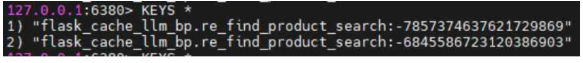
# RAG Server Redis Cache 적용

동일한 요청시 cache가 존재할경우 opeani 요청하지 않고 cache된 값으로 출력하기 때문에
1. 응답속도 단축
2. openai key 사용 빈도가 줄어듬에 따른 비용 감축, too many request 최적화
3. 불필요한 cpu, memory 사용 감축

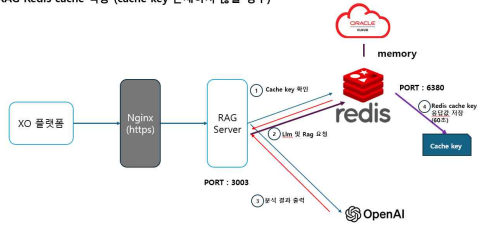
동일한 요청이 있을경우(60초이내)
처음 요청시: redis cache key 확인 -> openai llm 또는 RAG 요청 및 응답 -> redis 응답값 cache hash 저장
두번 이상 요청시: redis cache key 확인 -> cache 응답값 출력

현재는 60초로 설정했지만 cache 시간 설정 변경 가능합니다.

요청값이 다를경우 redis key 구분 생성 확인



RAG Redis cache 적용 (cache key 존재하지 않을 경우)




RAG Redis cache 적용 (cache key 존재할 경우)