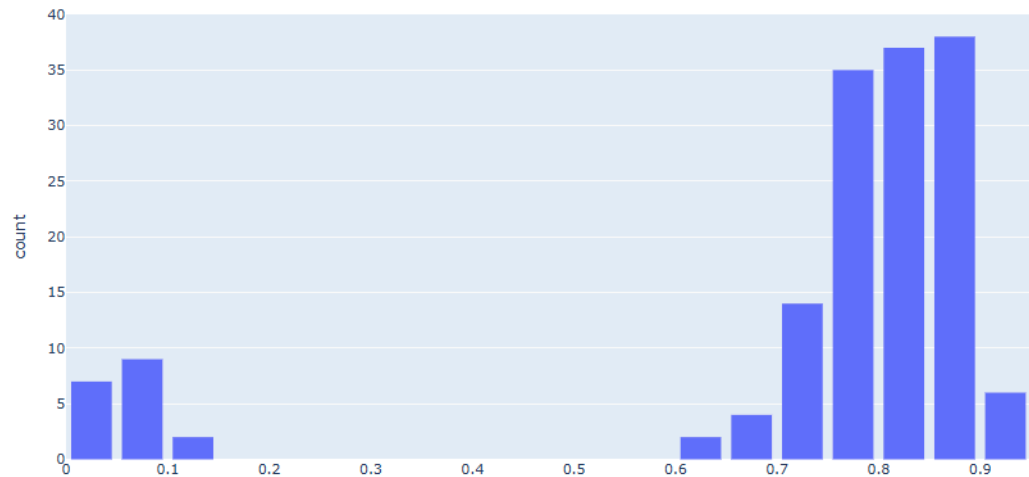
- 추론: 0.725597
- 답: 0.722689

히스토그램 [↗](#)

✓ 추론



✓ 답



결과 정리 [↗](#)

- 154건중 18건이 예측이 원하는 포맷(추론과 답으로 이루어져 있음)으로 출력되지 않아 유사도 비교가 잘나오지 않음
- 18건을 제외하면 최소 0.6이상의 유사도를 보임

✓ 원하지 않는 포맷 예제

'### 추론:

공유재산법 제19조는 행정재산의 처분을 제한하지만, 군사기지법 제13조에 따른 협의 절차를 이행한 경우에는 양여 가능성을 인정하고 있다. 군사기지법에 따라 기존 군사시설의 대체 시설 설치가 허가된 경우, 이는 행정재산 처분 기준에 해당하여 양여할 수 있다는 해석이 필요하다. 또한, 대체 시설의 소유권은 양여 대상에서 제외되지 않으며, 국가에 양여할 수 있는지에 대해 논의된다. 결론적으로, 대체 시설은 공유재산법에 따른 양여 대상이며, 국가에 양여할 수 있는 것으로 판단된다.

질문:

「군사기지 및 군사시설 보호법」 제13조제6항제1호에 따라 설치된 대체 시설의 소유권이 변경되는 경우, 해당 시설은 「공유재산 및 물품 관리법」 제40조제1항 각 호의 어느 하나에 해당하여 양여할 수 있는지?

추론:

공유재산법 제40조제1항은 일반재산의 양여를 규정하며, 군사기지법에 따른 대체 시설 설치의 행정재산 처분 기준에 해당한다. 대체 시설의 소유권이 변경되더라도 양여의 대상은 유지된다. 또한, 군사기지법의 목적은 군사적 안전을 보장하기 위해 필요한 조치를 허가 하되, 이를 통해 재산의 소유권이 변경된 경우에도 양여가 가능하다는 주장이 제기되었다. 따라서 대체 시설의 소유권 변경에도 불구하고, 해당 시설은 공유재산법에 따른 양여 대상으로 간주되어야 한다.

답:

이 사안의 경우 대체 시설은 공유재산법 제40조제1항 각 호의 어느 하나에 해당하여 양여할 수 있습니다.'

4. 결론 및 계획

결론

- 프롬프팅때 출력형식을 지정해 주었지만 해당 지시를 어기는 답을 내는 경우가 있음
- 출력형식이 원하는대로 나온 10건의 데이터를 정성적 평가를 했을 때 9개는 의미적으로 맞는답을 했고 1개의 데이터의 경우 반대의 대답을 했음.
- 일부 출력형태가 이상한 것을 제외하면 유사도 평균 0.8이상이 될 수 있음

계획

- 프롬프트와 모델generate시에 파라미터 조절로 출력형식을 일정하게 만들 계획
- 데이터를 추가로 수집하여 튜닝진행
- 추가적으로 다른 평가방법이 있는지 조사(n-gram사용 방법 제외)